Vlaeminck, Sven; Wagner, Gert G.

Working Paper

On the role of research data centres in the management of publication-related research data - Results of a survey among scientific infrastructure service providers

RatSWD Working Paper, No. 226

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Vlaeminck, Sven; Wagner, Gert G. (2013) : On the role of research data centres in the management of publication-related research data - Results of a survey among scientific infrastructure service providers, RatSWD Working Paper, No. 226, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/88148

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

226

On the role of research data centres in the management of publication-related research data

Results of a survey among scientific infrastructure service providers

Sven Vlaeminck and Gert G. Wagner

October 2013

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

On the role of research data centres in the management of publication-related research data

Results of a survey among scientific infrastructure service providers

Sven Vlaeminck

German National Library of Economics / Leibniz Information Centre for Economics (ZBW), Hamburg

Gert G. Wagner

German Data Forum (RatSWD), TU Berlin, MPI for Human Development, Berlin, and German Institute for Economic Research (DIW Berlin)

The findings presented in this article have been achieved in the course of the research project EDaWaX (European Data Watch Extended, www.edawax.de). EDaWaX is funded by the German Research Foundation (www.dfg.de). The institutions listed hereafter are involved in the project: The German Data Forum (RatSWD), the institute Inno-tec of the LMU Munich in cooperation with the Max Planck Institute for Intellectual Property and Competition Law (IMPRS-CI) as well as the German National Library of Economics / Leibniz Information-Centre for Economics (ZBW). Beside the authors the following persons are involved in the EDaWaX project: Prof. Klaus Tochtermann (ZBW), Prof. Joachim Wagner (Leuphana University, Lueneburg), Prof. Dietmar Harhoff (IMPRS-CI and MCIER), Dr. Brigitte Preissl (ZBW), Patrick Andreoli-Versbach (IMPRS-CI), Dr. Frank Mueller-Langer (IMPRS-CI and MCIER), Olaf Siegert (ZBW), Ralf Toepfer (ZBW) and Dr. Hendrik Bunke (ZBW).

Summary

This paper summarizes the findings of an analysis among scientific infrastructure service providers. These service providers have been evaluated in regard to their potential services for the management of publication-related research data. By conducting a desk research and an online survey, we found out that almost three quarters of all responding research data centres, archives and libraries generally store externally generated research data – what also applies to publication-related data. Almost 75% of all respondents also store and host the code of computation (the syntax of statistical analyses). If self-written software components have been used to generate research outputs, only 40% of all respondents accept these software components for storing and hosting. Eight in ten institutions also stated that they are taking specific actions for digital long-term preservation of their data. In regard to the documentation of stored and hosted research data almost 70% of all respondents claimed to use the metadata schema of the Data Documentation Initiative (DDI); Dublin Core was used by 30 percent (multiple answers were permitted). Almost two thirds also used persistent identifiers to facilitate citation of these datasets. Three in four respondents also stated to support researchers in creating metadata for their data. Application programming interfaces (APIs) for uploading or searching datasets currently have not been implemented by any of the respondents yet. Little widespread is the use of semantic technologies like RDF.

A German version of this paper is available as RatSWD Working Paper No. 225.

JEL Classification: C81, C88, H42, H54

Keywords: Research Data Centers, Libraries, Archives, Research Data Management, Journals, Replicability

Background

In economics more and more publications in scientific journals are empirical research papers, in which the authors evaluated either self-produced or externally available datasets along their own research interests.

Compared to other branches of empirical research the compilation of own datasets is not common in economics. A major exception is the field of experimental economics, where researchers often generate their own datasets in the course of investigations motivated by game theory. But these datasets are typically not documented appropriately or even archived for re-examination. Instead, empirical economists frequently use data received from official statistics or from surveys by specialised research bodies (e.g. from the ALLBUS of GESIS[1] or from the SOEP at DIW Berlin[2]). In addition, relevant data may often also be bought from companies like Thomson Reuters or Bloomberg.

Although a rising number of publications in economics (as in most of other scientific disciplines) is based on the analysis of datasets, there are currently few effective means to effectively replicate or re-examine the results of an empirical article, to verify it, or to make it available for re-utilisation and for the support of scholarly debates.

Even research data, that -in principle- is publicly available, will typically not be archived (e.g., in a final working-file) with respect to the specific selection and adjustment procedures. Thereby, replications will not necessarily be prevented, but they are extremely difficult in the cases of ambitious analysis based on specific data selections and calculations.

This current situation confronts both the scientific community and scientific infrastructure service providers like libraries and research data centres with multiple challenges.

[1] The ALLBUS (German General Social Survey) collects up-to-date data on attitudes, behaviour, and social structure in Germany. Since 1980 a representative cross section of the population is surveyed by GESIS every two years using both constant and variable questions. Cf. http://www.gesis.org/en/allbus/allbus-home/

[2] The German Socio-Economic Panel Study (SOEP) is a wide-ranging representative longitudinal study of private households, in which more than 20,000 persons and 11,000 households are interviewed on behalf of German Institute for Economic Research (DIW) Berlin each year. Cf. http://www.soep.de

Why is economic research often not replicable?

According to the literature the reasons for missing replicability of economic research may be located in different areas:

- First and most important is, that there is a lack of incentives for researchers to share their data with the community. The academic reward system does not honour the often time-consuming efforts of data sharing — in sharp contrast to publications, although "*[a]n applied economics article is only the advertising for the data and code that produced the published results*" (Anderson, Greene, McCullough and Vinod (2008), 101).

- Furthermore, economists may worry, that data sharing could lead to personal disadvantages. Because researchers who work up and share data with the community do not receive appropriate compensation, e. g. reputation, for their efforts and might even suffer from disadvantages in terms of academic career because data sharing takes time which cannot be spend on own research. In addition, many researchers suspect others to "misuse" their data, for example by faulty interpretations or by using a dataset without due reference to the creator of the dataset. Eventually, the legal status of research data with regard to data sharing is not sufficiently clear, which also leads to reservations in data sharing (Siegert, Toepfer and Vlaeminck, 2012).[3]

- Only few economic journals have currently implemented guidelines pledging their authors to provide the data and code of computation of their statistical analysis. So called "data availability policies" may in some instances oblige the authors of empirical research papers to supply the underlying data of their results and the code/syntax of their analysis along with the manuscript of the article. Those policies often are in line with the "replication standard" formulated by Gary King (1995).

[3] Indeed, various reports and legal opinions on research data handling have been published in recent years, but it remains questionable if the uncertainty on the part of researchers has thereby been reduced. Cf. Häder, M. (2009): Der Datenschutz in den Sozialwissenschaften. Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland. RatSWD Working Paper Series (90). Available on: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf; Hillegeist, T. (2012): Rechtliche Probleme der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten, Göttinger Schriften zur Internetforschung (8). Available on: http://webdoc.sub.gwdg.de/univerlag/2012/GSI8_Hillegeist.pdf (especially Chapter A); Spindler, G./Hillegeist T. (2008): KoLaWiss-Gutachten AP 4: Recht, Rechtsexpertise für das Projekt „Kooperative Langzeitarchivierung an Wissenschaftsstandorten" (KoLaWiss). Available on: http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf as well as De Cock Bruning, M. / van Dither, B. / Jeppersen de Boer, C.G. / Ringnalda, A. (2011): The legal status of research data in the Knowledge Exchange partner countries. Available on: http://www.knowledge-exchange.info/Default.aspx?ID=461.

- Useful infrastructure components for the management of publication-related research data are rarely being applied, which in turn prohibits any uniform way of citing the underlying data. Available technical solutions like Dataverse [4], a powerful tool for managing and documenting publication-related research data, are being adopted by few journals only. In this context a critical point is, how professional research data centres are handling research-related data and what kind of services, if any, they are offering.

Since autumn 2011 these issues are systematically being addressed by the project EDaWaX (European Data Watch Extended – www.edawax.de) funded by the German Research Foundation (DFG) (cf. Vlaeminck, Wagner, Wagner, Harhoff and Siegert, 2013). Some of the first findings are summarized in other publications from the project: One article describes the data sharing behaviour of applied economists (Andreoli-Versbach and Mueller-Langer, 2013), other publications deal with an analysis of data management in economics journals (cf. Vlaeminck, 2013).

The supplementary working paper at hand describes the results of an evaluation of scientific infrastructure service providers with regard to potential services for the management of publication-related research data in the social sciences and economics.

Do research data centres offer services for archiving publication-related research data?

Especially research data centres could actually be ideal institutions for managing publication-related research data published as attachments to articles within scholarly journals. These capacities originate from decades of expertise in the handling of social- and economic research data, from core-competencies in the creation and maintenance of metadata collected and tagged from surveys and, last but not least, from experiences in managing access to these data (cf. Research Information Network, 2011).

Therefore, the project EDaWaX conducted a study evaluating if such services for publication-related research data are currently available from scientific infrastructure service providers like research data centres, libraries and archives. For this purpose a list of 46 scientific infrastructure organisations was prepared. It includes all German research data centres and data service centres accredited by the German Data Forum (RatSWD) [5], research data centres organised within the Council of European Social Science Data Archives (CESSDA)[6], the library networks in Germany as well as single libraries and public archives.

[4] Webseite: http://www.thedata.org/
[5] The website of the German Data Forum can be found on: http://ratswd.de/eng/index.html
[6] The website of CESSDA can be found on: http://www.cessda.org

In a first step, the websites of these organisations have been examined with regard to potential services for storing and hosting publication-related research data.

The results of the inquiry showed that a publication-related archive[7] is existing at the ICPRS (Inter-university Consortium for Political and Social Research - University of Michigan), which is already used by numerous authors to deposit their publication-related data.[8] DANS EASY[9] – a service located in the Netherlands – does not offer a specific service for publication-related data, but can in principle also be used to deposit such data.[10] However, the desk research could not uncover other indications for further analysis, which is why more information had to be raised by an online questionnaire in order to start a more detailed evaluation of potential services by these organisations.

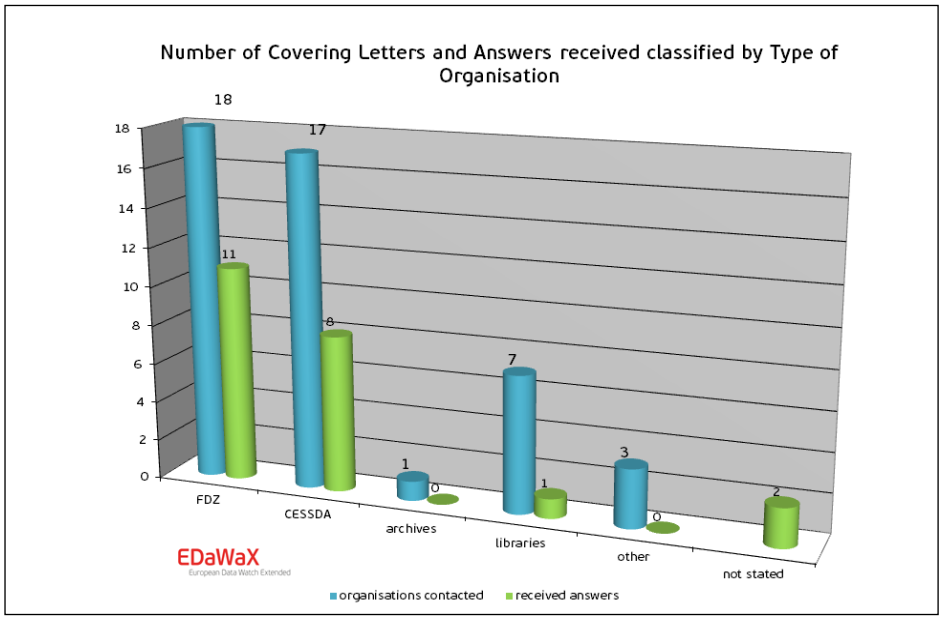
The online-survey

In October and November 2012 an online-questionnaire was sent to 46 organisations – among them 36 national and international research data and data service centres, 1 archive, 7 library networks and single libraries and three other organisations (non-European research data centres). 22 organisations responded to our survey (48%). This return rate may be considered as quite satisfactory, especially when compared to average return rates of written surveys.

Due to the structure of the questionnaire not all participating organisations responded to all questions, which explain deviations in the number of responses.

Certainly more important than the return rate is the structure of respondents and non-respondents. The big majority of all respondents came from research data centres in Germany and Europe (86.4%). Significantly under-represented were respondents from German library networks and archives, but also three research data centres from non-European areas did not respond.

[7] Meanwhile ICPSR's publication related archive has changed its name in „replication datasets".
[8] A list of all journals and articles in which data stored at the ICPSR-PRA is included are available at http://www.icpsr.umich.edu/icpsrweb/ICPSR/biblio/journals?collection=DATA
[9] The Website of DANS EASY can be found on: https://easy.dans.knaw.nl/ui/home
[10] Useful information for instance is provided in the document: „Deposit instructions for social and behavioural sciences" available on:
http://www.dans.knaw.nl/sites/default/files/file/EASY/Deponeerinstructie%20MaGw%20UK%20DEF.pdf



We can only presume that the library networks and the archive do not offer any relevant service for research data management, and therefore did not respond to our survey.

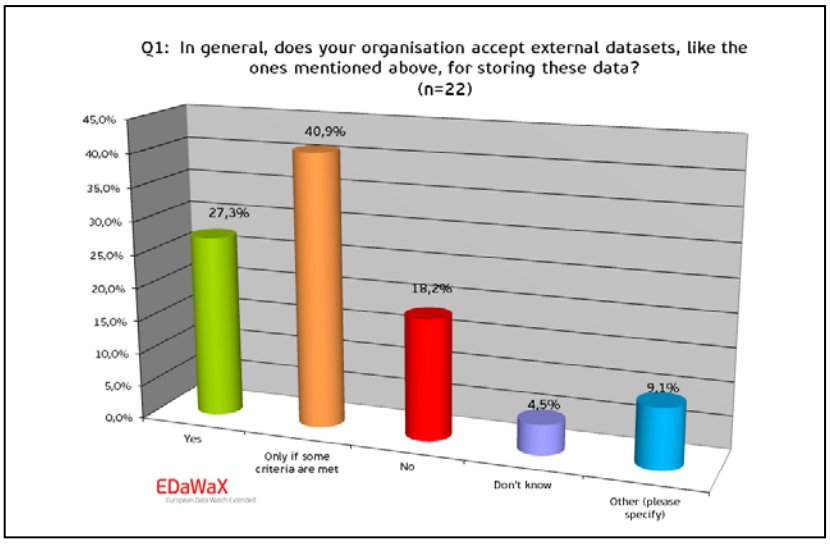
Empirical findings

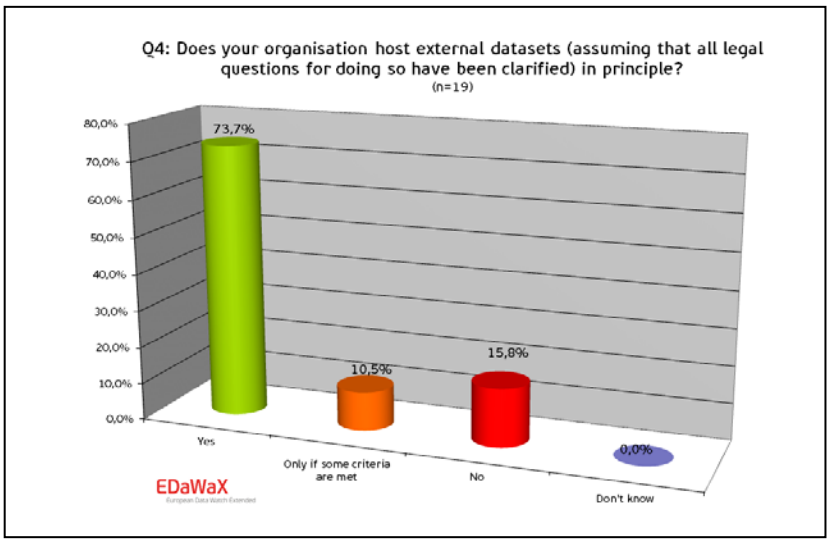
Initially, the survey asked, whether institutions would, in principle, host and store publication-related research data.[10] In addition, the survey also asked, whether organisations would also host and store (self-compiled) software components and the code of computation/syntax of statistical analyses. These three types of data often are part of empirical submissions to economic journals.[11]

Datasets

More than three-fourths of all organisations examined are generally accepting external datasets for storage. At the same time the lion's share of respondents reported, that research data would only be accepted, if certain criteria were met. Such criteria are subject to the specific competencies of many research data centres, but also to the specific regional/supra-regional or national competencies. Moreover, technical and organisational aspects (e.g. proper documentation, machine-readability…) and legal problems were cited as criteria. Approximately 74% of the respondents indicated, that their organisations would also host these types of data. If any criteria for hosting were mentioned, the subject-specific orientation of an institution was stated as main criterion.

[11] The required elements depend on the type of research. The data availability policy of the American Economic Review (AER) – available on http://www.aeaweb.org/aer/data.php – exemplifies such requirements exemplarily.

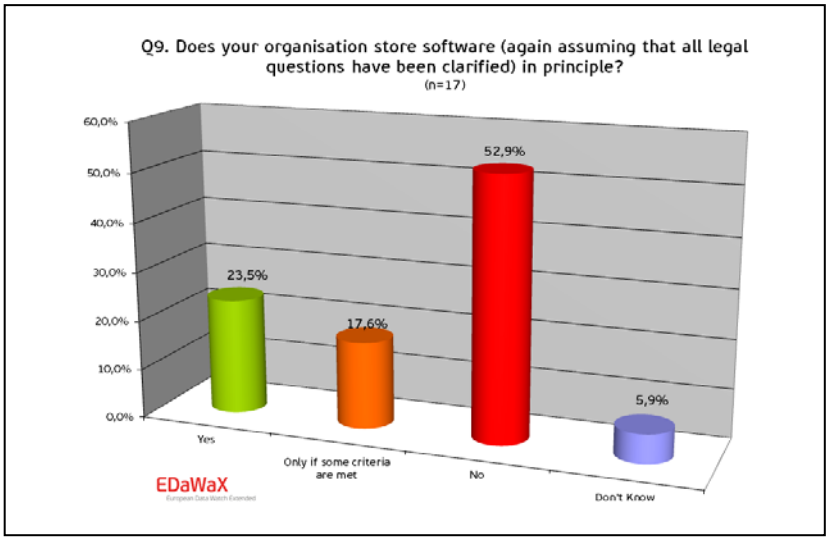




Software

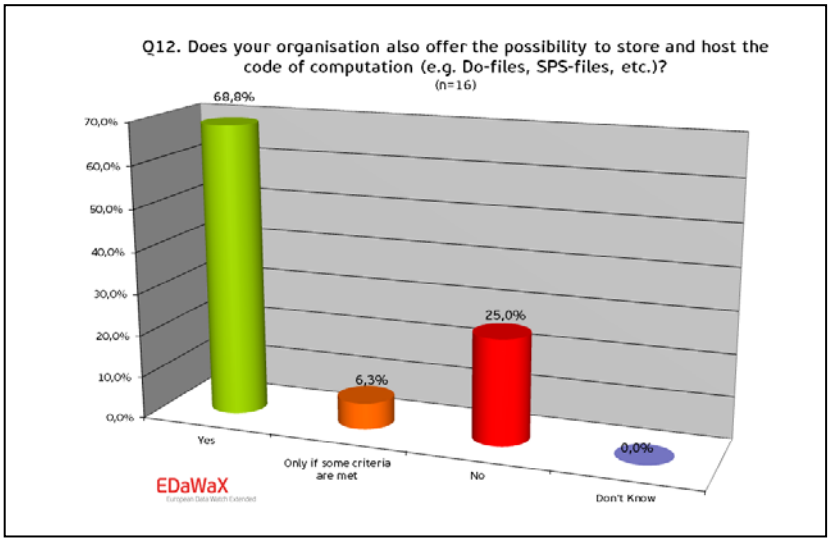
With regard to storing and hosting of (self-compiled) software components, which are often used for economic simulations, our survey indicates that only a minority of just under a fourth of the organisations accepts storing and hosting software components without restrictions. Another 17% pointed out that they established criteria for assessing, if software could be stored and hosted (e.g., *if essential for the analysis of the data*).

Therefore, hosting and storing software components can be considered as a gap. Only a limited number of organisations are offering this service.



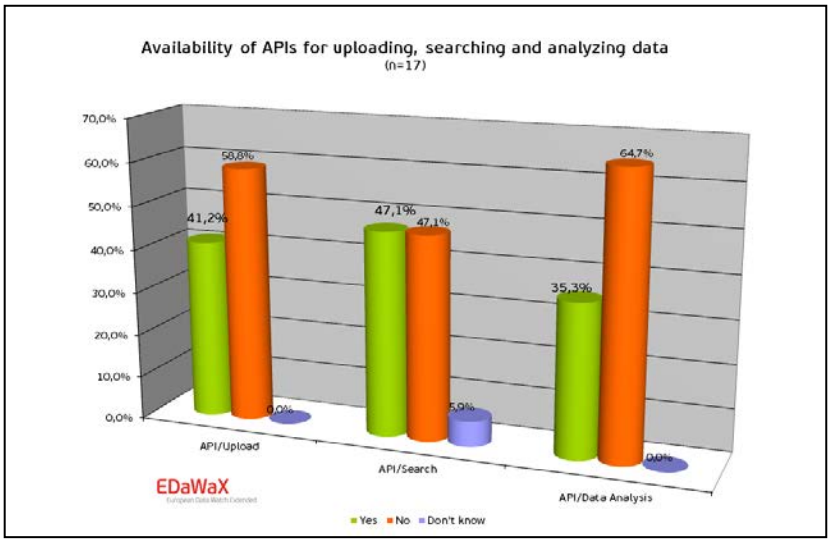
Code of computation

Almost 70% of the organisations examined offer options to store and host the code of computation. However, a quarter of all organisations is not considering to do so now or in the near future. One respondent also stated a criterion – he mentioned that storing and hosting of these data would only be useful in the case of derived variables.



APIs

Within our analyses we also examined the availability of application programming interfaces (APIs), which enable automated exchanges of data.Our results show that less than half of all organisations are having these interfaces at their disposal.



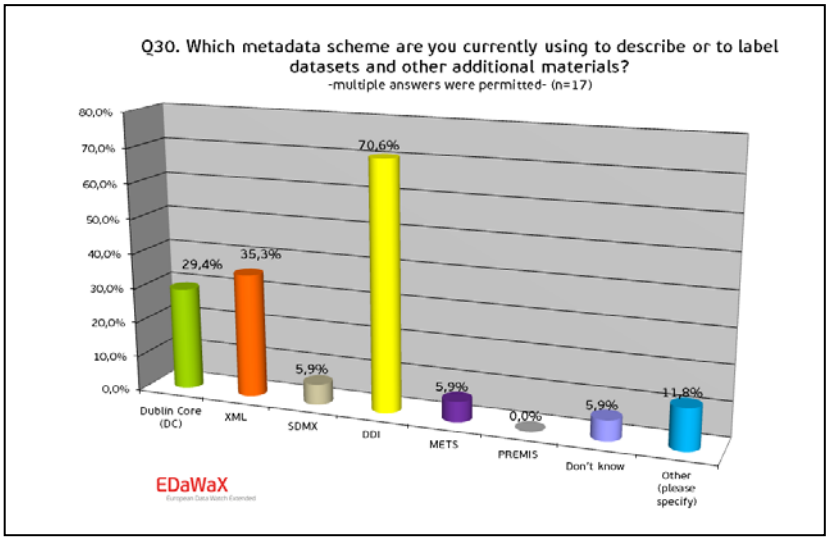
Most frequently APIs were mentioned as a device for data search (47%), followed by APIs used for uploading research data. Slightly more than a third (35%) of all respondents declared to have an API at their disposal to analyse research data.

Further analysis by EDaWaX showed, however, that the reported interface usage consists of searching and uploading interfaces on the respondents' websites only. We were not able to find an API. Presumably, APIs in terms of external reading and writing accesses are by and large unknown among our respondents and not available so far.

Metadata schema and the creation of metadata

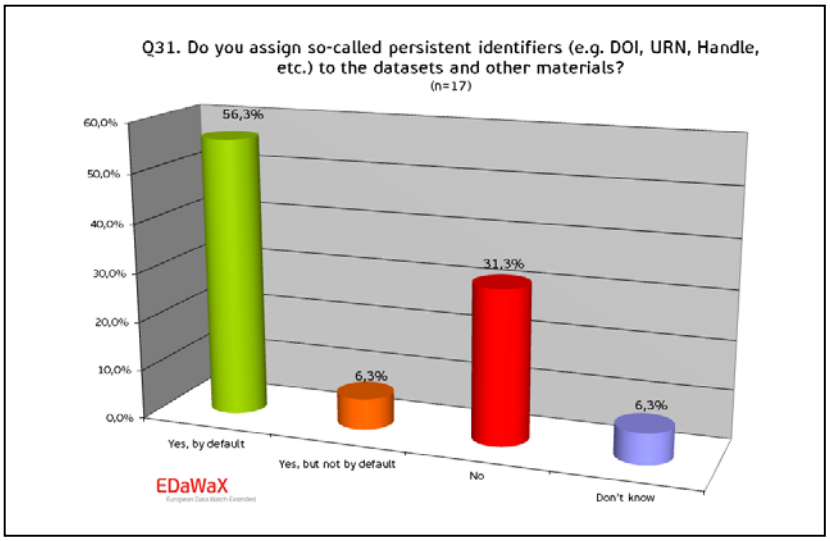
Employed metadata schemata

We were also interested in the metadata schemata currently used by the organisations in their daily work. Our survey shows that more than 70% of the respondents are using DDI. Other like XML or Dublin Core are being used quite rarely (35% and 29%). All other metadata schemata were used rather sporadically.



Persistent Identifiers (PI)

In addition, we asked, whether organisations are assigning persistent identifiers (e.g. handle, DOI, URN, etc…) to datasets and other materials. The persistent identification of research data is an important issue, for instance because it enables researchers to cite datasets. Organisations in our sample are assigning such identifiers in more than 56% per default, but almost a third is not.

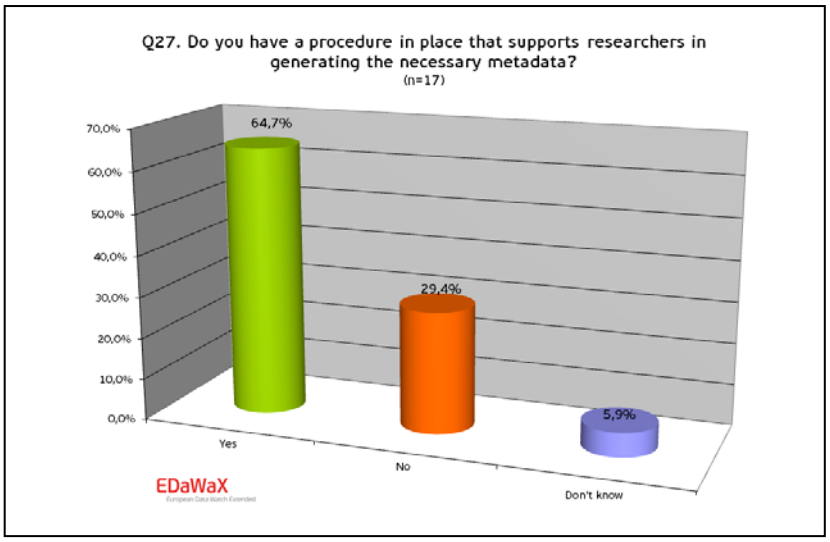


Support of Semantic Web Technologies

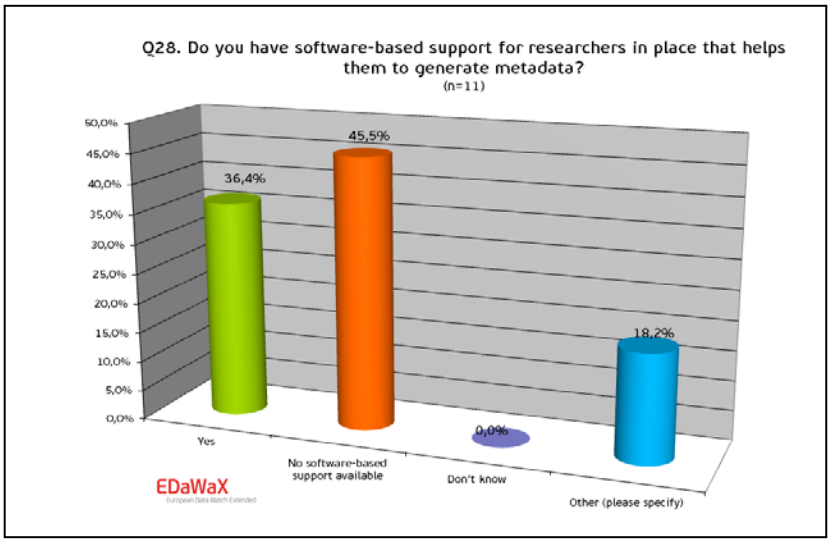
In our survey we also examined the implementation of RDF (Resource Description Framework). RDF is a general method for conceptual description or modelling of information implemented in web resources. Among the organisations answering this question only a minority of 6% stated to use and disseminate RDF-files. Almost a quarter of all respondents was not able to specify, whether their organisation is using RDF, which presumably indicates that RDF is largely unknown.

Support for creating metadata

Again and again, a critical issue regarding the reuse of research data is the quality of data documentation. Therefore, a matter of particular interest was to find out, whether and if how respondents support researchers in generating metadata.



Our survey shows, that the majority (almost 65%) of all organisations does so.

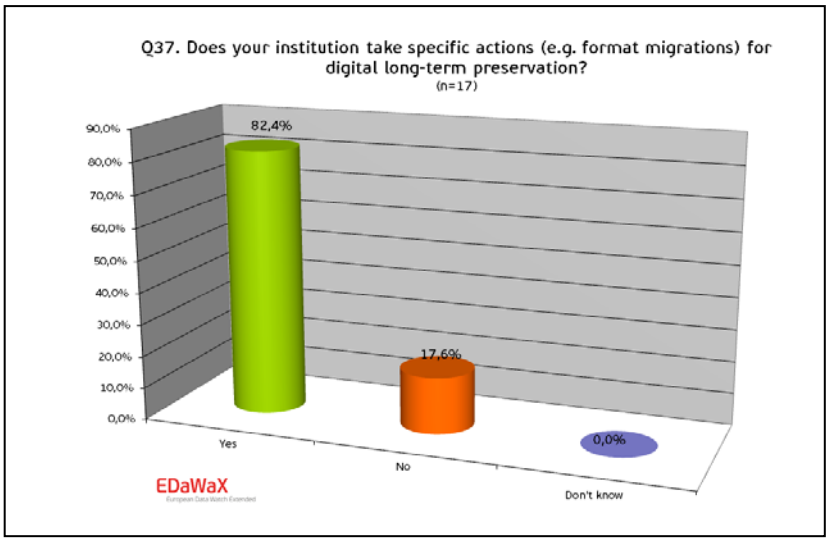


Furthermore, we were keen to know, whether this support is software-based – e.g., if there is a web frontend where researchers may type in the required information that is converted into a standardised metadata schema. We found more than 35% of the respondents to have such a software-based support for researchers in operation.There is a striking number of statements in the section *other*. Part of *other* support for researchers, for instance, consists of written *data deposit forms*. Our question regarding the software's names revealed that at least two institutions are using Nesstar.[12] Many organisations are also applying in-house developments.

[12] Website of nesstar, www.nesstar.com

Digital long-term preservation

In our survey we eventually wanted to identify to which extend the respondents' institutions have implemented specific measures for long-term preservation of research data. Our survey indicates, that more than 80% of all organisations have adopted procedures in this direction.



Conclusion

Our results show, that research data centres might be relevant places for hosting and storing publication-related research data, because they are already fulfilling many pre-requirements. Nevertheless, among the responding organisations there seems to be no institution, which is currently complying with all requirements with regard to storing and hosting publication-related research data.

In detail the outcome of our survey is:

- Almost three-fourths of all organisations in our sample is generally accepting external datasets – including publication-related research data. However, partial limitations exist – for instance, because of regional or subject-specific competence or because of the dataset's quality.

- Almost the same percentage (75%) of our sample is principally accepting the code of computation for storing and hosting. If (self-compiled) software is used for obtaining empirical results within an empirical research paper, only a minority of 40% will accept these data for storing and hosting.

- DDI is the most common metadata schema currently in use among our respondents (70%). XML and Dublin Core are following with shares of 35% and 30% respectively (multiple answers were permitted). Almost two thirds are using persistent identifiers for their datasets and, thereby, are facilitating citations of the data. Approximately three-fourths of all organisations support researchers in generating metadata for datasets though.

- Interfaces (APIs) for searching, analysing or uploading datasets and other materials currently do not seem to be available yet. Also the use of RDF is little popular among the responding organisations.

- Digital long-term preservation is wide-spread among our respondents. More than 80% reported that their institution takes measures for ensuring the long-term availability of their digital holdings.

References

Anderson, R. / Greene, W. H. / McCullough, B. D. / Vinod, H. D. (2008). The Role of Data/Code Archives in the Future of Economic Research. In: *Journal of Economic Methodology*, 15(1), 99-119

Andreoli-Versbach P. / Mueller-Langer, F. (2013). Open Access to Data: An Ideal Professed but not Practised, *RatSWD Working Paper Series*, No. 215, Berlin. Available on: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2224146

De Cock Bruning, M. / van Dither, B. / Jeppersen de Boer, C.G. / Ringnalda, A. (2011): The legal status of research data in the Knowledge Exchange partner countries. Available on: http://www.knowledge-exchange.info/Default.aspx?ID=461

Häder, M. (2009): Der Datenschutz in den Sozialwissenschaften. Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland. RatSWD Working Paper Series, No. 90, Berlin. Available on: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf

Hillegeist, T. (2012): Rechtliche Probleme der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten, Göttinger Schriften zur Internetforschung (8). Available on: http://webdoc.sub.gwdg.de/univerlag/2012/GSI8_Hillegeist.pdf

King, G. (1995). Replication, replication. In: *PS: Political Science and Politics*, 28, 443–499. Available on: http://gking.harvard.edu/gking/files/replication.pdf

McCullough, B.D. (2009): Open Access Economics Journals and the Market for Reproducible Economic Research. In: *Economic Analysis and Policy*, 39 (1), 117-126

Research Information Network (2011). Data centres: their use, value and impact. A Research Information Network report. September 2011. Available on: http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf

Spindler, G. / Hillegeist T. (2008): KoLaWiss-Gutachten AP 4: Recht, Rechtsexpertise für das Projekt „Kooperative Langzeitarchivierung an Wissenschaftsstandorten“ (KoLaWiss). Available on: http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf

Siegert, O./ Toepfer R./ Vlaeminck, S. (2012). Forschungsdatenmanagement in den Wirtschaftswissenschaften – Ausgewählte Dienste und Projekte der Deutschen Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft (ZBW). In: R. Altenhöner / C. Oellers (Hrsg.): *Langzeitarchivierung von Forschungsdaten - Standards und disziplinspezifische Lösungen*, Berlin, Scivero Publishing.

Vlaeminck, S. (2013). “Data Management in Scholarly Journals and possible Roles for Libraries – Some Insights from EDaWaX.” *LIBER Quarterly*, 23 (1). Available on: http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-114595

Vlaeminck, S. / Siegert, O. (2012). “Welche Rolle spielen Forschungsdaten eigentlich für Fachzeitschriften? Eine Analyse mit Fokus auf die Wirtschaftswissenschaften.” *RatSWD Working Papers*, No. 210, Berlin. Available on: http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_210.pdf

Vlaeminck, S. / Wagner, G. G. / Wagner, J. / Harhoff, D., Siegert, O. (2013). Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen. In: *LIBREAS. Library Ideas*, 23: Forschungsdaten. Metadaten. Noch mehr Daten. Forschungsdatenmanagement. Available on: http://edoc.hu-berlin.de/libreas/23/vlaeminck-sven-1/PDF/vlaeminck.pdf (urn:nbn:de:kobv:11-100212694)