

# Data Availability Policies: Ideal and Practice

---

Leen Breure  
Maarten Hoogerwerf



*The Hague 2013 – version 1.1*

## Contents

Preface.....	4
1. Introduction.....	6
1.1. Open science, open data policies.....	6
1.2. Only a limited portion of data is published.....	7
1.3. Promoting a data availability infrastructure .....	8
2. Data availability in economics and economic history .....	10
2.1. Data submission and replication of research.....	10
2.2. Data Availability Policy in economic journals.....	10
3. The journals: ideal and practice .....	14
3.1. Criteria for a good DAP.....	14
3.2. DAP examples in economic journals .....	14
3.3. Problems with data sharing and supplementary material.....	15
4. Data repositories .....	18
4.1. A broad choice of repositories .....	18
4.2. The workflow: data ingest and data enhancement .....	19
4.2.1. Information flow: Dryad and Pensoft.....	20
4.2.2. Enhancement with metadata.....	22
4.2.3. Smart data ingest: BioMed Central and LabArchives.....	23
4.3. Empowering the user .....	25
4.3.1. Access to data in the context of the research process.....	25
4.3.2. Running programs .....	26
5. Conclusions, questions and recommendations.....	29
5.1. Summary .....	29
5.2. Issues concerning the CLIO-INFRA DAP proposal.....	29
5.3. Issues concerning the role of DANS .....	30
5.4. Recommendations .....	32
5.4.1. Assumptions .....	32
5.4.2. A simple DAP .....	32
5.4.3. Separate data review through an online interactive data paper .....	32
Works Cited .....	35

*“The digital era transformed how science was disseminated and in so doing the word ‘paper’ became synonymous with the term ‘PDF’—the same content just delivered differently.*

*We are at a point where the word PDF will soon be replaced by something else; let's just call it an interactive PDF. What I am suggesting is that one day the interactive PDF will be replaced by the scientific workflow as the entity by which we get credit as scientists.*

*The workflow will make science more reproducible and more open, and this is how I want the publisher of the future to handle my scientific output—I want publishers to publish my workflows.*

*The notion of a workflow here is perhaps slightly different than that defined by many of this readership. It is less of a computational workflow, but part process and part container for content (or pointers to that content) that is significantly broader and more integrated than what is sent for publication today, namely, a manuscript and supplemental information in an essentially computationally unusable form.”*

Philip E. Bourne, *What Do I Want from the Publisher of the Future?*  
(Bourne, 2010)

## Preface

This report provides an overview of the current state of Data Availability Policies (DAPs) in various scholarly domains in general and in economics in particular. It was created as a background study to the development of a demonstrator for the promotion of a DAP in economic history, which is part of the CLARIAH Seed Money Projects 2012.

In this context 'economics' and 'economic history' are considered exchangeable, which may be justified by the overlap of journals in which publications on economic historical subjects are published<sup>1</sup>. A much broader perspective was chosen in order to collect useful experience from scientific fields where data submission has already been successful for many years and to sketch the state of the art in data availability in general. Passages from a variety of web documents and publications have been selected, ordered and merged into a running text to provide ample information across different domains to answer the following core questions:

1. Why is a DAP considered as useful? See [chapter 1](#).
2. What is the current practice of DAPs and compliance with DAPS, in particular in economics / economic history? See [chapter 2](#) and [3](#).
3. What is a good DAP? See [chapter 3.1](#), [3.2](#) and [5.4.2](#).
4. How to handle and store the data that are submitted as a consequence of a DAP? This concerns the role of repositories, because journals do not like to store all the material themselves. See [chapter 4](#).
5. What are the choices to be made for the design of the CLARIAH demonstrator? As an aid in answering this question a fully worked-out example has been added, which may be customized. See [chapter 5](#).

The first section of chapter 5 may be read as a brief management summary of the preceding pages of the report (see [chapter 5.1](#)).

The text consists of two main parts, each taking a different perspective: (1) DAP and scientific journals and (2) DAP and repositories. The discussion of the concept of DAP and the compliance of journals with it, is followed by criteria for and examples of good data submission policies in the field of economics. Some lessons can be learned from other disciplines where successful data submission led to serious problems in the review process, which finally made a few journals stop accepting supplementary material at all. The conclusion at the end should support the decision process that precedes the design of the demonstrator by presenting alternative choices and comprises suggestions for the further implementation of a widely acceptable DAP for economic history.

Not everything which is proposed in this report, can be realized in the current pilot project. However, agreeing upon a well-founded shared vision may make it easier to attract new funding in the near future.

Finally, we want to thank our colleague Marjan Grootveld for critical reading of the text and valuable comments.

---

<sup>1</sup> CLARIAH Newsletter 1, 21 December 2012:  
<http://www.clarin.nl/system/files/CLARIAH%20Nieuwsbrief%201%202012-12-21.pdf>

**Please, note:**

- The text is mainly a compilation of web document fragments containing the original wording of the author(s) where required only slightly modified for stylistic reasons (tense, single, plural, etc.). The reference at the end of each fragment refers to the source.
- A few passages in chapter 1 through 4 have been added by the authors; these end with (Breure & Hoogerwerf).
- Only chapter 5 is entirely written by the authors / editors of this report.
- Therefore, to avoid plagiarism, the text in its current state is not suitable for publication and intended **for internal use only**.
- Citations must be made from the source texts only and **not directly from this report**.

Leen Breure  
Maarten Hoogerwerf

The Hague, 2013

# 1. Introduction

## 1.1. Open science, open data policies

Sharing data which is generated by research projects is increasingly being recognized as an academic priority by funders, researchers and publishers (JoRD, 2013). Scientists should communicate the data they collect and the models they create, to allow free and open access, and in ways that are intelligible, assessable and usable for other specialists in the same or linked fields wherever they are in the world. Where data justify it, scientists should make them available in an appropriate data repository. Where possible, communication with a wider public audience should be made a priority, and particularly so in areas where openness is in the public interest (Boulton & etc., 2012).

A valuable distinction is made in some data availability policies between two categories of data: *integral*, which directly support the arguments and conclusions of the article, and *supplementary*, which enhance the article, but are not essential to its argument (JoRD, 2013). As for how best to make data available to non-specialists too many politicians have the illusion that scientific data can be made readily available through an Excel spreadsheet. Data need to be not just accessible, but also

- *intelligible*, so might need to be cast in multiple forms to meet the needs of specialist and lay audiences;
- *assessable*, so that disclosure of sources, funding, methods and other influences allow audiences to make a judgment of the trustworthiness of claims; and
- *usable*, meaning that data are accompanied by explanatory metadata (Noorden, 2012).

Journal	Impact Factor	Policy of Required Public Deposition for Types of Data				Policy of Provision of Materials and Methods			Full data deposited Percentage of papers
		Microarray	Nucleic Acid	Protein	Macromolecular	Materials upon request	Protocols upon request	Condition of publication	
New England Journal of Medicine	52.389								0
Cell	29.887								1
Nature	28.751								0
Lancet	28.638								0
Nature Medicine	26.382								0
Science	26.372								1
Nature Immunology	26.218								9
Nature Genetics	25.556								0
JAMA	25.547								1
Nature Biotechnology	22.848								5
Nature Materials	19.782								0
Immunity	19.266								0
Nature Cell Biology	17.623								0
Journal of Clinical Investigation	16.915								0
Archives of General Psychiatry	15.976								0
Journal of the National Cancer Institute	15.678								0
Nature Neuroscience	15.664								1
Journal of Experimental Medicine	15.612								0
Annals of Internal Medicine	15.516								0
Journal of Clinical Oncology	15.484								0
Nature Methods	15.478								6
Genes and Development	14.795								3
Nature Physics	14.677								0
PLoS Biology	13.501								2
Neuron	13.41								0
Molecular Cell	13.156								0
Circulation	12.755								0
PLoS Medicine	12.601								0
Developmental Cell	12.436								0
Gastroenterology	11.673								0
Genome Research	11.224								6
American Journal of Human Genetics	11.092								3
Nature Structural and Molecular Biology	11.085								0
Journal of the American College of Cardiology	11.054								0
Blood	10.896								0
Hepatology	10.734								0
Current Biology	10.539								0
Gut	10.015								0
British Medical Journal	9.723								0
Circulation Research	9.721								1
Plant Cell	9.653								0
Nano Letters	9.627								0
Journal of Cell Biology	9.598								0
PNAS	9.598								1
Molecular and Cellular Proteomics	9.425								7
PLoS Pathogens	9.336								0
American Journal of Psychiatry	9.127								0
American Journal of Respiratory and Critical Care Medicine	9.074								0
Annals of Neurology	8.813								0
PLoS Genetics	8.721								0

Breakdown of journal policies for public deposition of certain data types, sharing of materials and/or protocols, and whether this is a condition for publication and percentage of papers with fully deposited data <sup>2</sup> – [click to enlarge](#)

As a condition of publication, scientific journals should enforce a requirement that the data on which the argument of the article depends should be accessible, assessable, usable and traceable through

<sup>2</sup> See <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168487/>

information in the article. This should be in line with the practical limits for that field of research. The article should indicate when and under what conditions the data will be available for others to access (Boulton & etc., 2012) (Boulton, Rawlins, Vallance, & Walport, 2011).

Similar observations have been made for economic history, where better coordination has been proposed through the introduction of collaboratories. When the results are published in a DAP journal, the researchers are also requested to deposit their data. Not all journals with such a policy do use internationally accepted data-archiving methods or metadata protocols, but simply put the data on an accessible web site. At the end of the information cycle, the researcher sometimes submits the entire data collection to the data archives, allowing third parties to use it for new research. In the ideal collaboratory information cycle, researchers discuss and fine-tune their ideas about research questions from the start. They exchange ideas about the necessary data and data format and set up a database format that can include data in different but comparable formats (Moor & Zanden, 2008).

## **1.2. Only a limited portion of data is published**

Many academic journals have already explicit policies that require authors to make their data available, but rates of compliance are low. A team led by John Ioannidis, an epidemiologist at Stanford University in California, looked at 500 papers published in 50 top biomedical journals in 2009, and found that of the 351 papers covered by a data-availability policy, 59% didn't adhere to that policy, and only 47 papers deposited full raw primary data online. So, a substantial proportion of original research papers published in high-impact journals are either not subject to any data availability policies, or do not adhere to the data availability instructions in their respective journals (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2012).

The journals *Nature* and *Science* have similar requirements to ensure the availability of data and code to its readers. Replication and robustness studies have been difficult to conduct because they usually require cooperation from the author(s). Researchers frequently fail to keep documented, well organized, and complete records of data and data processing programs underlying published articles, and are less than enthusiastic when asked to help replicate their work.

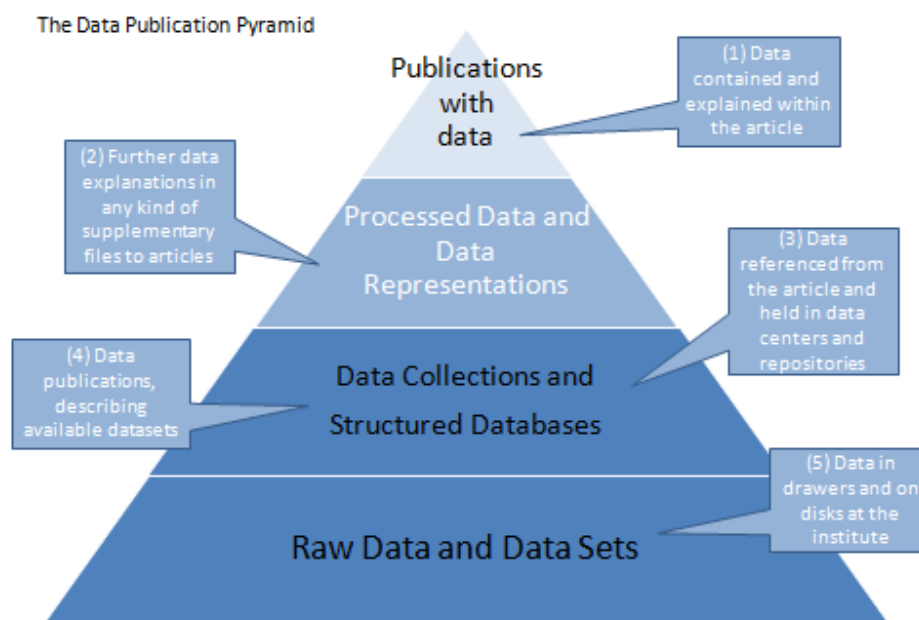
In 2011 PloS (Public Library of Science)<sup>3</sup> conducted a survey in which a total of 1329 scientists participated and explored current data sharing practices, perceptions of the barriers and enablers of data sharing: scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding. Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle (collecting their research data; searching for, describing or cataloging, analyzing, and short-term storage of their data) but are not satisfied with long-term data preservation. Many organizations do not provide support to their researchers for data management both in the short- and long-term. If certain conditions are met (such as formal citation and sharing reprints) respondents agree they are willing to share their data. There are also significant differences and approaches in data management practices based on primary funding agency, subject discipline, age, work focus, and world region (Tenopir, Allard, Douglass, Aydinoglu, & Wu, 2011).

---

<sup>3</sup> PloS: <http://www.plos.org/>

The relationship between data and publications can be illustrated with a modified version of Jim Gray's e-science pyramid, here presented as the Data Publication Pyramid, see the graph below (Hey, Tansley, & Tolle, 2007). As we descend the pyramid the exclusive relationship between data and publication diminishes. At the top, for example, the journal (and author/researcher) takes full responsibility for the publication including the aggregated data embedded in it and the way it is presented.

For data published in the second layer, as supplementary files to articles, the link to the published Record of Science remains strong, but it is not always clear at what level the data is curated and preserved and if the criteria for discoverability and re-usability are met. At the Data Collections and Structured Database layer, the publication includes a citation and links to the data, but the data resides in and is the responsibility of a separate repository. The publication of data becomes collaborative. At the bottom layer of the pyramid, most datasets remain unpublished and hence unfindable and not re-usable. As Jim Gray already made clear, the data published now within or with publications, is only the tip of the data iceberg (Reilly, Schallier, Schrimpf, & Smit, 2011).



The Data Publication Pyramid, developed on the basis of the Jim Gray pyramid, to express the different manifestation forms that research data can have in the publication process.

### 1.3. Promoting a data availability infrastructure

In the United Kingdom a structural monitoring system has been set up in the form of the **JoRD** Policy Bank<sup>4</sup> project, which aims at conducting a feasibility study into the scope and shape of a sustainable service that will collate and summarize the relevant journal policies (JISC-JoRD, 2012). The project will deliver requirements and specifications for a service that will provide researchers, managers of research data and other stakeholders with an easy source of reference to understand and comply with the research data policies of journals and publishers.

A preliminary conclusion is that, although the idea of making scientific data openly accessible for share is widely accepted in the scientific community, the practice confronts serious obstacles. The most immediate of these obstacles is the lack of a consolidated infrastructure for the easy sharing of

<sup>4</sup> JoRD: <http://crc.nottingham.ac.uk/projects/jord.php>



data. In consequence, researchers quite simply do not know how to share their data. At the present juncture, when policies are either not available, or provide inadequate guidance, researchers acknowledge a need for the kind of information that a policy bank would supply. Most of the people interviewed thought that they would use a basic facility: an online searchable database of journal data policies (JoRD, 2013).

**DataCite** is an organization which brings together the datasets community to collaboratively address the challenges of making research data visible and accessible. Members of DataCite meet in person every six months at summer and winter conferences, and collaborate in established working groups.

Through collaboration, it supports:

- researchers by helping them to find, identify, and cite research datasets with confidence
- data centers by providing persistent identifiers for datasets, workflows and standards for data publication
- journal publishers by enabling research articles to be linked to the underlying data.

By working with data centers to assign persistent identifiers to datasets, it helps to develop an infrastructure that supports simple and effective methods of data citation, discovery, and access. In addition, DataCite is developing a number of [services](#) and [resources](#) to support its aims<sup>5</sup>.

---

<sup>5</sup> DataCite: <http://datacite.org/>

## 2. Data availability in economics and economic history

### 2.1. Data submission and replication of research

In economics, as in many other research disciplines, there is a continuous increase in the number of papers where authors have collected their own research data or used external datasets. However, so far there have been few effective means of replicating the results of economic research within the framework of the corresponding article, of verifying them and making them available for repurposing or use in the support of the scholarly debate. One exception are time series analyses based on the national economic accounts: here, the observation points are so scarce that replication and the improvement of research results by applying new methods, are necessarily part of the academic routine (Huschka & Wagner, 2012).

In the light of these findings B.D. McCullough pointed out in 2006:

“Results published in economic journals are accepted at face value and rarely subjected to the independent verification that is the cornerstone of the scientific method. Most results published in economics journals cannot be subjected to verification, even in principle, because authors typically are not required to make their data and code available for verification.” (McCullough, McGeary, & Harrison, 2006)

Harvard professor Gary King also asked:

“If the empirical basis for an article or book cannot be reproduced, of what use to the discipline are its conclusions? What purpose does an article like this serve?” (King, 1995)

Therefore, the management of research data should be considered an important aspect of the economic profession (Vlaeminck (2), 2012).

### 2.2. Data Availability Policy in economic journals

However, there are also successful precedents regarding data availability. In 2005, the *American Economic Review* (AER) imposed such a mandate on authors and found, as one might expect, that it improved the accuracy of research results. In summer 2008, the AER conducted a project to evaluate the quality of the data and processing code contained in its online data archive. The objectives of the project were:

1. to assess the extent to which authors *complied* with the AER’s data submission policy;
2. to evaluate how *easily* results could be *replicated*; and when the materials supplied were complete,
3. to attempt *detailed replications* without contacting the author(s).

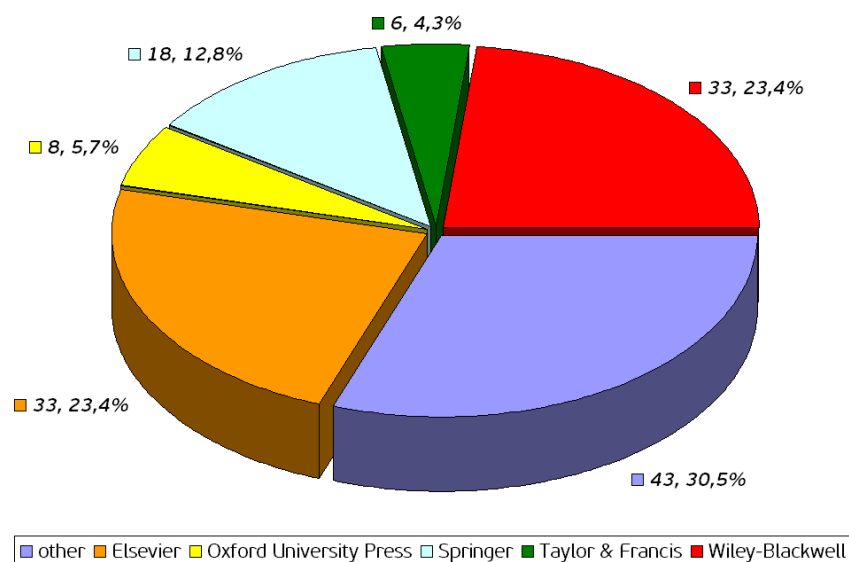
All authors submitted something to the data archive. Roughly 80% of the submissions satisfied the spirit of the AER’s data availability policy, which is to make replication and robustness studies possible independently of the author(s). The replicated results generally agreed with the published results. There remains, however, room for improvement both in terms of compliance with the policy and the quality of the materials that authors submit (Glandon, 2010).

Table I: Data and code submission results by year of publication

	2006	2007	Mar-08	Total
Articles Published <sup>7</sup>	98	100	22	220
Articles Subject to Data Policy	61	63	11	135
Articles Investigated	13	24	2	39
With Readme File	12	23	1	36
	(92%)	(96%)	(50%)	(92%)
With complete submission <sup>8</sup>	7	12	1	20
	(54%)	(50%)	(50%)	(51%)
With proprietary data instructions	1	10	0	11
	(8%)	(42%)	(0%)	(28%)
Articles Investigated believed replicable without contacting the author(s)	8	22	1	31
	(62%)	(92%)	(50%)	(79%)

#### Results replication study AER (Glandon, 2010)

In 2011 EDaWaX (European Data Watch Extended)<sup>6</sup> conducted an evaluation study in which about 140 economic scholarly journals were analyzed regarding their data availability policies (Vlaeminck (2), 2012).



Publishers (number and percentage) in the EDaWaX sample<sup>7</sup>.

Several questions came up concerning the reasons why economics papers may not be replicable in many cases:

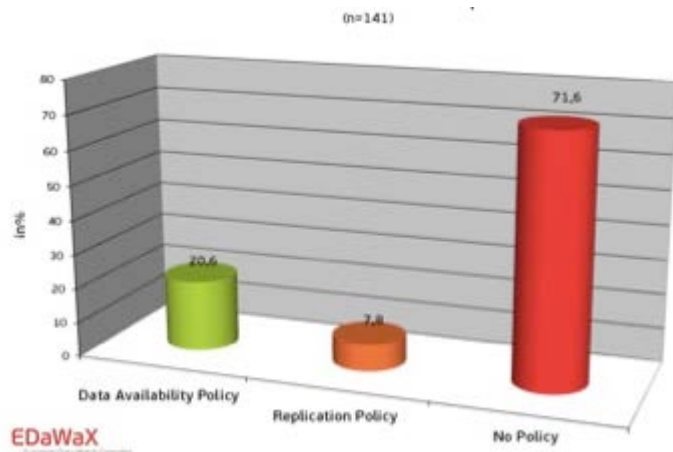
1. What kind of data is needed for replication attempts?
2. When publishing an empirical paper, do economists have to provide their data to the journal?
3. How many scholarly journals commit their authors to do so?
4. Do these journals require their authors to submit only the datasets, or also the code of computation?
5. Do they pledge their authors to provide programs used for estimations or simulations?

<sup>6</sup> EDaWaX: <http://www.edawax.de/>

<sup>7</sup> For details, see: <http://www.edawax.de/2012/04/edawax-wp2-some-information-about-journals-and-selection-of-our-analysis/>

6. What about descriptions of datasets, variables, values or even a manual on how to replicate the results?

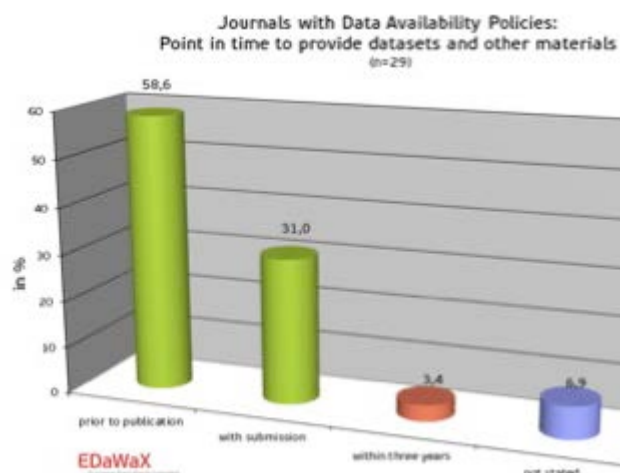
The sample used was *not* representative for economic journals in general, because it mainly consisted of high-ranked journals. Furthermore, by adding some journals explicitly owning a data policy, the percentage of journals that is equipped with such guidelines also is much higher than one might expect for economic journals in general.



In the sample 29 journals were equipped with a data availability policy (20.6%) and 11 journals (7.8%) had a so called “replication policy”<sup>8</sup>. It may be assumed that the majority of economic journals with data (availability) policies were included in the sample.

Results found in this EDaWaX survey:

- Mandatory:** More than 82% of the data policies were mandatory; 26 of the 29 policies (89.7%) pledged authors to submit datasets used for the computation of their results. The remaining journals do not pledge their authors to do so; the journal’s focus often is more oriented towards experimental economic research.
- Documentation:** 65.5% of the journals’ data policies require their authors to provide descriptions of the data submitted and some instructions on how to use the single files submitted.
- Point in time:** 90% of the data availability policies pledge authors to provide their data prior to the publication of an article.
- Exemption:** In economic research it is quite common to use proprietary datasets. Normally, if researchers want to publish an article based on these data, they have to request for an exemption from the data policy. More than 72% of the journals analyzed offered this possibility. More than 60% of the 28 journals allowing exemptions from the data policy had rules for these cases.



<sup>8</sup> “Replication policies” are pledging authors to provide “sufficient data and other materials” on request only, so there are no files authors have to provide to the journal. This approach sounds good in theory – but it does not work in practice because authors often simply refuse to honor the requirements of these policies.

5. **Open formats:** only two journals made recommendations for open formats.<sup>9</sup>
6. **Versions:** the results achieved in economic research are often influenced by the statistical package that was used for calculations. Also the operating system has a bearing on the results. Therefore both the version of the software and the OS used for calculations should be specified. Most of the journals did not mandate their authors to provide these specifications. But there are differences: For example almost every journal that has adopted the data availability policy of the American Economic Review (AER) requires its authors to “document[...] the purpose and format of each file provided” for each file they submit to the journal. In sharp contrast not a single policy required the specification of the operating system used for calculations.
7. **Replication section in journal:** only a very limited number of journals have a section for results of replications.

In summary, it can be stated that the management of publication related research data in economics is still at its early stages. One found 29 journals with data availability policies. That is many more than other researchers found some years ago but compared to the multitude of economic journals in total the percentage of journals equipped with a data availability policy is still quite low. The 20.6% might be the major proportion of all journals equipped with a data policy. Nevertheless, editors and journals in economics seem to be in motion – the topic of data availability seems to become more and more important in economics. This is a positive signal and it will be an interesting aspect to monitor whether and how this upward trend continues.

A large portion of the analyzed data availability policies are mandatory. Moreover, the finding that 90% of the journals are pledging their authors to submit the data prior to the publication of an article shows that many of them have appreciated the importance of providing data at an early stage in the publication process (Vlaeminck (2), 2012).

---

<sup>9</sup> Open formats are important for two reasons: The first is that the long-term preservation of these data is much easier, because the technical specifications of open formats are known. A second reason is that open formats offer the possibility to use data and code in different platforms and software environments.

### 3. The journals: ideal and practice

#### 3.1. Criteria for a good DAP

The EDaWaX survey made clear, that DAPs that aim to ensure the replicability of economic research results, have to:

1. be mandatory,
2. pledge authors to provide datasets, the code of computation, programs and descriptions of the data and variables (in form of a data dictionary at best),
3. assure that the data is provided prior to publication of an article,
4. have defined rules for research based on proprietary or confidential data,
5. provide the data, so other researchers can access these data without problems.

Besides, journals should:

6. have a special section for the results of replication attempts or should at least publish results of replications in addition to the dataset(s),
7. require their authors to provide the data in open formats or in ASCII-format,
8. require their authors to specify the name and version of both the software and the operation system used for analysis (Vlaeminck (2), 2012).

#### 3.2. DAP examples in economic journals

Without claiming that the list below is complete, it seems to be quite easy to get an adequate impression of what a DAP in economics and economic history means. Noteworthy examples of DAPs in economics are:

1. American Economic Review<sup>10</sup>
2. Canadian Journal of Economics<sup>11</sup>
3. Economics, the Open Access, Open Assessment E-journal<sup>12</sup>
4. European Economic Association<sup>13</sup>
5. IMF Economic Review<sup>14</sup>
6. Journal of Political Economy<sup>15</sup>
7. Review of Economic Studies<sup>16</sup>

A detailed comparison of these DAPs doesn't make much sense. Most of the journals in this list follow completely the rather extensive DAP of the *AER*, which comprises mandatory submission of data, programs (with documentation), more specific information about experimental designs, and provides rules for exemption in case open access is not allowed. Only the *IMF Economic Review* and the *Review of Economic Studies* have more concise instructions. So, the DAP of *AER* may be

---

<sup>10</sup> American Economic Review: <http://www.aeaweb.org/aer/data.php>

<sup>11</sup> Canadian Journal of Economics: <http://economics.ca/cje/en/datapolicy.php>

<sup>12</sup> Economics DAP: <http://www.economics-ejournal.org/submission/data-availability-policy>

<sup>13</sup> European Economic Association: <http://www.eeassoc.org/index.php?site=JEFA&page=42>

<sup>14</sup> IMF Economic Review: [http://www.palgrave-journals.com/imfer/author\\_instructions.html#data-availability-policy](http://www.palgrave-journals.com/imfer/author_instructions.html#data-availability-policy)

<sup>15</sup> Journal of Political Economy: <http://www.press.uchicago.edu/journals/jpe/datapolicy.html?journal=jpe>

<sup>16</sup> Review of Economic Studies: <http://sfx.cceu.org.cn/cgi-bin/tgxx.cgi?issn=0034-6527>

considered as the ‘mother of all DAPs’ in this field and can be used as starting point for further discussion (Breure & Hoogerwerf).

### 3.3. Problems with data sharing and supplementary material

One would not expect that a successful policy of publishing supplementary material may become a mixed blessing and a serious problem in the publication workflow. Two examples have been documented on the web.

**CASE 1:** In fall 2010 *The Journal of Neuroscience* announced that it would stop hosting and peer-reviewing supplementary material for articles, so authors were no longer allowed to include any additional materials when they submitted new manuscripts (Vlaeminck (1), 2012).

The major motivation for removing supplementary material from their websites is the following:

1. The amount of material associated with a typical article has grown dramatically. While the size of articles has grown gradually over the past decade, the supplemental material associated with an article grew exponentially. The biggest problem is not the storage itself, but it starts at the point where a journal is not only peer reviewing the article but also the supplementary material .
2. Another troubling problem associated with this context is that the extensive use of supplementary material in journals encourages reviewers to demand even more material and details. Reviewers increasingly insist that authors have to add further analysis, proofs, experiments etc. – even if these additions were subordinate or tangential. For the authors, it is real work and sometimes unjustified burden. Additionally, reviewer’s demands delay publication.

So the journal solved the problem by removing the supplemental material from the peer review process and by requiring that each submission be evaluated and approved as a complete, self-contained scientific report. By allowing the authors to include a link to supplemental material on their own site, readers will continue to have access to any amount of additional material that the authors consider interesting, but with the clear warning that the material has not gone through peer review.

**CASE 2:** In 2011 a similar argument was heard from the *Journal of Experimental Medicine* (JEM)<sup>17</sup>. It decided to accept only “essential” supplementary tables and figures for publication.

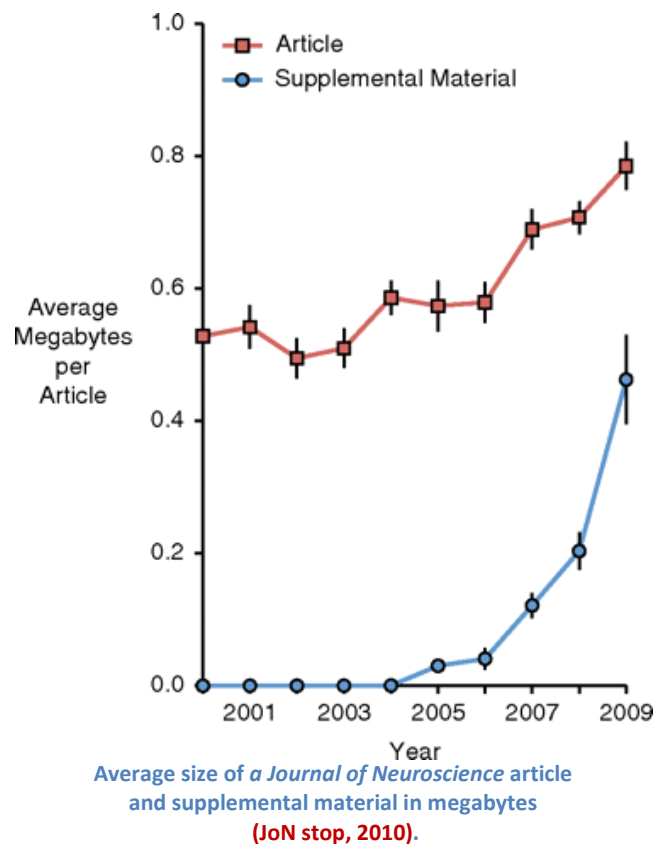
“Complaints about the overabundance of supplementary information in primary research articles have increased in decibel and frequency in the past several years and are now at cacophonous levels. Reviewers and editors warn that they do not have time to scrutinize it. Authors contend that the effort and money needed to produce it exceeds that reasonably spent on a single publication. How often readers actually look at supplemental information is unclear, and most journal websites offer the supplement as an optional download.

The abundance of supplemental material in *The Journal of Experimental Medicine* is growing. In the May 2011 issue of JEM, all research papers have a supplement, with an average of 5.9 supplementary items per paper. Only 3 years ago (May 2008 issue), 16% of JEM papers had no supplementary items, and the overall average was 4.6 items per paper. 3 years before

---

<sup>17</sup> Journal of Experimental Medicine: <http://jem.rupress.org/>

that (May 2005 issue), 57% of papers had no supplementary items, and the average was only 2.4 items per paper.” (Borowski, 2011)



Why the increase in the prevalence of supplementary data? Reviewers frequently asked for it. Editors generally allowed it. So authors were compelled to provide it, although some did so grudgingly (Davis, 2011).

However, problems with data availability are often due to the opposite, a lack of cooperation. Reluctance to data sharing may stem from understandable motives, for example in the domain of health studies. In recent years in North America data sharing has become to be the norm rather than the exception in this field. However, one has defended that it is inappropriate to prescribe exactly when or how researchers should preserve and share data, since these issues are highly specific to each study, the nature of the data collected, who is requesting it, and what they intend to do with it.

The level of ethical concern will vary according to the nature of the information, and the way in which it is collected – analyses of anonymised hospital admission records may carry a quite different ethical burden than analyses of potentially identifiable health information collected directly from the study participants. It is striking that most discussions about data sharing focus almost exclusively on issues of ownership (by the researchers or the funders) and efficiency (on the part of the funders). There is usually little discussion of the ethical issues involved in data sharing, and its implications for the study participants.

It has been noted that simply stripping a data set of unique identifiers such as names, addresses, and identification numbers may not suffice. For example, 97% of records in voter registration lists for Cambridge, MA, could be uniquely identified using birth date and 9-digit zip code. In fact similar



methods have been used on census data in New Zealand to create cohorts that can be followed over time, and linked with cancer and death registration data (Pearce & Smith, 2011).

In other fields negative answers to request for data range from no response from original data owners (public bodies where bureaucratic processes make it hard to give permission for data access), to complex ownership within one dataset, desire to publish on the dataset before making it public, and political reasons or the requirement to receive payment for the data. Some reluctance has been found also in a Dutch survey (Dillo & Doorn, 2011). It is also important to recognize that it has been a long and arduous task for a few individuals to compile these datasets, and that these people have not always been sufficiently remunerated or recognized for their work. This adds even more value to the data, making people reluctant to pass it on (Biofresh, 2011).

## 4. Data repositories

Scholarly journals may want to collect data sets but are usually not equipped to archive and to curate them in a durable manner. So the adoption of a DAP cannot be separated from the role data repositories.

### 4.1. A broad choice of repositories

Open data repositories (public databases, data warehouses, data hosting centers) are subject- or institution-oriented infrastructures, usually based at large national or international institutions. These provide data storage and preservation according to widely accepted standards, and provide free access to their data holdings for anyone to use and re-use under the minimum requirement of attribution, or under an open data waiver (Penev, Mietchen, Chavan, & Hagedorn, 2011).

The landscape of data repositories is very heterogeneous. In addition to national data archives (e.g. DANS) discipline-based repositories have been set up such as Dryad for biological sciences, ChemSpider for chemistry, SPASE for space physics and The Cell for cell biology images.



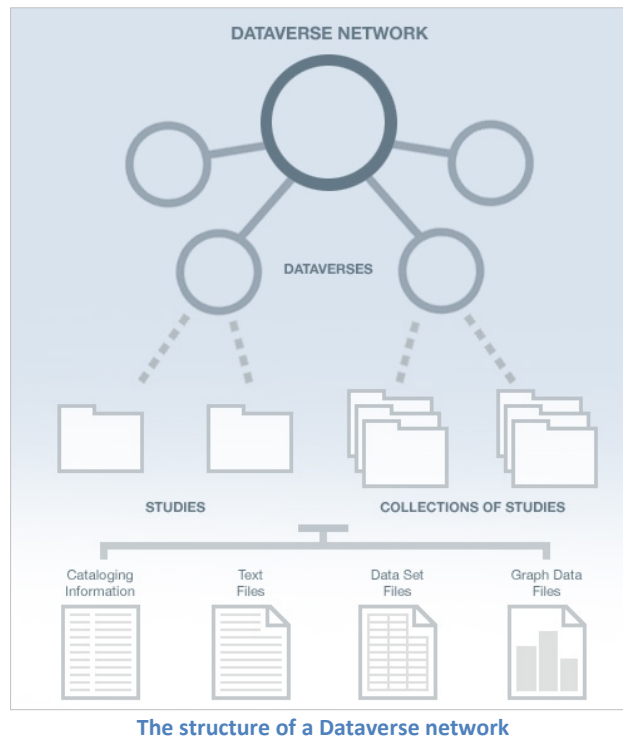
Thus it is difficult for researchers, funding bodies, publishers and scholarly institutions to select appropriate repositories for storage and search of research data. In this respect **Re3data.org**<sup>18</sup> may be helpful, which is a joint project of the Berlin School of Library and Information Science, the German Research Centre for Geosciences (GFZ) and the Karlsruhe Institute of Technology (KIT). It aims at the creation of a global registry of research data repositories and will cover research data repositories from different academic disciplines. Re3data.org will present repositories for the permanent storage of and access to data sets to researchers, funding bodies, publishers and scholarly institutions. It cooperates with DataCite (see [section 1.3](#)) to enhance accessibility and visibility of data sets.

More and more universities and research centers are starting to build their own research data repositories allowing permanent access to data sets in a trustworthy environment. With regard to the storage system **Dataverse**<sup>19</sup> is a popular choice. The Dataverse Network is an open source application to publish, share, reference, extract and analyze research data. It facilitates making data available to others, and allows to replicate others' work. A Dataverse Network hosts multiple dataverses. Each dataverse contains studies or collections of studies, and each study contains cataloging information that describes the data plus the actual data and complementary files.

---

<sup>18</sup> Re3data.org: <http://www.re3data.org/>. See also the list of data repositories: [http://oad.simmons.edu/oadwiki/Data\\_repositories](http://oad.simmons.edu/oadwiki/Data_repositories)

<sup>19</sup> Dataverse Network: <http://thedata.org/book/about-project>



For economic and social history **CESSDA** (Council of European Social Science Data Archives) is relevant. CESSDA<sup>20</sup> is an umbrella organization for social science data archives across Europe. Since the 1970s the members have worked together to improve access to data for researchers and students. CESSDA research and development projects and Expert Seminars enhance exchange of data and technologies among data organizations. Preparations are underway to move CESSDA into a new organization known as CESSDA European Research Infrastructure Consortium (CESSDA ERIC).



Collectively the constituent CESSDA member organizations serve some 30,000+ social science and humanities researchers and students within the European Research Area each year, providing access to 25,000 data collections, delivering over 70,000 data collections per annum and acquiring a further 1,000 data collections each year. The CESSDA Catalogue enables users to locate datasets, as well as questions or variables within datasets, stored at CESSDA archives throughout Europe. Data collections include sociological

surveys, election studies, longitudinal studies, opinion polls, and census data. Among the materials are international and European data such as the European Social Survey, the Eurobarometers, and the International Social Survey Programme.

#### 4.2. The workflow: data ingest and data enhancement

A DAP is about more than just getting the data. There are many organizational points concerning the workflow: to whom should an author send his data: to the editors of the journal, or directly to a repository? Will the journal always correctly forward submitted data, or, if the data archive acts as recipient, will it timely inform the journal? What happens if the publication is rejected? Are the data

<sup>20</sup> CESSDA: <http://www.cessda.org/>

kept in the repository or returned to the authors? Are the data at submission good enough for reuse, or is any further processing required? (Breure & Hoogerwerf)

#### **4.2.1. Information flow: Dryad and Pensoft**

Before starting with a DAP it may be wise to see if there is something to learn from experience elsewhere. One of the cases rather well documented on the web, is that of Dryad and the publishing house Pensoft (Interview Dryad, 2012) (Penev L, 2011) (Shotton, 2011). Dryad<sup>21</sup> is a nonprofit organization and an international repository of data underlying scientific and medical publications. The mission of Dryad is to promote the availability of data underlying findings in the scientific literature for research and educational reuse.

The data publishing workflow of eight journals published by Pensoft, a publisher specialized in biodiversity science and natural history, has now been integrated with the Dryad Digital Repository, facilitating data archiving for data files associated with articles in these journals. The workflow is highly automated thanks to a module of the online editorial management platform, Pensoft Journal System (PJS). Integration with Dryad allows journals to facilitate data archiving by setting up automatic notifications to Dryad from the journals' manuscript submission system.

Upon acceptance of a manuscript in any of these Pensoft journals, the article submission metadata are sent automatically to Dryad, creating a provisional record for the article data. In another automated message, the authors are invited to archive the data files underpinning that particular article, using the provisional record in Dryad to expedite the data upload process. Upon publication, the article metadata and all the associated data files will be available on the Dryad website. Currently, Dryad and Pensoft are exploring the possibilities for a more automated workflow of notification and status changes, from the accepted manuscript through to the published article, enabling the complete bibliographic information about the published article to be available in Dryad on the day of publication.<sup>22</sup> Journal integration with Dryad is available at no cost for any journal that wishes to implement low-burden data archiving and enhance their published articles with links to data.

In more detail, this workflow consists of the following steps. To make archiving as low-burden as possible for authors, data files are archived in conjunction with the journal's manuscript submission process, resulting in permanent 2-way linking between an article and its data:

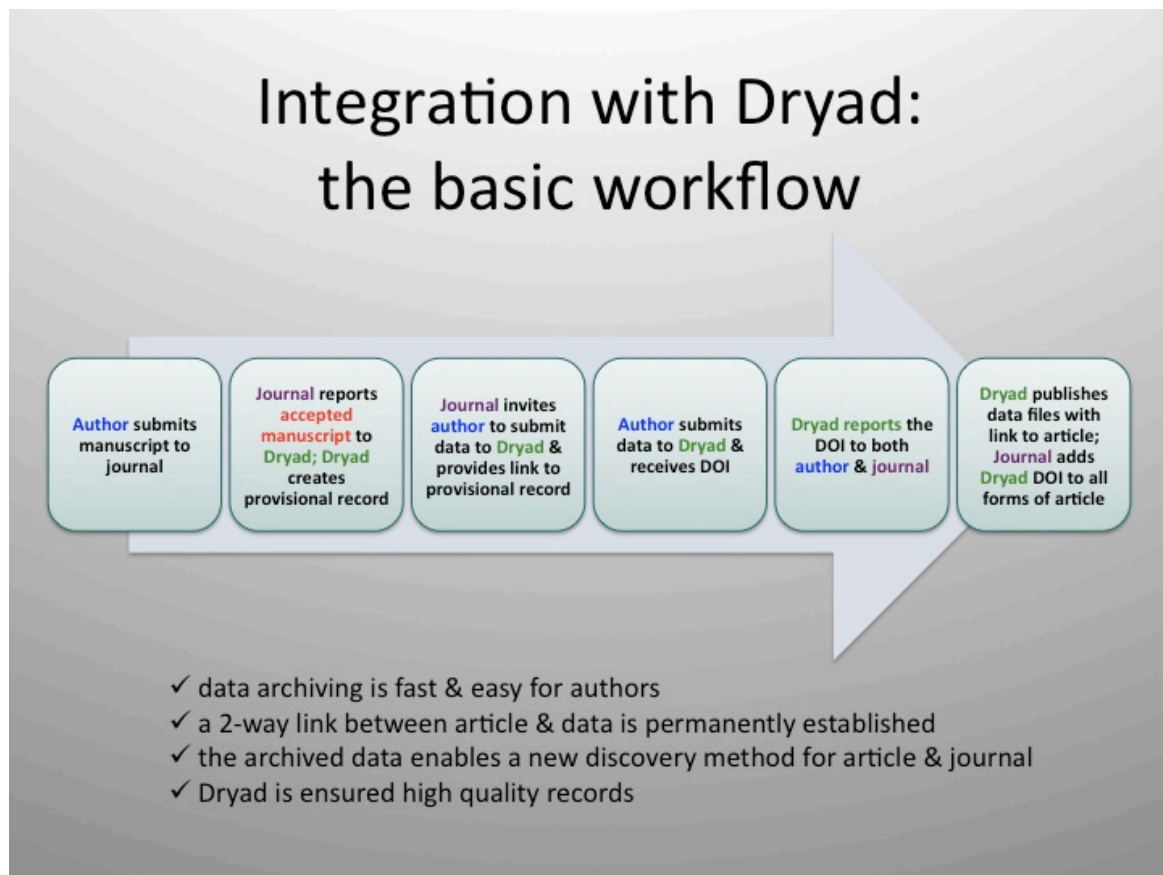
1. Journal provides information about manuscripts to Dryad through automated notices from the manuscript processing system, and invites authors to archive data in Dryad.
2. Authors upload their files to Dryad through a custom submission link supplied by the journal; no redundant information need be entered and the article details are correct.
3. Dryad Curators approve the data files and register its Digital Object Identifier (DOI), a persistent identifier that allows the data to be cited and tracked.
4. Journal and publisher add the Dryad DOI to all forms of the final article, enabling readers of the article to access the data.

---

<sup>21</sup> Dryad: <http://datadryad.org/>

<sup>22</sup> Pensoft: <http://www.pensoft.net/news.php?n=86>

5. Dryad stores the data files, including spreadsheets, images, videos, audio recordings, and many other formats, and links to the article on the journal website. Dryad also provides links to data in other repositories, including sequences in GenBank and phylogenetic trees in TreeBASE.

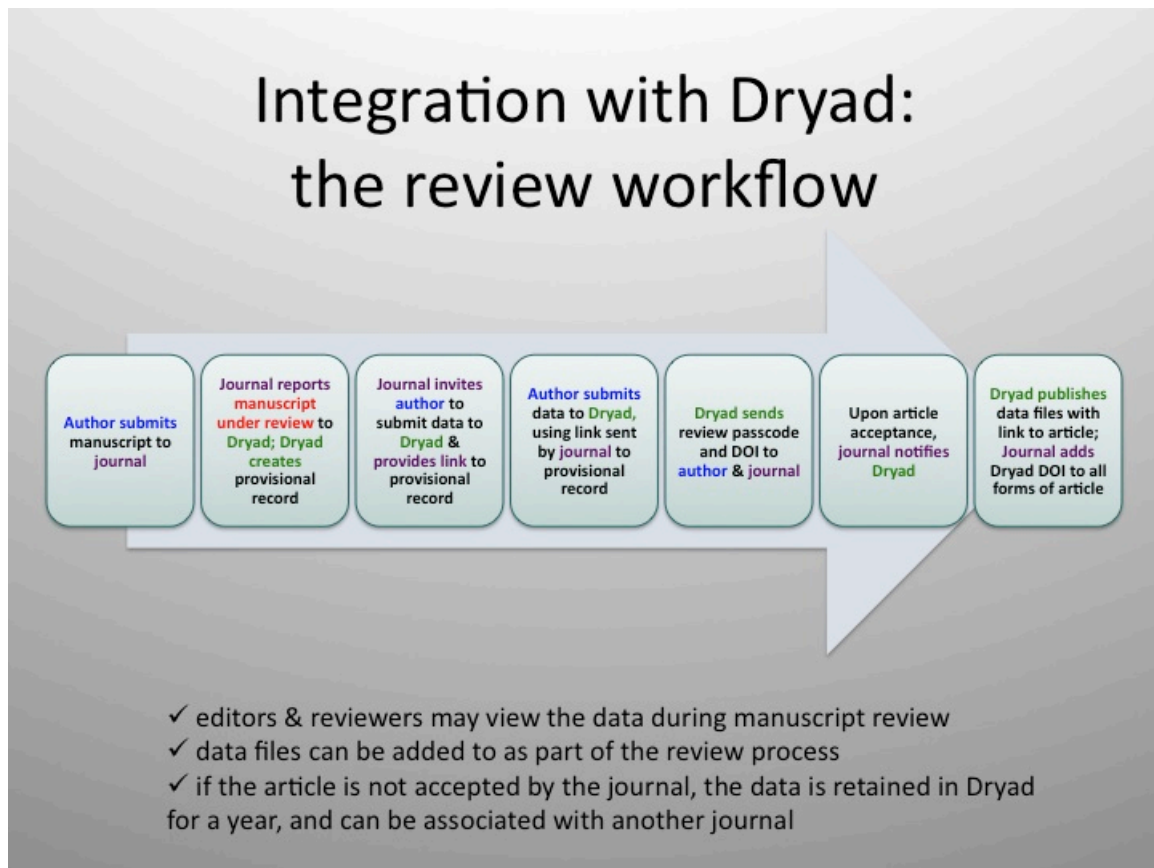


Journals and publishers tell Dryad how they wish to configure integration to meet their needs. Journal editors maintain all communications with authors. Dryad staff only contact authors to approve their data deposits, convey the DOI, or handle questions. These are some of the critical decisions and options for integrating journals:

1. to invite all authors to deposit data, or only selected authors
2. to make data archiving voluntary, or a condition of publication
3. to request author permission before manuscript notifications are sent to Dryad
4. to allow or disallow authors' ability to set a one-year embargo for their data
5. to allow editors to establish custom-length embargoes in special cases
6. to offer anonymous and secure access to the data for editors and reviewers during the manuscript review process
7. to require Dryad to suppress all information about the article until it has been published
8. for journals that publish articles immediately upon acceptance, Dryad can provide a provisional DOI.<sup>23</sup>

---

<sup>23</sup> [http://wiki.datadryad.org/Submission\\_Integration](http://wiki.datadryad.org/Submission_Integration)



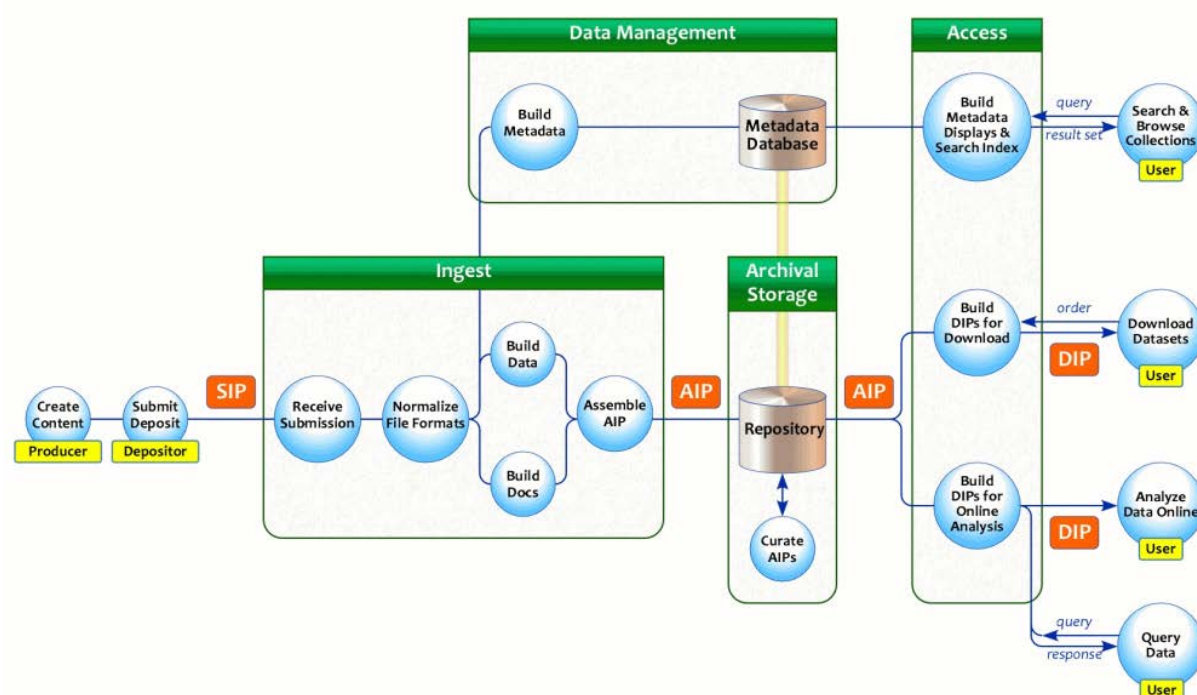
#### 4.2.2. Enhancement with metadata

Some data archives, such as Inter-university Consortium for Political and Social Research (ICPSR)<sup>24</sup> will further enhance the data submitted. This workflow is based on the Reference Model for an Open Archival Information System (OAIS) and comprises, among others, adding metadata according to the DDI-standard<sup>25</sup> (i.e. the Data Documentation Initiative metadata specification, a standard for the content, presentation, transport, and preservation of documentation expressed in XML ). DDI is also used by CESSDA data archives<sup>26</sup>.

<sup>24</sup> ICPSR: <http://www.icpsr.umich.edu/icpsrweb/landing.jsp>

<sup>25</sup> DDI: <http://www.ddialliance.org/>; DDI with ICPSR: <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/metadata.html>

<sup>26</sup> See data documentation with CESSDA: <http://www.cessda.org/sharing/managing/3/index.html>



The internal workflow of ICPSR

#### 4.2.3. Smart data ingest: BioMed Central and LabArchives

The ingest of data into a repository requires at least a formal check and requires always some attention from an archivist to guarantee a minimum quality. A higher standard means even more human intervention, which makes it worthwhile to consider smart software solutions.

*BioMed Central*<sup>27</sup> is an STM (Science, Technology and Medicine) publisher of 243 open access, online, peer-reviewed journals. Recently it has partnered with LabArchives to work together for the shared goal of making datasets supporting peer-reviewed publications available and permanently linked to online publications. LabArchives is the producer of Electronic Lab Notebook, which is used by scientists throughout the world to store, organize, share, and publish their laboratory data. It is a web-based application, which may also be installed on a local server.

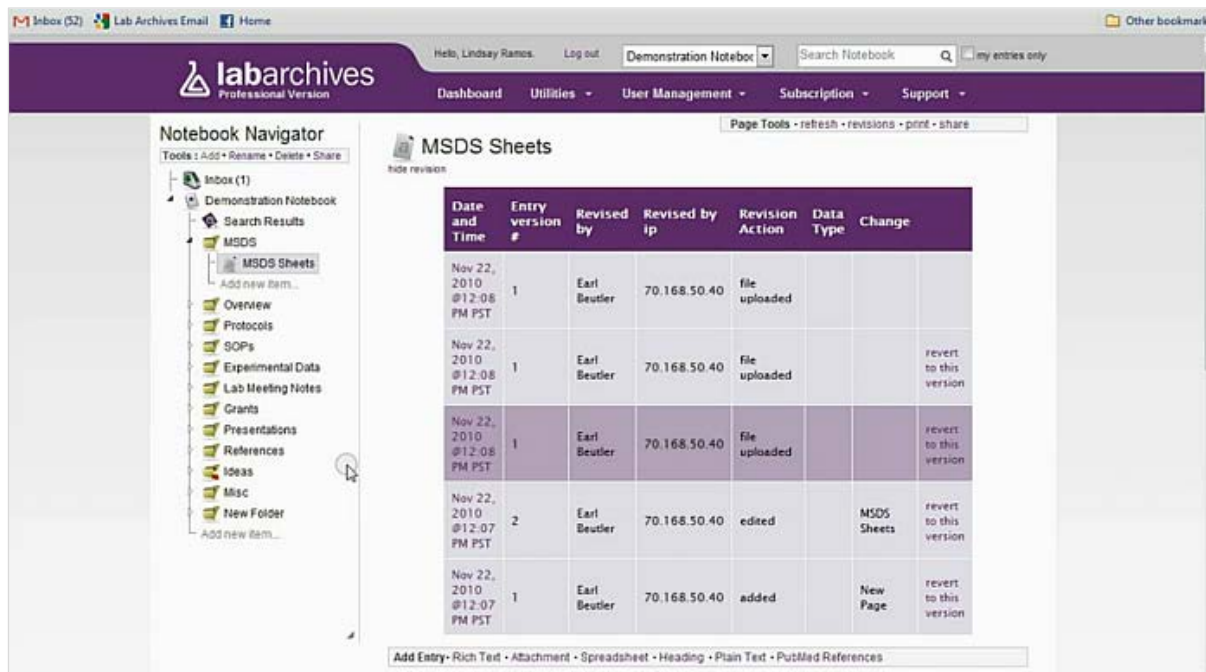
The concept of a lab notebook has benefits, which *mutatis mutandis* are also interesting for non-experimental disciplines. The researcher can directly store and retrieve all his data in a secure environment. All data is stored with multiple redundancy – so, one should never worry about losing information again. Intellectual property is protected. It comes with a versioning system, which keeps all revisions, enabling the researcher to view a complete history of his work. Lab Notebook stores all types of files and supports export facilities as well. It includes viewing software for a variety of file types. This feature enables those who discover the information to be able to see the data, even when stored in certain proprietary formats. Readers (and reusers) of data published and shared through Lab Notebook can view files in context. In this way the user can easily collaborate within the laboratory or across the globe. The application works on any web-connected computer on any platform: Windows, Mac, or Linux.<sup>28</sup>

<sup>27</sup> BioMed Central: <http://www.biomedcentral.com/>

<sup>28</sup> LabArchives: lab notebook, see <http://www.labarchives.com/features.php>



From the point of view of data archiving the lab notebook application is interesting because data are captured directly where they are produced, which may simplify the process of data safe storage and ingest. The application can enforce a basic formal quality, add already some metadata and supports elementary data sharing facilities. For example, in 2011 LabArchives introduced the ability to assign digital object identifiers (DOIs) to datasets stored and shared with the software.



The screenshot shows the LabArchives Professional Version interface. On the left is the 'Notebook Navigator' with a tree view containing 'Inbox (1)', 'Demonstration Notebook', 'Search Results', 'MSDS', 'Overview', 'Protocols', 'SOPs', 'Experimental Data', 'Lab Meeting Notes', 'Grants', 'Presentations', 'References', 'Ideas', 'Misc', and 'New Folder'. The main area displays the 'MSDS Sheets' table. The table has columns: Date and Time, Entry version #, Revised by, Revised by ip, Revision Action, Data Type, and Change. The table contains five rows of data, all from November 22, 2010. The first three rows are 'file uploaded' and the last two are 'edited' and 'added'. Each row has a 'revert to this version' link in the 'Change' column.

Date and Time	Entry version #	Revised by	Revised by ip	Revision Action	Data Type	Change
Nov 22, 2010 @12:08 PM PST	1	Earl Beutler	70.168.50.40	file uploaded		
Nov 22, 2010 @12:08 PM PST	1	Earl Beutler	70.168.50.40	file uploaded		revert to this version
Nov 22, 2010 @12:08 PM PST	1	Earl Beutler	70.168.50.40	file uploaded		revert to this version
Nov 22, 2010 @12:07 PM PST	2	Earl Beutler	70.168.50.40	edited	MSDS Sheets	revert to this version
Nov 22, 2010 @12:07 PM PST	1	Earl Beutler	70.168.50.40	added	New Page	revert to this version

Sample page from a Lab Notebook (taken from video)<sup>29</sup>

A LabArchives user can choose to share a data set as it exists at the time of publication, or they may enable users to continue to view changes as they are made, while, importantly, maintaining the version which supports a peer-reviewed publication. Datasets published via the LabArchives platform and assigned DOIs are available under a Creative Commons CC0 waiver. CC0 helps dispel legal uncertainties about what a person or machine can do with data – or any other content – they discover on the web. CC0 enables cultural (scholarly) norms of citation to take precedence over legal conditions, such as requirements for attribution, for ensuring scientists receive appropriate credit for their contributions.

For publishers to speed publication and reduce barriers to data sharing it is important to better integrate with scientists' workflows and tools, upstream of journal submission and publication. The LabArchives – BioMed Central Edition includes integrated manuscript submission to BioMed Central journals. Authors submitting research manuscripts can, directly from LabArchives, choose the most appropriate of any BioMed Central journal, and authors preparing data notes can link directly to BMC Research Notes' submission system. The manuscript templates for research and data notes are incorporated in LabArchives' integrated Office documents feature, to help speed the process of manuscript preparation (Hrynaskiewicz, 2012).

<sup>29</sup> <http://www.labarchives.com/demo-video/demo.php>



### 4.3. Empowering the user

#### 4.3.1. Access to data in the context of the research process

Particularly in the sciences the argument has been often formulated, that the current state of publication, even in digital format as PDF, is much too limited and do not do sufficient justice to the research process as a whole. Solutions are to be found in a closer cooperation between journals and data archives. The latter should enhance their service and technical capabilities in terms of access and tools (Breure & Hoogerwerf).

A few years ago, Philippe Bourne his made a clear statement in this respect (Bourne, 2010). He represents research as a workflow (but don't confuse his notion of 'workflow' with the one concerning a data archive as discussed above!). First, there is an idea that then is formulated as a hypothesis. An experiment is designed to test that hypothesis. The experiment produces data that are analyzed, generating results. Those results are discussed and conclusions drawn. Today, much of the product of that workflow is in digital form.

Then comes the barrier that the authors climb over to publish. Everything done in the research process needs to be retrofitted to a medium that really does not represent the work in the best possible way. For example, the data from which the conclusions were drawn and the conclusions themselves may now be disjointed, perhaps presented in two separate public repositories (journal and database) with only a tenuous, if any, link between them. Much of the work may have to be omitted to meet restrictions imposed by page limits (or page charges) that do not really make sense in an electronic medium. Visualization of the data, which was so easily accomplished in the laboratory, is impossible in the final published article. In summary, the final published work does not map well to the workflow of the scientific endeavor used to create it.

"In the digital era there is no excuse for not doing better. The digital era transformed how science was disseminated and in so doing the word "paper" became synonymous with the term "PDF"—the same content just delivered differently. We are at a point where the word PDF will soon be replaced by something else; let's just call it an interactive PDF. What I am suggesting is that one day the interactive PDF will be replaced by the scientific workflow as the entity by which we get credit as scientists. The workflow will make science more reproducible and more open, and this is how I want the publisher of the future to handle my scientific output—I want publishers to publish my workflows. The notion of a workflow here is perhaps slightly different than that defined by many of this readership. It is less of a computational workflow, but part process and part container for content (or pointers to that content) that is significantly broader and more integrated than what is sent for publication today, namely, a manuscript and supplemental information in an essentially computationally unusable form." (Bourne, 2010)

Bourne wants the publisher of the future, or the publisher in collaboration with a third party, to be the guardian of these workflows in the same way that today the authors entrust them with the finished product of research. Does publishing more data make any sense in a world weighed down by information overload? The response is that one person's trash is another person's treasure. What is important is that the *tools* exist for a consumer to efficiently make their own judgment between the treasure and the trash. Those tools need to be able to navigate and summarize the workflows and in fact make associations that are just not possible today, but lead to new discoveries.

The PDF is a single static interface, while a workflow is more dynamic and can be viewed from a variety of perspectives in the same way a database or content management system presents multiple views of the content. This flexibility could be very powerful, but would represent a major change for most scientists. A change of work habit is only one major barrier to the workflow vision. There is something comforting about the simple organization of a paper and the relatively brief description of the work relative to what is proposed here. A counter view is that the workflow as content container could include audio and video discussions by the authors that would make the content potentially more accessible (Bourne, 2010).

This idea puts a heavy strain on the user interface of interactive digital publications (and, in the second instance, on repositories which have to feed the interactive components with data). The simple bottom line is that the type of publication we are used to mainly consists of a persuasive discourse supported by static figures with data. Simply linking the full article to complete data sets works fine as long as the amount of data is documented, directly associated with the line of discourse and thus in balance with the reader's capacity to analyze and to understand the information. Otherwise, the PDF tends to become a village in the middle of a data jungle, through which the user himself has to clear a path. So, usability does matter!

One feasible solution is substitution of static tables and graphs with more powerful data presentation components, which present different views on data and connect different data sets, by which the author "can make the data jungle more accessible". In this concept linking data sets occurs in closer connection with the presentation of arguments and on a smaller scale ("the network of paths through the data jungle becomes much denser"). This may also ease the review process of publications with submitted data, because more questions can be answered by the hidden power of the data interface (Breure & Hoogerwerf).

The use of interactive graphs and tables is already standing practice in many Elsevier journals that have implemented the concept of *Article of the Future*<sup>30</sup> and is a great step forward in comparison with the concept of enhanced publication as a bundle of publications, data sets and other information, hold together by meta data. In an ideal world this would require much more service from data repositories. They should do much more than allow users to download a data set, and provide facilities to embed and use data snapshots in the publication itself. This opens the way to the article as a computational document<sup>31</sup>, or executable paper<sup>32</sup>, which may come close to what Bourne had in mind (Breure, Voorbij, & Hoogerwerf, 2011).

#### 4.3.2. Running programs

A related question is to what degree data repositories should or can support the execution of program code that is submitted together with the data sets. A recent solution is RunMyCode<sup>33</sup>. RunMyCode is a web service, launched in January 2012, allowing people to run program codes associated with a scientific publication (articles and working papers) using their own data and parameter values.

---

<sup>30</sup> Article of the Future: <http://www.articleofthefuture.com/>

<sup>31</sup> See, for example, Wolfram CDF: <http://www.wolfram.com/cdf/>

<sup>32</sup> Executable paper: <http://www.executablepapers.com/>

<sup>33</sup> <http://www.runmycode.org/>

RunMyCode.org has three main objectives:

1. to allow researchers to quickly disseminate the results of their research, including their data and code, to an international audience,
2. to provide a very large community of users with the ability to use the latest scientific methods in a user-friendly environment, and
3. to allow members of the academic community (researchers, editors, referees, etc.) to replicate scientific results and to demonstrate their robustness. Doing so permits RunMyCode.org to develop coder profiles, and enables the formation of a collaborative social network.

It is intended to support reproducible research (initially in computational economics). Authors create companion web sites for papers that include the software they used; other people can then re-run their models, and (crucially) play with parameters, using cloud-based instances of those environments. Currently, it only supports MATLAB, R, and SAS right now, but plans to add more tools.

The service only requires a web browser as all calculations are done on a dedicated cloud computer. Once the results are ready, they are automatically displayed to the user. It is also possible to only make the code downloadable, but not executable. In that case, users can download the code and run it on their own computer.

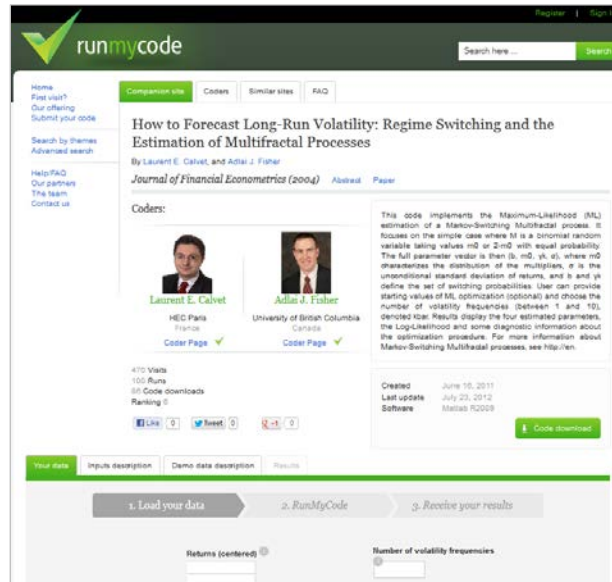


**The RunMyCode architecture.**

The RunMyCode website is operated by a not-profit scientific association called the RunMyCode Association. The mission of the Association is to make research in economics and business easier to use and easier to replicate. RunMyCode is currently funded by several national research agencies and universities.

The RunMyCode.org concept can be viewed as a novel attempt to provide, on a large scale, an executable paper solution. The difference between this and the executable paper approach proposed by the scientific publishers (see previous section) is that the companion webpage is not encapsulated within the text of a scientific publication. In that sense, a companion webpage can be considered as

providing additional material for a scientific publication, in particular the digital objects that permit verification and replication of the published computational results (Stodden, Hurlin, & Pérignon, 2012). A drawback of RunMyCode is that the durability of data storage is unclear; it doesn't seem to be a real data repository having acquired a Data Seal of Approval<sup>34</sup> (Breure & Hoogerwerf).



Example of a RunMyCode companion website<sup>35</sup>.

<sup>34</sup> Data Seal of Approval: <http://datasealofapproval.org/>

<sup>35</sup> Companion website RunMyCode: <http://www.runmycode.org/CompanionSite/site.do?siteId=18>

## 5. Conclusions, questions and recommendations

### 5.1. Summary

The effective use of a DAP implies much more than getting and storing research data. There is a gap between ideal and practice. Although the idea of making scientific data openly accessible for share is widely accepted in the scientific community, the practice confronts serious obstacles. There are many reasons for reluctance among scholars. The most immediate of these obstacles is the lack of a consolidated infrastructure for the easy sharing of data and an effective system of rewarding (see [section 1.3](#) and [section 3.3](#)).

In the field of economics the management of publication related research data is still at its early stages. The 20.6% of journals that has a DAP, might be the main proportion of all journals equipped with a data policy. Nevertheless, editors and journals in economics are in motion – the topic of data availability seems to become more and more important in this field (see [section 2](#)).

But even a successful DAP is not without problems as we have learned from the biomedical field. The data review process may become a bottle neck of the regular publication review process (see [section 3.3](#)). We have concluded that enhanced publications with powerful interactive data exploration and analysis components may ease the data review process (see [section 4.3.1](#)). If replication and re-analysis of data by the user is a serious goal, the digital publication should evolve into the direction of an executable paper. In this respect an organization such as RunMyCode is an interesting experiment (see [section 4.3.2](#)).

### 5.2. Issues concerning the CLIO-INFRA DAP proposal

The most complete DAP in the field of economics and economic history is the one of the *American Economic Review*, which contains as core:

1. “Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Review, *prior to* publication, the *data, programs*, and other *details of the computations* sufficient to permit *replication*.”
2. For econometric and simulation papers, the minimum requirement should include the data set(s) and programs used to run the final models, plus a description of how previous intermediate data sets and programs were employed to create the final data set(s).
3. If a request for an exemption based on proprietary data is made, authors should inform the editors if the data can be accessed or obtained in some other way by independent researchers for purposes of replication<sup>36</sup>.

We shall skip here the instructions for experimental papers, because these are less relevant in this context.

CLIO-INFRA has to negotiate with different journals, which makes that the DAP of the *AER* cannot be adopted without further discussion. If CLIO-INFRA wants to promote a widely acceptable DAP, some questions are to be considered, which concern the work of editors and reviewers:

---

<sup>36</sup> See DAP AER: <http://www.aeaweb.org/aer/data.php>

**Question 1 – Publication review and data review?** A DAP for economic history should be mandatory (see [section 2.2](#) and [3.1](#)), but must data review be implicit? Or, would it be wise to separate both and make data review optional, because it has shown to be a bottleneck in the process (see [section 3.3](#))? Separation may more easily convince reluctant journals to adopt a DAP and further better compliance.

**Question 2 – Replication:** Is replication a serious goal, for example, for reviewers (see [section 2](#))? In economic history much computation seems to be made through spreadsheets. Scripting code in spreadsheets is usually not very transparent. What kind of documentation is required? What about documentation of computation through other software?

**Question 3 – Journal and repository:**

- a. **One-to-one or one-to-many?** Authors may have a preference for a certain data repository or may be limited in their choice by demands from their institution or funding agency. Some journals publish a list of recommended repositories<sup>37</sup>. In addition, a journal may decide to collaborate with a certain repository in particular, which seems to be easier and more efficient communication than with a number of data archives. Should the CLIO-INFRA DAP make any recommendations in this respect?
- b. **Workflow:** if data review is an integral part of the review process of the article, the data concerned may be best submitted to the editors of the journal. This is the standard rule in most data policies. However, Pensoft and Dryad have organized it differently: the publication is submitted to the editors, while the data go to Dryad (see [section 4.2.1](#)). This, of course, requires a good communication between journal and repository and is related to the previous question. In that case a one-to-one match is most obvious.

### 5.3. Issues concerning the role of DANS

**Question 4 – DANS as preferred data repository:** DANS could take the position of one of the many data repositories to be used for submission of CLIO-INFRA data and only play a special role during this project. But the opposite is conceivable as well: DANS could be presented as the preferred repository, not only during this project, but as long term partner of CLIO-INFRA. Of course, data could be also deposited elsewhere if such is required by local regulations (i.e. double deposition).

**Question 5 – DANS as super repository:** In the course of time DANS may acquire the status of super repository by simply being much better than the rest. For example, it could present tools and facilities in the spirit of LabArchives (see [section 4.2.3](#)) and RunMyCode ([section 4.3.2](#)). In this way, DANS may become a super repository by offering more than others do. This extra effort will require a business model, e.g. an annual subscription fee as with many other web services. For example, these extra facilities could get the following implementation:

- a. **Data gathering:** Assuming that many economic historians use spreadsheets for data storage and computation, one can conceive a counterpart of the Electronic Lab Notebook, in the form of a *dedicated spreadsheet* with scripts, facilities for (semi-)automatic adding meta data, and formal checks on completeness, which may support the work of researchers and facilitate the data ingest for both DANS and CLIO-INFRA.

---

<sup>37</sup> For example the *Journal of Open Archaeology Data*: <http://openarchaeologydata.metajnl.com/repositories/>



Currently, the concept of *research data management* is in fashion in the humanities and the social sciences<sup>38</sup>. However, it is noteworthy that the discussion is mainly limited to checklists concerning procedures, responsibilities, metadata, backup procedures etc., while (plans for) supporting, time saving and error prevention software<sup>39</sup> are still missing. This is in contrast with the biomedical field and the hard sciences, where data capture platforms with automated services exist<sup>40</sup>.

Software that could be considered in this context is DATAup<sup>41</sup>, which is an open source tool, helping researchers document, manage, and archive their tabular data. It operates within the scientist's workflow and integrates with Excel. The DataUp tool will parse an .xlsx or .csv file to detect the presence of potential issues that do not comply with data management best practices. In addition to identifying the locations of these problems, DataUp explains why they are potentially problematic, and offers suggested alternatives (of course, users also have the ability to ignore these suggestions). In addition, it creates metadata, obtains persistent identifiers and connects directly to a data repository for uploading. Beforehand there seems no good reason why such an approach is not feasible for economic history.

- b. **Data publishing in context:** Following the notion of *companion pages* (RunMyCode, [section 4.3.2](#) and see also [next section below](#)) and the idea of an *executable paper* (see [section 4.3.1](#)), DANS could facilitate authors who want to create a companion page consisting of (a summary of) the text of the paper with embedded interactive data components like StatPlanet<sup>42</sup> graphs and an interactive spreadsheet such as EditGrid<sup>43</sup>. This would be more complete and more user-friendly than the RunMyCode companion website. For the time being it could be offered as an experimental service, illustrated by one or two articles converted to this format. This would transform the *enhanced* publication (i.e. digital publication linked to data set) into a really *enriched* publication (digital publication with linked companion page containing embedded data).

The latter suggestions may be partly beyond the scope of this project, but could be reckoned with if they are part of plans for the near future. Most of these questions are interrelated and must be discussed in relation to each other. The answers have consequences for both, the presentation and promotion of a DAP for economic historians and for the design of the demonstrator.

---

<sup>38</sup> See ICPSR (<http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html>), the UK Data Archive (<http://www.data-archive.ac.uk/create-manage/planning-for-sharing/data-management-checklist>), DANS (<http://dans.knaw.nl/content/categorieen/nieuws/informatiepakket-voor-datamanagement-biedt-hulp-bij-databaseheer-en-%E2%80%9393opslag> and <http://www.dans.knaw.nl/sites/default/files/file/EASY/Datamanagementplan%20NL%281%29.pdf>).

<sup>39</sup> There are also tools to create a data management plan, e.g. DMP: <https://dmp.cdlib.org/>.

<sup>40</sup> See a few examples: *Research Data Management Tools*, Stanford Center for Clinical and Translational Education and Research ([http://spectrum.stanford.edu/page\\_listings/detail/research-data-management-tools-2](http://spectrum.stanford.edu/page_listings/detail/research-data-management-tools-2)). Project *DaMaRO* at Oxford University (<http://www.edawax.de/tag/research-data-management-tools/>). REDCap (<http://www.ctsi.ufl.edu/research/research-support/redcap/>).

<sup>41</sup> DATAup: [http://dataup.cdlib.org/dataup\\_features.html](http://dataup.cdlib.org/dataup_features.html)

<sup>42</sup> StatPlanet: <http://www.statsilk.com/software/statplanet>

<sup>43</sup> EditGrid: <http://www.editgrid.com/>

## 5.4. Recommendations

### 5.4.1. Assumptions

Answering the above questions is probably easier when one can refer to a sample solution, which may be customized as desired. For recommendations in this respect we make the following assumptions:

1. The goal of this project is to let as many journals as possible accept a DAP. We may expect that the less work this entails, the more success we can expect. So, a DAP should be *simple*, contain essential things only, and should not be a burden on the shoulders of the editorial board.
2. Requiring data together with the submission of a paper implies at least a *moral obligation* to review the data as well. However, we have learned that this combination not only means more work, but may slow down the publication process and may even lead to a bottleneck in the workflow. Notwithstanding that an appropriate solution has to be found for this issue.
3. Compliance with a DAP requires some kind of award for all people involved doing the extra work. Therefore, *credits* should be part of this process.
4. Journals are free to choose what they want; CLIO-INFRA is not in a position to make any demands. So, the request for a DAP should be presented as an attractive offer.

### 5.4.2. A simple DAP

The following text contains a basic regulation only (no details concerning the workflow) and is based partly on the short Joint Data Archiving Policy of Dryad<sup>44</sup> and partly on the DAP of the *American Economic Review*<sup>45</sup>:

1. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Therefore, << journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate data archive preferably having a Data Seal of Approval<sup>46</sup>. Other recommended data repositories are: <<list>>
2. Data together with programs and scripts for computation are to be clearly and precisely documented to allow replication and to be submitted prior to publication of the article.
3. Exceptions may be granted at the discretion of the editor, especially for proprietary data. Authors will have to supply written information on the conditions and procedures by which these data may be obtained.

### 5.4.3. Separate data review through an online interactive data paper

We propose data submission without immediate and mandatory data review. As compensation the system will comprise ample opportunity to make comments on the data and to review them afterwards. This is the point where DANS can play a decisive role.

In 2010 the SURF Share report *Over kwaliteit van onderzoeksdata* was published (Graaf & Waaijers, 2010). It presents the results of an exploratory study carried out to the quality assurance of research datasets. The study consisted of a literature review, 16 interviews with key figures in the field and an

---

<sup>44</sup> JDAP: <http://datadryad.org/pages/jdap>

<sup>45</sup> DAP AER: <http://www.aeaweb.org/aer/data.php>

<sup>46</sup> Data archives having acquired a Data Seal of Approval: <https://assessment.datasealofapproval.org/seals/>



online survey among a representative sample of university professors and associate professors in the Netherlands. It has two conclusions important in this context:

- *Data reviews as part of the peer review process of the journal article:* “It can be concluded from the results of the survey that many scientists deem this desirable, but unfeasible because of the overload of the peer review system”. Only 26.8 % of the interviewees in the social sciences and humanities agreed with a combined peer review process.
- *Commentaries about quality by reusing scholars.* “Scholars who reuse a research dataset will be asked to tag their comments on the quality of (parts of) the dataset on it. These commentaries can be used by other scholars who want to reuse the dataset. Many scientists participating in the survey found this a desirable option”. In the same disciplines about 70% supported this idea and was willing to do the work.

These conclusions are in line with other facts in literature reported above. Therefore, we propose the following:

1. DANS offers the facility of a simple, lightweight interactive *data publication*, based on a template which also contains criteria for reviewing, which are to be respected by the authors. Data quality metrics for business data have been described (Pipino, Lee, & Wang, 2002), and something similar for economic historical data may be feasible as well. The specification of criteria will provide an identical structure to all data publication papers and will ease the job of peer reviewing the data.
2. This data paper contains *interactive data controls* which provide different views on the data set and, perhaps, embedded scripts which helps to detect anomalies easily, for example, unrealistic outliers in the set of values, and other, discipline dependent indicators (if any). Some statistical tests may even run automatically (Data Quality Assessment: A Reviewer's Guide, 2006). If most social and economic historians put their data in Excel, an online equivalent such as EditGrid<sup>47</sup> could be embedded in such a data paper.
3. The interactive data papers will be published by DANS as digital *companion webpages* (article supplements) to the peer reviewed publications in the journal. Nowadays many books have got a companion website, and articles in the sciences and biomedical disciplines have a lot of material which cannot be printed. For example, cardiovascular computed tomography wants to show videos as supplement to a publication and the *Journal of American Folklore* has a multimedia website to distribute audio files. These are only a few examples out of many, so the idea of a companion webpage is not new.
4. Each data paper will get a *persistent identifier* (DOI) and should be counted as a regular publication, which rewards the authors for extra work.
5. The interactive data paper will contain a facility for making *comments* on the data set. So, the data review is not mandatory but the scholarly community is invited to review the data: researchers who want to reuse the data or who are working in the same field, may want to review them (using the criteria specified) or the editorial board may ask somebody to do so. This voluntary effort will also provide to them a reward in the form of greater visibility in their own domain, exactly as activity in social media and blogs does.

---

<sup>47</sup> EditGrid: <http://www.editgrid.com/>

DANS has already got experience with a form of data review (Grootveld & Egmond, 2012), however, this has been rather a data consumer review than a peer review. Some of the questions used allowed entering free text, but most of the questions are five point scales ranging from “bad” (1) to “very good” (5). An example of such a rating is the question to evaluate the quality of the downloaded data. Open questions were used to ask for comments, keywords or tags; these questions are optional. This type of survey therefore yields both quantitative and qualitative information. Our proposal goes one step further into the direction of a real quality review.

Below you can see how users have responded to the dataset 'Bestand Bodemgebruik 2006 - BBG'06'. The legend to the right explains the ratings.

**Ratings:**

- 5: very good
- 4: good
- 3: neither good nor bad
- 2: insufficient
- 1: bad

Aspect	Rating						Average rating
	(5)	(4)	(3)	(2)	(1)	(n/a)	
data quality	2	5	0	0	0	0	★★★★★ (4.29/5)
quality of the documentation	1	4	2	0	0	0	★★★★☆ (3.86/5)
completeness of the data	2	2	2	1	0	0	★★★★☆ (3.71/5)
consistency of the dataset (if applicable)	1	3	1	0	0	2	★★★★☆ (4/5)
structure of the dataset (if applicable)	0	4	1	0	0	2	★★★★☆ (3.8/5)
usefulness of the file formats	2	4	1	0	0	0	★★★★☆ (4.14/5)

6 out of 8 reviewers of this dataset recommend the use of it.

0 out of 8 reviewers of this dataset have published using this dataset.

3 out of 8 reviewers of this dataset intend to use this dataset for a publication.

**Results of consumer data review for a single data set by DANS.**

As noted earlier, journals are free to adopt a complete, integrated publication-plus-data-review, or to skip the data review entirely. By offering a simple but attractive package such as the one outlined above we may expect to obtain a maximum support for a DAP in social and economic history.

## Works Cited

- Data Quality Assessment: A Reviewer's Guide*. (2006). Retrieved from <http://www.epa.gov/QUALITY/qs-docs/g9r-final.pdf>
- Announcement Regarding Supplemental Material*. (2010, August 11). Retrieved from The Journal of Neuroscience: <http://www.jneurosci.org/content/30/32/10599/F1.expansion.htm>
- Requesting data and dealing with complex intellectual property rights issues*. (2011, September 29). Retrieved from The Biofresh Blog: <http://biofreshblog.com/2011/09/29/requesting-data-and-dealing-with-complex-intellectual-property-rights-issues/>
- Bioscientists publish their research data*. (2012, March 6). Retrieved from OpenAIRE Newsletter: <http://www.openaire.eu/en/component/content/article/345-interview-with-a-data-repository-dryad>
- Journal Research Data Policy Bank (JoRD)*. (2012). Retrieved from JISC: [http://www.jisc.ac.uk/whatwedo/programmes/di\\_researchmanagement/managingresearchdata/research-data-publication/jord.aspx](http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata/research-data-publication/jord.aspx)
- A rather long post, but quite a brief summary*. (2013, February 1). Retrieved from JoRD: <http://jordproject.wordpress.com/>
- Alsheikh-Ali, A., Qureshi, W., Al-Mallah, M., & Ioannidis, J. (2012, June 28). *Public Availability of Published Research Data in High-Impact Journals*. Retrieved from Watts Up With That?: <http://wattsupwiththat.com/2012/06/28/editorial-in-nature-calls-for-open-access-data-sciences-capacity-for-self-correction-comes-from-its-openness-to-scrutiny-and-challenge/>
- Borowski, C. (2011, June 6). Enough is enough. *The Journal of Experimental Medicine*.
- Boulton, G., & etc. (2012). *Science as an Open Enterprise*. The Royal Society.
- Boulton, G., Rawlins, M., Vallance, P., & Walport, M. (2011, May 14). Science as a public enterprise: the case for open data. *The Lancet*, 377 (9778), 1633 - 1635. doi:10.1016/S0140-6736(11)60647-8
- Bourne, P. (2010, May 27). What Do I Want from the Publisher of the Future? *PLoS Computational Biology*, 6(5). doi:10.1371/journal.pcbi.1000787
- Breure, L., Voorbij, H., & Hoogerwerf, M. (2011). Rich Internet Publications: 'Show What You Tell'. *Journal of Digital Information*, 12(1). Retrieved February 21, 2013, from <http://journals.tdl.org/jodi/article/view/1606/1738>
- Davis, P. (2011, July 11). *A Journal Is Not a Data Dump*. Retrieved from The Scholarly Kitchen: <http://scholarlykitchen.sspnet.org/2011/07/11/journal-not-data-dump/>
- Dillo, I., & Doorn, P. (2011). *The Dutch data landscape in 32 interviews and a survey*. The Hague: DANS.

- Glandon, P. (2010). *Report on the American Economic Review Data Availability Compliance Project*. Vanderbilt University. Retrieved from [http://www.aeaweb.org/aer/2011\\_Data\\_Compliance\\_Report.pdf](http://www.aeaweb.org/aer/2011_Data_Compliance_Report.pdf)
- Graaf, M. v., & Waaijers, L. (2010). *Over kwaliteit van onderzoeksdata*. SURF Share. Retrieved from [http://www.surf.nl/nl/publicaties/documents/surfshare\\_over\\_kwaliteit\\_van\\_onderzoeksdata\\_dec2010def.pdf](http://www.surf.nl/nl/publicaties/documents/surfshare_over_kwaliteit_van_onderzoeksdata_dec2010def.pdf)
- Grootveld, M., & Egmond, J. v. (2012). Peer-Reviewed Open Research Data: Results of a Pilot. *The International Journal of Digital Curation*, 7(2). doi:doi:10.2218/ijdc.v7i2.231
- Hey, T., Tansley, S., & Tolle, K. (2007). *Jim Gray on eScience: A Transformed Scientific Method. Based on the transcript of a talk given by Jim Gray*. Retrieved from [http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th\\_paradigm\\_book\\_jim\\_gray\\_transcript.pdf](http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf)
- Hrynaskiewicz, I. (2012, April 4). *LabArchives and BioMed Central: a new platform for publishing scientific data*. Retrieved from BioMed Central blog: <http://blogs.biomedcentral.com/bmcblog/2012/04/04/labarchives-and-biomed-central-a-new-platform-for-publishing-scientific-data/>
- Huschka, D., & Wagner, G. G. (2012). *Data accessibility is not sufficient for making replication studies a matter of course*. RatSWD Working Paper Series. Retrieved from [http://www.ratswd.de/download/RatSWD\\_WP\\_2012/RatSWD\\_WP\\_195.pdf](http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_195.pdf)
- King, G. (1995). Replication, Replication. *Political Science and Politics*, 28, 443–499. Retrieved from <http://gking.harvard.edu/files/gking/files/replication.pdf>
- McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006, June). Lessons from the JMCB Archive. *Journal of Money, Credit, and Banking*, 38(4), 1093-1107.
- Moor, T. d., & Zanden, J. v. (2008). Do ut ses (I Give So That You Give Back). Collaboratories as a New Method for Scholarly Communication and Cooperation for Global History. *Historical Methods*, 41(2), pp. 67-78.
- Noorden, R. v. (2012). *Royal Society urges era of open research data*. Retrieved February 16, 2013, from Nature News Blog: <http://blogs.nature.com/news/2012/06/royal-society-urges-era-of-open-research-data.html>
- Pearce, N., & Smith, A. H. (2011). Data sharing: not as simple as it seems. *Environmental Health*, 10(107). doi:10.1186/1476-069X-10-107
- Penev, L. M. D. (2011). *Pensoft Data Publishing Policies and Guidelines for Biodiversity Data*. Pensoft Publishers. Retrieved from [http://www.pensoft.net/J\\_FILES/Pensoft\\_Data\\_Publishing\\_Policies\\_and\\_Guidelines.pdf](http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf)
- Penev, L., Mietchen, D., Chavan, V., & Hagedorn, G. (2011). *Pensoft Data Publishing Policies and Guidelines for Biodiversity Data*. Pensoft. Retrieved from [http://www.pensoft.net/J\\_FILES/Pensoft\\_Data\\_](http://www.pensoft.net/J_FILES/Pensoft_Data_)

- Pipino, L., Lee, Y., & Wang, R. (2002, April). Data quality assessment. *Communications of the ACM*, 45(4). Retrieved from <http://web.cba.neu.edu/~ywlee/publication/PipinoLeeWangCACMApr02.pdf>
- Reilly, S., Schallier, W., Schrimpf, S., & Smit, E. (2011). *Report on Integration of Data and Publications. Opportunities for Data Exchange (ODE) – 7th Framework Programme*. Retrieved from [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-exesummary\\_final.pdf](http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-exesummary_final.pdf)
- Shotton, D. (2011, June 30). *Pensoft Journals policy and author guidelines on data publication and citation*. Retrieved from Open Citations and Related Work: <http://opencitations.wordpress.com/2011/06/30/pensoft-journals-policy-and-author-guidelines-on-data-publication-and-citation/>
- Stodden, V., Hurlin, C., & Pérignon, C. (2012, September 15). RunMyCode.org: a novel dissemination and collaboration platform for executing published computational results. *Social Science Research Network*. Retrieved from [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2147710](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2147710)
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., & Wu, L. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6). doi:10.1371/journal.pone.0021101
- Vlaeminck (1), S. (2012, March 23). *Journal in Neurosciences banned Supplementary Materials*. Retrieved from EDaWaX blog: <http://www.edawax.de/2012/03/journal-in-neurosciences-banned-supplementary-materials/#more-752>
- Vlaeminck (2), S. (2012). *Research Data Management in Economic Journals*. Retrieved February 16, 2013, from <http://openeconomics.net/resources/data-policies-of-economic-journals/>
- Vlaeminck (3), S. (2012, April 2). *EDaWaX WP2: Some Information about Journals and Selection of our Analysis*. Retrieved from EDaWaX blog: <http://www.edawax.de/2012/04/edawax-wp2-some-information-about-journals-and-selection-of-our-analysis/>