Evaluation of the EDaWaX Online Survey on Hosting Options for publication-related Research Data

Background

The online-survey conducted by the EDaWaX project (European Data Watch Extended[1]) aims to evaluate the services of research data centres (in particular research data centres accredited by the German Data Forum (RatSWD)[2] and the CESSDA[3] research data centres), archives, library networks and single libraries in regard to the general possibility to store and host publication-related research data.

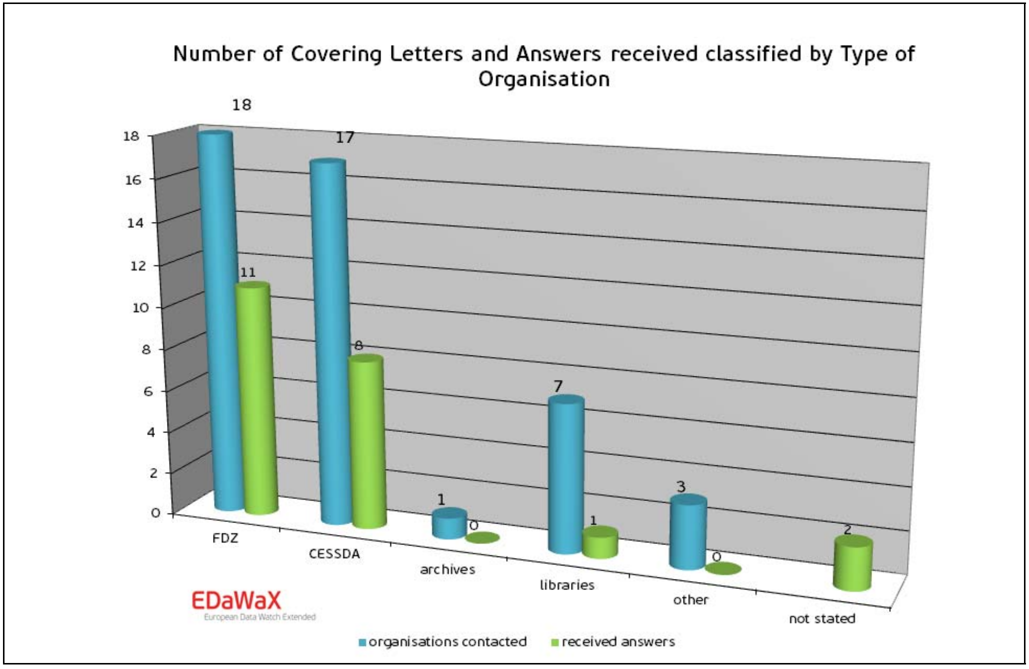
The implementation of such a publication-related research data archive is one of the main objectives of EDaWaX.

The Online-Survey

The online-questionnaire was send to 46 organisations in October and November 2012. Among the recipients were 36 national and international research data centres, 1 archive, 7 library networks and single libraries and three other organisations (non-European research data centres). 22 organisations participated in our survey (47.8%). The return rate can be considered as very good, compared to the average return rate of surveys in written form.

Due to the structure of the questionnaire not all organisations responded any question. Differences in the number of respondents (among other things) are explainable thereby.

Certainly more important than the return rate is the structure of the respondents and non-respondents. For our survey the big majority of all responds are working in research data centres in Germany and Europe (86.4%). Clearly underrepresented were respondents from German library networks and archives, but also three research data centres from non-European areas did not respond.



We can only suppose that the library networks and the archive to not own appropriate services or offers for the management of research data and therefore these organisations did not participate.

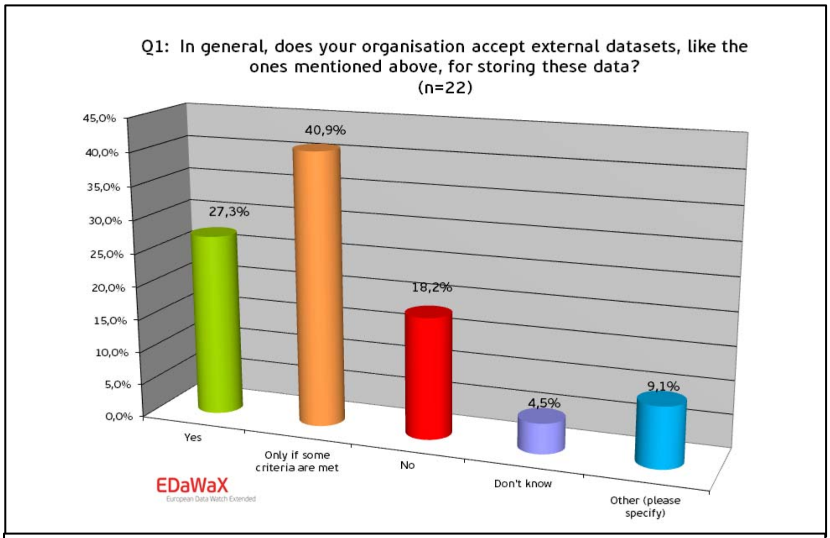
[1] European Data Watch Extended Project: http://www.edawax.de (in English)

[2] German Data Forum: http://www.ratswd.de/eng/index.html

[3] Council of the European Social Science Data Archives: http://www.cessda.org

Interpretation of our Findings

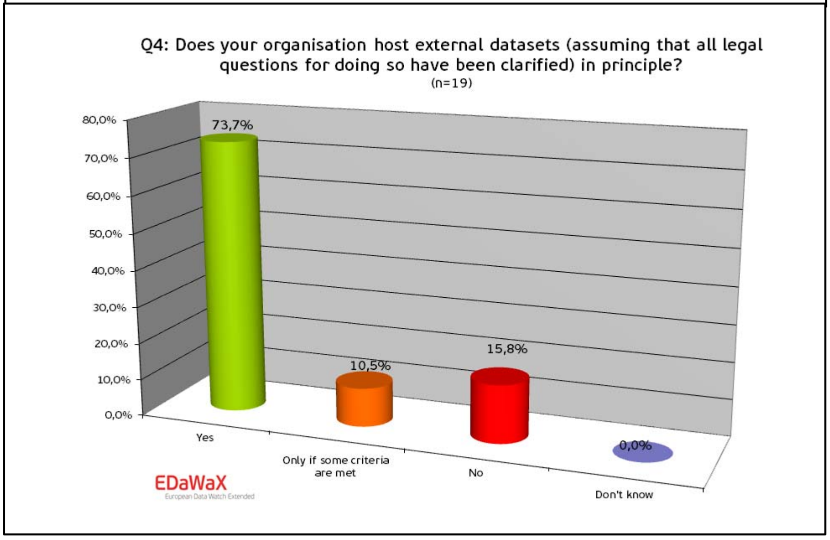
Initially all surveyed were asked, whether their institutions host and store publication-related research data in general. In addition, they were asked whether their organisations also host and store (self-written) software components and the code of computation of statistical analyses. All of these types of data often are part of empirical submissions to economics journals.


Q1: In general, does your organisation accept external datasets, like the ones mentioned above, for storing these data?
(n=22)

Datasets

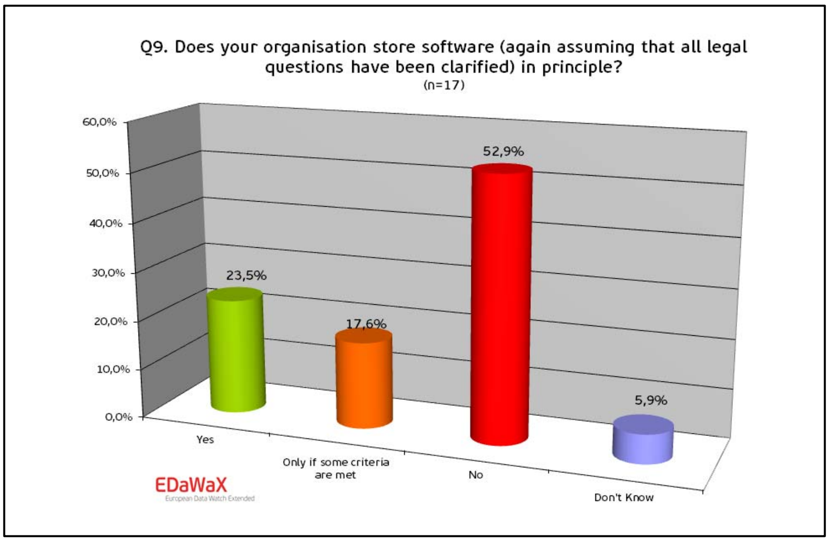
Of all organisations evaluated more than three-fourths generally accept external datasets for storing. The largest portion of respondents reported that these types of data only are accepted if certain criteria are met. Such criteria consist in form of subject specific competencies of many research data centres, but also in form of regional/supra-regional or national competencies. Besides, technical and organisational aspects (e.g. proper documentation, machine-readability…) and judicial questions were mentioned.

Approximately 74% of the respondents indicated, that their organisations also host these types of data. If some criteria for hosting were mentioned again the subject-specific orientation of an institution was stated as criterion.


Q4: Does your organisation host external datasets (assuming that all legal questions for doing so have been clarified) in principle?
(n=19)

Software

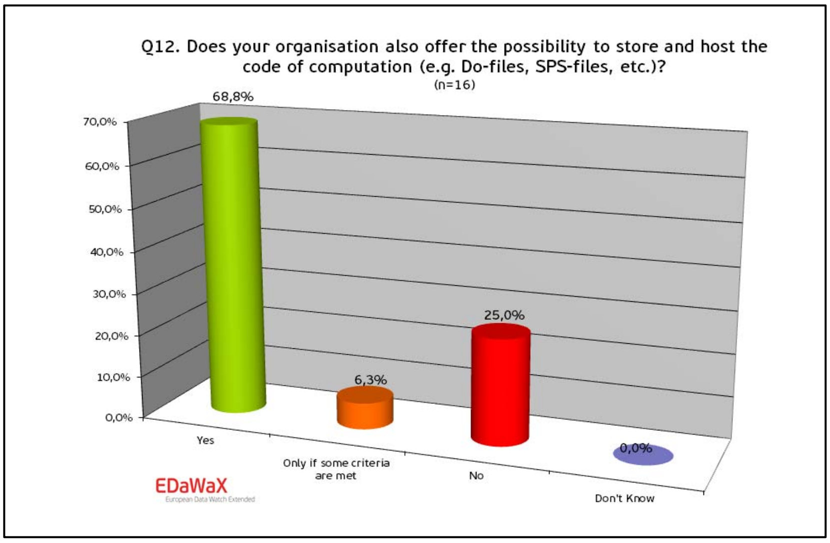
In regard to storing and hosting of (self-written) software components that are often used for the purpose of economics simulations, our survey shows that only a minority of almost a fourth accepts storing and hosting software components without restrictions. Another 17% pointed out that criteria exist (e.g. *essential for the analysis of the data*) for assessing if software could be stored and hosted. Therefore hosting and


Q9. Does your organisation store software (again assuming that all legal questions have been clarified) in principle?
(n=17)

storing software components can be considered as a gap. Only a limited number of organisations offer this service.

Code of Computation

Almost 70% of the organisations analysed offer the possibility to store and host the code of computation. One fourth of all organisations though does not and is not plan to do so in the near future. One respondent also stated a criterion – he mentioned that storing and hosting of these data are only useful in case of derived variables.



APIs

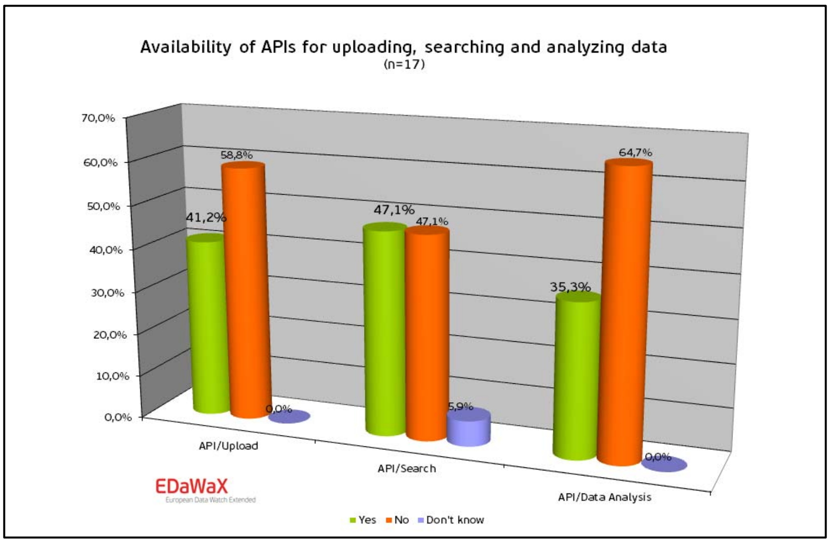
In the course of our analyses we also asked for the availability of application programming interfaces (APIs). With these APIs automated exchanges of data are enabled.

Our results show that less than half of all organisations reported to have those interfaces at their disposal.

Most frequently APIs for searching data were mentioned (47%), followed by APIs for uploading research data. Slightly more than a third (35%) of all respondents declared to own an API for analysing purposes.

Further analysis of these APIs surprisingly showed that the interfaces reported consist of searching and uploading interfaces on the respondents' websites only. We were not able to find an API.

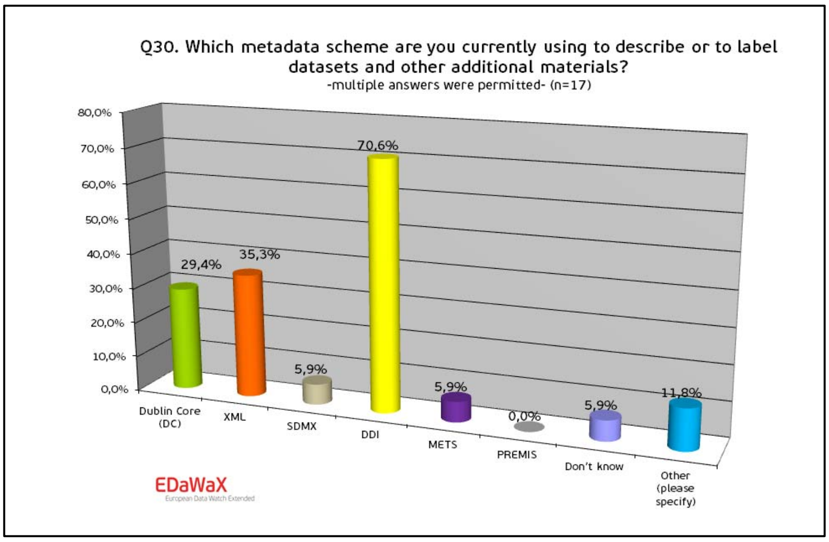
It can therefore be assumed that APIs are not well known among the respondents and are currently not available yet.



Metadata Schemata and Creation of Metadata

Metadata Schemata in Use

We were also interested in the metadata schemata that are currently used by the organisations for their daily work. Our survey shows that more than 70% of the



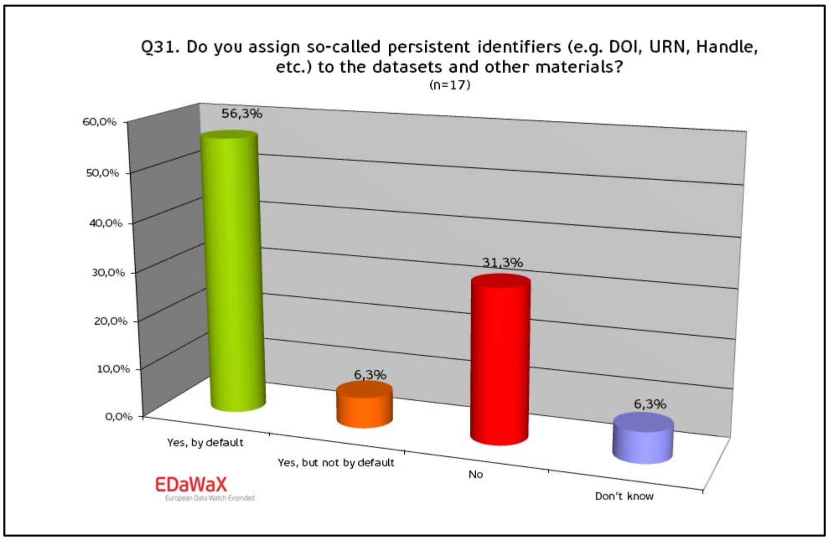
respondents used DDI. Other schemata like XML or Dublin Core were used considerably less (35% and 29%). All other metadata schemata were used sporadically only.

Persistent Identifiers (PI)

In addition the respondents were asked whether their organisations assign 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 assigned such identifiers in more than 56% per default, but almost a third does not.



Support of Sematic Web Technologies

In our survey we also asked for the implementation of RDF (Resource Description Framework). RDF is a general method for conceptual description or modelling of information that is implemented in web resources. Of all organisations that answered this question only a minority of 6% stated to use and disseminate RDF-files. Almost a quarter of all respondents did not know whether their organization uses RDF, what probably indicates that RDF is little-known.

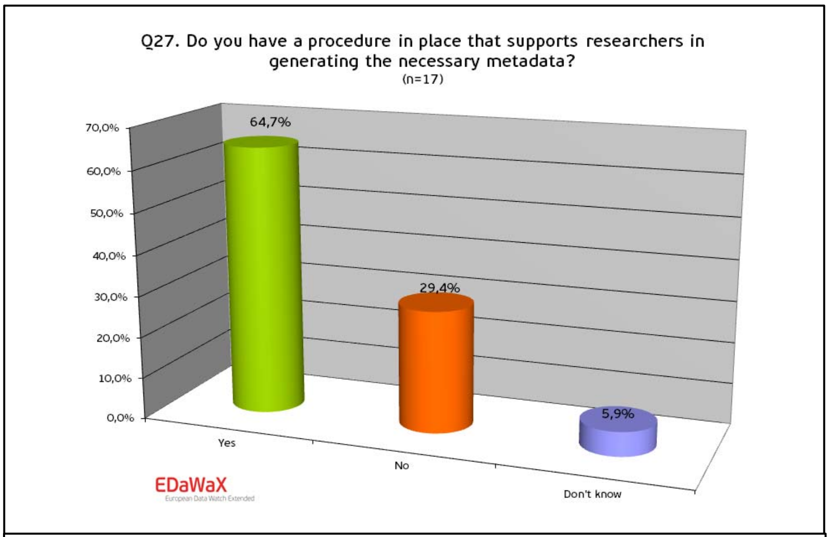
Support for Metadata Creation

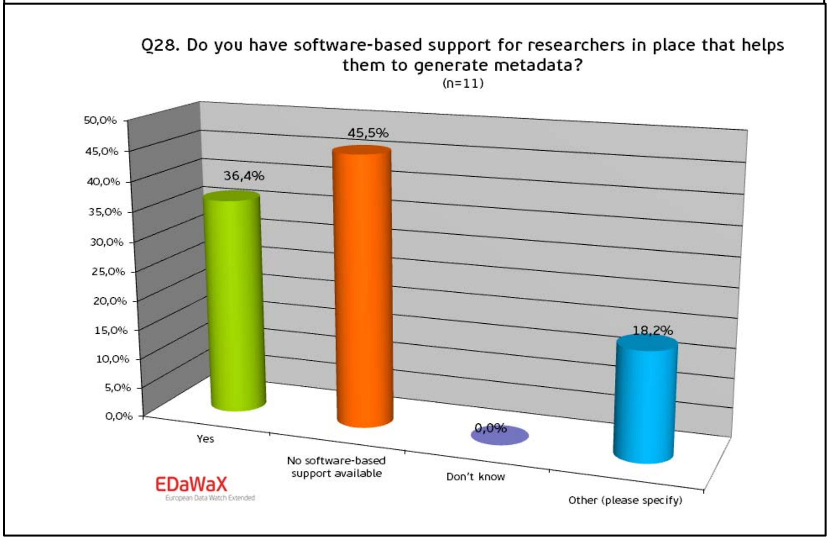
A very important topic for reusing research data often is the quality of the data's documentation. Therefore it was of note to know if and how the respondents support researchers in generating metadata.

Our survey shows that the majority (almost 65%) of all organisations support researchers in creating metadata.

In addition we wanted to know whether this support is software-based – e.g. if there is a web frontend where researchers can type in the required information that is converted into a standardized metadata schema.

For software-based support we see that more than 35% of the respondents have such a

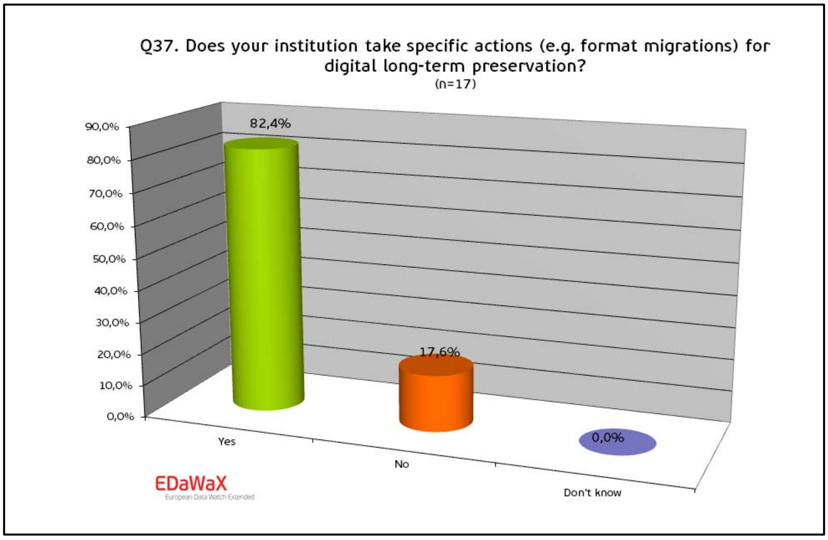




software-based support for researchers in place.

Noticeable is the number of statements in the section *other*. Part of *other* support for researchers for instance consists of *data deposit forms* in written form.

Our request for the software's names showed that at least two institutions used Nesstar.[4] Many organisations also used in-house developments.


Q37. Does your institution take specific actions (e.g. format migrations) for digital long-term preservation?
(n=17)

Digital long-term Preservation

We also wanted to know to which extend the respondent's institutions implemented specific actions for the long-term preservation of research data. Our survey shows that more than 80% of all organisations have adopted corresponding arrangements.

Conclusion

Our results show that research data centres are relevant places for hosting and storing publication-related research data, because they already fulfil many requirements for doing so. Nevertheless among the responding organisations currently there seems to be no institution that entirely complies with all requirements in 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 accept external datasets in principle – including publication-related research data. However partial limitations exist - for instance because of regional or subject-specific responsibilities or because of the dataset's quality.
- Almost the same percentage (75%) of organisations accepts the code of computation for storing and hosting in principle. If (self-written) software was used for obtaining empirical results claimed in an empirical article only a minority of 40% accepts these data for storing and hosting.
- DDI is the most common metadata schema among our respondents currently in use (70%). XML and Dublin Core follow with 35% respectively 30% (multiple answers were permitted). Almost two thirds used persistent identifiers for their datasets and thereby facilitate citations. Approximately three-fourths of all organisations though support researchers in generating metadata for datasets.
- Interfaces (APIs) for searching, analysing or uploading datasets and other materials currently doesn't 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 actions for ensuring the long-term availability of their digital holdings.

[4] www.nesstar.com