Insight into digital preservation of research output in Europe

survey report

Project Number	223758
Project Title	PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Survey report
Deliverable Number	D3.4
Contributing Work package	WP3: Community Insight
Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31 August 2009 (M18)
Actual Delivery Date	9 December 2009 (M22)
Author(s)	Tom Kuipers, Jeffrey van der Hoeven

PARSE.Insight (INSIGHT into issues of Permanent Access to the Records of Science in Europe) is a two-year project co-funded by the European Commission under the Seventh Framework Programme (FP7), Research Infrastructures. This document is created by PARSE.Insight and is deliverable D3.4 Survey Report.

December 2009, PARSE.Insight

Abstract

This report (deliverable 3.4 of PARSE.Insight) describes the results of the surveys conducted by PARSE.Insight to gain insight into research in Europe. Major surveys were held within three stakeholder domains: research, publishing and data management. In total, almost 2,000 people responded; they provided us with interesting insights in the current state of affairs in digital preservation of digital research data (including publications), the outlook of data preservation, data sharing, roles & responsibilities of stakeholders in research and funding of research.

Keyword list

Insight, preservation, survey, questionnaire, results, research, publishing, data management, funding, discipline, data, sharing, publication

Contributors

Person	Role	Partner	Contribution
Tom Kuipers	Author	KB	Document owner and author
Jeffrey van der Hoeven	Co-author	KB	Contributing author
Eefke Smit	Reviewer	STM	Review of section publishing
Simon Lambert	Reviewer	STFC	Review of section research
Barbara Sierman	Reviewer	KB	Review of all sections

Document Approval

Person	Role	Partner
Simon Lambert	Reviewer	STFC
Eefke Smit	Reviewer	STM

Revision History

Issue	Author	Date	Description
0.1	Tom Kuipers	01-02-2009	Initial version
0.2	Jeffrey van der Hoeven	22-05-2009	Internal review
0.3	Tom Kuipers	29-07-2009	Changed structure and improved illustrations
0.4	Jeffrey van der Hoeven	30-07-2009	Added text and selected diagrams
0.5	Tom Kuipers	14-08-2009	Improved section research
0.6	Tom Kuipers	13-10-2009	Finalised section research + start publishing
0.7	Tom Kuipers	28-10-2009	Incorporated comments reviewers
0.8	Jeffrey van der Hoeven	24-11-2009	Rerun some graphs + integrated all sections
0.9	Tom Kuipers Jeffrey van der Hoeven	30-11-2009	Major update on all sections + update all graphs
1.0	Tom Kuipers Jeffrey van der Hoeven	09-12-2009	Final version

1 Executive Summary

PARSE.Insight is a two-year project co-funded by the European Commission under the Seventh Framework Programme (FP7), Research Infrastructures. It is concerned with the preservation of digital information in research, from primary data through analysis to the final publications resulting from this research. Ultimately, the PARSE.Insight project will develop a roadmap for an e-Science infrastructure, intended to guide the European Commission's strategy about research infrastructures.

To define the needs for an e-Science infrastructure for long-term availability of research data, a number of surveys were conducted to gather information on the practices, ideas, and needs. The surveys were targeted at four stakeholders we identified as key figures in the research communities: research, data managers, publishers, and funders.

This report describes the results gained via these surveys. In total 1,840 people responded: 1,389 responses on the researchers' survey; the data managers surveys yielded 273 responses, and the 178 publishers started the publishers survey. The funders survey gained only a few responses. In-depth analysis of this stakeholder will be done in follow-up research by PARSE.Insight.

1.1 Key Findings

1.1.1 Researchers

- Researchers consider the *possibility of re-analysis of existing data* as the most important driver for the preservation of research data; 91% of the respondents thought this to be either important or very important.
- Researchers regard *the lack of sustainable hardware, software or support of computer environment may make the information inaccessible* as the most important threat to digital preservation. 80% believe this to be either important or very important.
- 58% of the research respondents believe that an *international infrastructure for data preservation and access* should be built to help guard against some of the above-mentioned threats.
- 25% of the researcher make their data openly available for everyone.
- Major barriers for sharing research data are *the fear of researchers regarding legal issues* and *the misuse of their data*.

1.1.2 Data Managers

- Data managers think that public funding is the most important reason for data preservation. 98% of the respondents think that *if research is publicly funded, the results should become public property and therefore properly preserved.*

- Data managers also regard *the lack of sustainable hardware, software or support of computer environment may make the information inaccessible* as the most important threat to digital preservation. 86% believe this to be either important or very important.
- 60% of the respondents to the data managers' survey believe that *an international infrastructure for data preservation and access* should be built to help guard against some of the above-mentioned threats.
- 59% of the respondents to the data managers' survey don't think that the tools and infrastructure available to them will suffice for the digital preservation objectives they have to achieve.
- 71% of the respondents to the data managers' survey believe that funding for preservation will be an issue now and in five years time.

1.1.3 Publishers

- The most important reason for preservation marked by publishers is that *it will stimulate the advancement of science*. 96% of the small and large publishers regarded this either important or very important.
- Regarding threats to preservation, 78% of the small publishers fear *the sustainability of data when the current custodian of the data ceases to exist in the future*. For large publishers this percentage is even 80%.
- Both small (75%) and large (64%) publishers think that an international infrastructure for data preservation and access should be built.
- Of the respondents to the publishers' survey, 55% of the small publishers stated to have a preservation policy in place compared to 84% of the large publishers.
- However, both 69% of the respondents of large and small publishers stated that they have no preservation arrangement in place for underlying research data.
- 71% of the large publishers stated that authors can submit underlying research data with their publication. For small publishers only 57% stated to accept it.
- Large publishers often have a preservation strategy (e.g. normalisation, outsourced preservation service, emulation) in place while 28% of the small publishers stated not to have such a strategy in place.
- The majority of the respondents to the publishers' survey stated that publishers are responsible for the preservation of publications (73% of the small publishers, 69% of the large publishers), but that the author is responsible for the underlying data.
- Regarding current movements in scholarly communications, publishers stated that *a hybrid model, combining subscription-based journals and open access journals, while the journal model remains dominant* is the most likely scenario for the future (32% of small publishers, 43% of large publishers).

2 Acknowledgements

This research would not have been possible without the generous help of several organisations and individuals.

In special, we would like to thank the following people: Henk Voorbij (KB), Inge Angevaare (NCDD), Barbara Sierman (KB), Hans Krabbendam (RSC) and Wouter Schallier (LIBER) who as reviewers outside the consortium helped us compose and analyse the surveys.

We would also like to thank the following organisations that helped us with the distribution of the surveys:

All European Academies (ALLEA)
Alliance for Permanent Access
CASPAR project
Digital Curation Centre (DCC)
Digital Humanities Observatory (DHO)
Digital Preservation Europe (DPE)
Digital Preservation Coalition (DPC)
Directory of Open Access Journals (DOAJ)
D-Lib Magazine
EURODOC
European Science Foundation (ESF)
International Association of STM Publishers
LIBER
Marie Curie Fellowships Association (MCFA)
Max Planck Institute (MPI)
Nationale Coalitie Digitale Duurzaamheid (NCDD)
Reed Elsevier
Young European Associated Researchers (YEAR)

On behalf of the PARSE.Insight project,

Tom Kuipers (KB)
Jeffrey van der Hoeven (KB)

3 Table of Contents

4 Introduction

The growing multitude of digital resources forms the basis of the intellectual capital of European research. Retrieving information from these resources and allowing new generations of researchers to "stand on the shoulders of giants" is the very essence of research. These digital resources must persist and remain findable, accessible, and understandable. Data re-use (by users in a different discipline, for example) may happen immediately when the data is produced or may not happen for an extended period of time. There is a very real risk that much of the research data and documentation that exist today may be lost to future generations unless permanent access is secured.

The European project PARSE.Insight focuses on the infrastructure needed to support persistence and the ability to understand these key assets over the long term.

4.1 About PARSE.Insight

PARSE.Insight is a two-year project co-funded by the European Union under the Seventh Framework Programme. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research.

Many initiatives are already under way in this area, and the aim of the PARSE.Insight project is to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe.

With surveys, case studies, desk research and interviews we aim to gain insight into the practices, needs and requirements of research communities. A gap analysis is performed to measure the gaps between the today's practices and the future ideal. The roadmap, the ultimate product of PARSE.Insight, is intended to guide the European Commission's strategy about research infrastructure.

PARSE.Insight is closely linked to the Alliance for Permanent Access to the Records of Science[1].

4.2 About this report

To define the needs for an e-Science infrastructure for long-term availability of research data, a better understanding is needed of the current and future challenges in that field. Therefore, we developed and distributed a number of surveys to gather information on the practices, ideas, and needs of research communities regarding the preservation of digital research data. In this report digital preservation denotes digitally encoded objects are specifically curated to be re-usable in the long term.

[1] Alliance website: http://www.alliancepermanentaccess.eu

Over the past years, several national and international surveys have been conducted that cover much of the digital preservation territory, ranging from the "Mind the Gap" survey of preservation readiness in the UK to the e-IRG request for input. Although the outcomes of these surveys are very useful, each of them form only a small piece of the larger puzzle. We studied the existing reports and felt the need to conduct a new survey that is more targeted to the various communities and stakeholders in research across Europe.

4.3 Objectives

The main objective of the survey is to provide information that is needed to perform a gap analysis and refine the roadmap. For this we need to gather information on:

- perceptions of importance of preservation and the funding expected to be available and factors influencing those decisions;
- information about what is in place in terms of policies tools and services both for preservation of publications as for preservation of data and the links in between;
- ideas and evidence of cross-disciplinary use;
- plans for new policies, tools and processes;
- expectations about roles and responsibilities.

4.4 Scope

This report focuses on research communities in a broad sense. It includes all major research disciplines, senior and junior researchers, major research institutes, research schools, and universities. We did not specifically target applied research as a separate community. Yet since boundaries are fluid, the results do include applied research as well to some extent, especially in the Technology disciplines.

Of course research communities include more players than researchers alone. So, besides the researchers, we identified data managers, publishers and funders as major stakeholders of the research communities.

4.5 Terminology

4.5.1 Survey

In this report a *survey* denotes a method of gathering information from a sample of individuals. The sample represents the total population being studied. There are different methods for survey data collection—telephone interviews, in-person interviews, mail—but in this document survey means an online questionnaire.

4.5.2 Digital research data

In PARSE.Insight the term *digital research data* is used for all output in research. In practical terms, raw data, processed data and publications are all covered by the same term. A distinction

between these sorts of research data is only made when necessary (for example when policies for publications are compared with other data).

4.5.3 Digital preservation

Digital preservation denotes the process of storing digital information in such a way that it remains accessible, understandable and usable over the long term (usually 5, 10, 50 or more years). This means that data needs to be specifically curated and enriched with extra information (metadata). For example: where did the data come from? How have they been stored? Which file formats have been used? What special terminology or other information is needed to interpret and use the data?

4.6 Target groups

This report is a public deliverable of the PARSE.Insight project on the issues of permanent access to the records of science in Europe. The target group of this report are people with a stake in the issues surrounding access to research data, especially (but not limited to) the stakeholders we identified.

4.7 The structure of the document

Following the Introduction and the Appendix, the document is divided into five chapters. Chapter 5 contains a discussion of the method applied in the surveys. Chapters 6 through 8 provide an analysis of the research results for three stakeholders: researchers, data managers, and publishers. All chapters are structured in a similar way. Each starts with an introduction that discusses some of the general facts of the survey. These facts include the number of responses and distribution channels used. Next, the important questions (highlights) of the survey for the specific stakeholder are dealt with. Finally, attention is paid to the impact of these results on the PARSE.Insight roadmap.

The analysis part of each of these chapters is subdivided into 6 themes based on the objectives discussed in 4.3. The themes are:

- Perceptions of preservation
- Preservation – the state of affairs
- Preservation – the outlook
- The cross-disciplinary use of research data
- Funding
- Roles and responsibilities

At the end of the report, the section on conclusions brings all separate analyses together and weighs the needs, ideas and practices of the stakeholders against each other.

4.8 Data set and graphics

The data set used to write this report and all graphs are available at our public website: www.parse-insight.eu

5 Method

5.1 Workflow

The method by which we conducted research has four components:

- Desk research
- Surveys
- Interviews
- Case studies

This chapter only records the method for the survey. The other three strategies will be described in more detail in the final *Insight Report* (deliverable D3.6 of PARSE.Insight, to be expected in early 2010).

5.2 Stakeholders & Distribution

The PARSE.Insight surveys aimed at (European) stakeholders in research. This encompasses stakeholders from all member states of the European Union (see appendix 2) and all disciplines. While our main focus was on Europe, we did not exclude responses from outside the EU.

As stated before, research communities encompass more than researchers and research institutes. Research involves a number of actors working individually or in groups, but who quite often have different—sometimes conflicting—agendas. This has to be taken into account when trying to map the practices, knowledge and needs of research communities regarding digital preservation.

We recognised four major stakeholders in research (see Figure 1):

- Researchers
- Data Managers (data centres, digital archives, etc.)
- Publishers
- Funders (national and European)

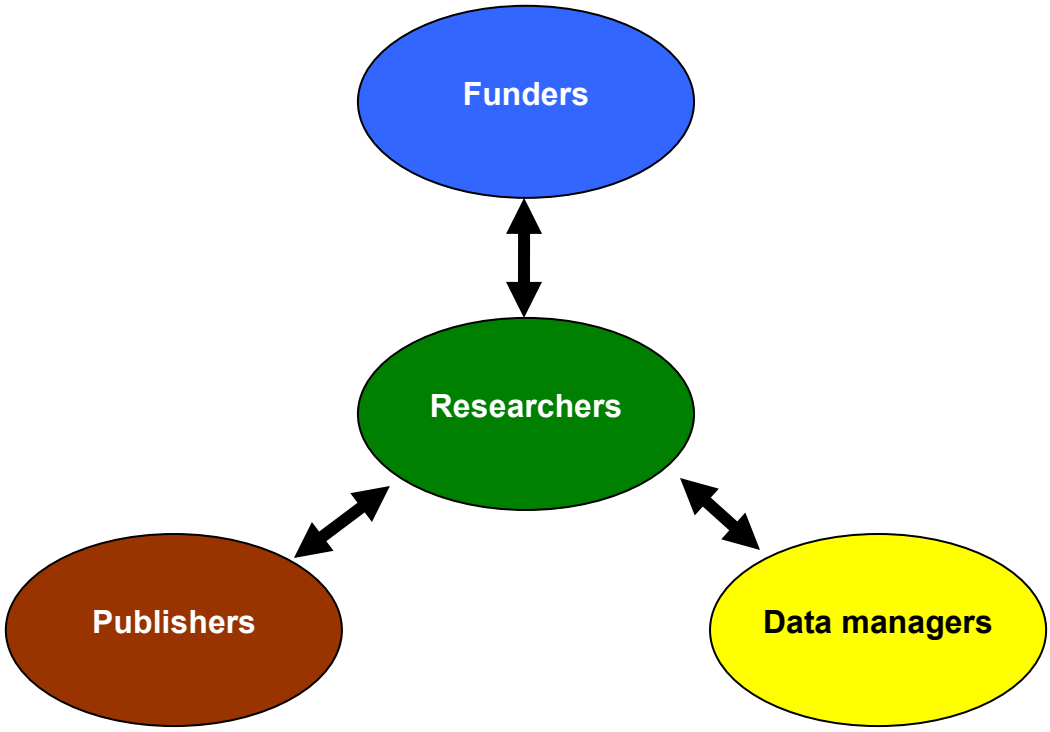


Figure 1: generalised view on stakeholders in research

Separate surveys were developed for each stakeholder, sharing a common core of questions, because we wanted to ask stakeholder-specific questions next to more general questions that appeared in all surveys. Thus, four surveys were developed, which were sent out through a number of different distribution channels.

Surveys were sent separately through stakeholder-specific distribution channels as well as combined into one merged survey. The merged surveys were mostly sent to mailing lists which could attract responses from more than one type of stakeholder. In the merged survey a question was added about the role of the respondent, so that the respondents only had to answer questions relevant to themselves. Built-in skip logic made sure that respondents would only see the relevant questions.[2]

Due to the relatively small number of responses we received from funders their responses are not analysed in this report. They will be taken into account—together with the interviews that will be conducted—in the *Insight Report (D3.6)*.

For the distribution of the surveys, the collection of responses, and the (basic) analysis of data we used Survey Monkey,[3] a web application that enabled us to design and distribute the surveys as well as to collect and analyse the responses.[4]

[2] For an overview of all distribution channels, see the introductions of the stakeholder-specific sections in this report.

[3] SurveyMonkey: http://www.surveymonkey.com

[4] PARSE.Insight *Deliverable 3.1: Survey and Forum platforms* explains and justifies why we choose Survey Monkey as the survey tool for our online questionnaires. See http://www.parse-insight.eu/publications.php

The research survey was distributed through different channels to a large number of researchers from all disciplines and all European Union member states[5] (and beyond). When we noticed a lack of responses from certain disciplines or countries, we tried to locate new distribution channels that specifically targeted group(s) of people whose responses were lacking.

5.3 Research disciplines

Due to the number of research disciplines, it would have been impossible to analyse the data separately for each discipline. Also the boundaries between disciplines are amorphous; research can often not be tied down to one specific discipline. Therefore we categorised the research disciplines into a number of main categories based on the categorisation that the KNAW (Royal Netherlands Academy of Arts and Sciences) uses for their online Dutch research database.[6] It recognises nine main categories or groups of disciplines:

- Agriculture & Nutrition
- Behavioural Sciences
- Humanities
- Life Sciences
- Medicine
- Social Sciences
- Physical Sciences
- Socio-Cultural Sciences
- Technology

Each main category has a number of subcategories (or disciplines falling under the main category). For instance Agriculture & Nutrition is subdivided into:

- Agricultural technology
- Agriculture and Horticulture
- Animal feeds
- Animal husbandry
- Fisheries
- Forestry
- Foods and stimulants
- Nutrition

Sometimes these categories are subdivided one level further.[7].

[5] For a detailed description of the distribution channels employed and the response rates, see chapters 6 through 8.
[6] KNAW database: http://www.onderzoekinformatie.nl/en/oi/nod/
[7] See appendix 1 for a visual representation in the form of a tree.

5.4 Validity of results

This section deals with issues of validity and representation. These issues are dealt with separately for each stakeholder survey – research, data management, and publishing.

5.4.1 Research survey

The total number of researchers in Europe is estimated at 1.33 million[8]. As it is impossible to reach the whole population we defined a random sample. Based on statistical measures, a minimum of 385 responses is needed to give an adequate representation of Europe's research community (with confidence interval = 95%, bias = 5%)[9].

In total the surveys elicited 1,389 responses of which 609 from EU member states. We checked the results for possible biases. For instance, certain groups may be overrepresented. In trying to identify alleged biases we found two instances of overrepresentation that may influence the results, one related to disciplines and one related to experience. Compared to other disciplines there were significantly more responses from the physical sciences (33% of total responses). In addition (senior) researchers (70%) with more than 20 years of research experience are much better represented in the surveys' results than researchers with less than 20 year of experience.

To examine whether the large number of physical scientists introduced a bias in our results we looked at the normal distribution of research disciplines in Europe[10] and compared this with the distribution of the discipline in our survey results. From this we learned that there **are** significantly more researchers in physical sciences in Europe than other disciplines. This is discussed in more detail in chapter 6.

The second possible bias we investigated is perhaps less obvious, but comparing the results of experienced researchers to the results of non-experienced researchers did not reveal major discrepancies. This will also be explained in more detail in chapter 6.

This strengthens our confidence that in general the results of the research survey paint an adequate picture of the whole population. If doubt exist in certain instances this is explicitly stated in the text.

5.4.2 Data management survey

We do not know the total number of people actually working in data management in Europe and the number of responses we received for our data management survey is significantly lower than for the researchers' survey (262). The results are indicative rather than giving an adequate representation of the real population.

[8] http://ec.europa.eu/research/era/pdf/key-figures-report2008-2009_en.pdf p51. We do not have exact numbers for the USA or other continents. Even if we estimate that the number of researchers are roughly the same in the USA and other continents, this does not alter minimum number of responses necessary.

[9] See for example http://www.stats.gla.ac.uk/steps/glossary/sampling.html for more information on statistics.

[10] In this comparison we assumed that the distribution of researchers amongst disciplines is the same on European and global scale.

5.4.3 Publishing survey

The publishers' survey elicited 178 responses. It is difficult to determine the exact number of publishers operating worldwide.[11] It all depends on the definition of publishers. When only considering publishers who publish peer-reviewed journals, the estimate is roughly 2,000. When taking a broader approach the total may be twice that number or even more.

Yet basing the analysis on the number of publishers skews the results. A better and more realistic approach is to base the analysis on the number of journals incorporated in the survey results. The survey responses represent roughly 8,800 of the peer-reviewed scholarly journals, or 35% of the market. This approach prevents overrepresentation of small publishers. We therefore believe that we have adequate coverage of the market of peer-reviewed scholarly journals.[12]

5.4.4 Other concerns

We can be sure that the respondents are those willing to fill in surveys but those too busy or otherwise unwilling to complete the surveys will be underrepresented. In other words, we may only hear the loudest voices. Another concern is that we have had to provide some structure to the responses by means of multiple choice questions. There is clearly a danger that by doing so we could have pre-determined the answers to some extent. However, we have tried to eliminate these concerns by means of the following techniques and checks:

- We have provided free text options to allow respondents to express their own ideas, and we have then analysed these free text responses to see what ideas we have missed.
- We did relative comparisons between different groupings (e.g. disciplines, traditional publishers, open access) which will not be affected by absolute figures.

[11] It depends very much on how one defines publisher. Mark Ware in a recent report on the journal publishing market mentions an estimate of 2,000 publishers operating globally:
http://www.stm-assoc.org/about.php?PHPSESSID=5a0ce8c1d23246500dd5a6fc3042ea99
Ulrich's periodical directory, on the other hand, contains information about journals of roughly 90,000 publishers.

[12] See also the introduction of chapter 4, which deals with these issues in more detail.

6 Researchers

6.1 Introduction

For the Research Survey several distribution channels were employed. Initially the survey was sent to Elsevier's mailing list of journal editors, but due to a lack of responses from young researchers (< 10 years of professional research experience) and humanities researchers, other distribution channels were used as well. These included the mailing lists of organisations such as European Science Foundation (ESF), Marie Curie Fellowships Association (MCFA), EURODOC, and All European Academies (ALLEA).[13]

It is difficult to gauge the exact response rate, because we do not know the total number of members to these different lists. We do know that the initial invitation was sent to approximately 35,000 Elsevier editors and that 1,081 people (3.1%) responded to this invitation. In total 1,389 people responded to the researchers' survey. This analysis includes responses from the merged surveys.

6.1.1 Geographic Spread of Respondents

The geographical spread of the respondents was quite large. This is not surprising considering that the majority of results were collected from the survey that was distributed to the editors of the globally operating company Elsevier. Still, the majority of responses came from the EU (see Table 1 and Figure 2), but the single country with largest number of responses was the USA (33%)—far outweighing the United Kingdom as the largest European country (9%).

Table 1: geographic spread of responses

Country/Region	Numbers of respondents	Percentage [14]
EU	609	44%
USA	465	33%
Other	315	23%
Total	**1389**	**100%**

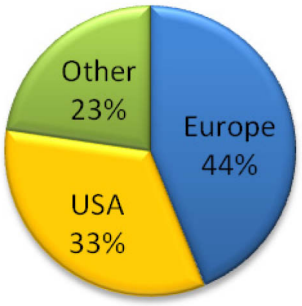


Figure 2: geographic spread

Decomposing the *EU* and *other* to single countries shows that the majority of responses came from countries that are active in the field of digital preservation and/or well-represented in (international) digital preservation projects and communities.

[13] The complete list of distribution channels: European Federation of National Academies of Sciences and Humanities (ALLEA), Digital Humanities Observatory (DHO), D-Lib Magazine, EURYI awardees through ESF, EURODOC, Humanities in the European Research Area (HERA), Marie Curie Fellows Association (MCFA), Max Planck Institute (MPI), Young European Associated Researchers (YEAR), several maling lists through the Digital Curation Centre (DCC), WePreserve, Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR), UNESCO World Heritage, and the Alliance for Permanent Access (APA).

[14] The basis for the percentages is the total number of respondents who answered the question concerned.

The following table represents the top five countries from *Europe*.

Table 2: top 5 listed countries in Europe

Country	Numbers of respondents	Percentage of total
United Kingdom	129	9%
Germany	67	5%
Italy	51	4%
France	50	4%
Netherlands	46	3%

For the *other* category the top five looks as follows.

Table 3: top 5 listed countries non-Europe

Country	Numbers of respondents	Percentage of total
Canada	66	5%
Australia	58	4%
Japan	35	3%
China	27	2%
Israel	24	2%

6.1.2 Research disciplines

In addition to comparing the advance of digital preservation in different countries, we wanted to be able to cross-analyse the results by research disciplines. We recognised nine different categories of disciplines (see chapter 5 and Appendix 1). As Figure 3 illustrates, the main categories can roughly be divided into three groups of disciplines of similar size:

1. Physical Sciences
2. Technology / Life Sciences / Social Sciences
3. Humanities / Socio-Cultural Sciences / Agriculture & Nutrition / Behavioural Sciences

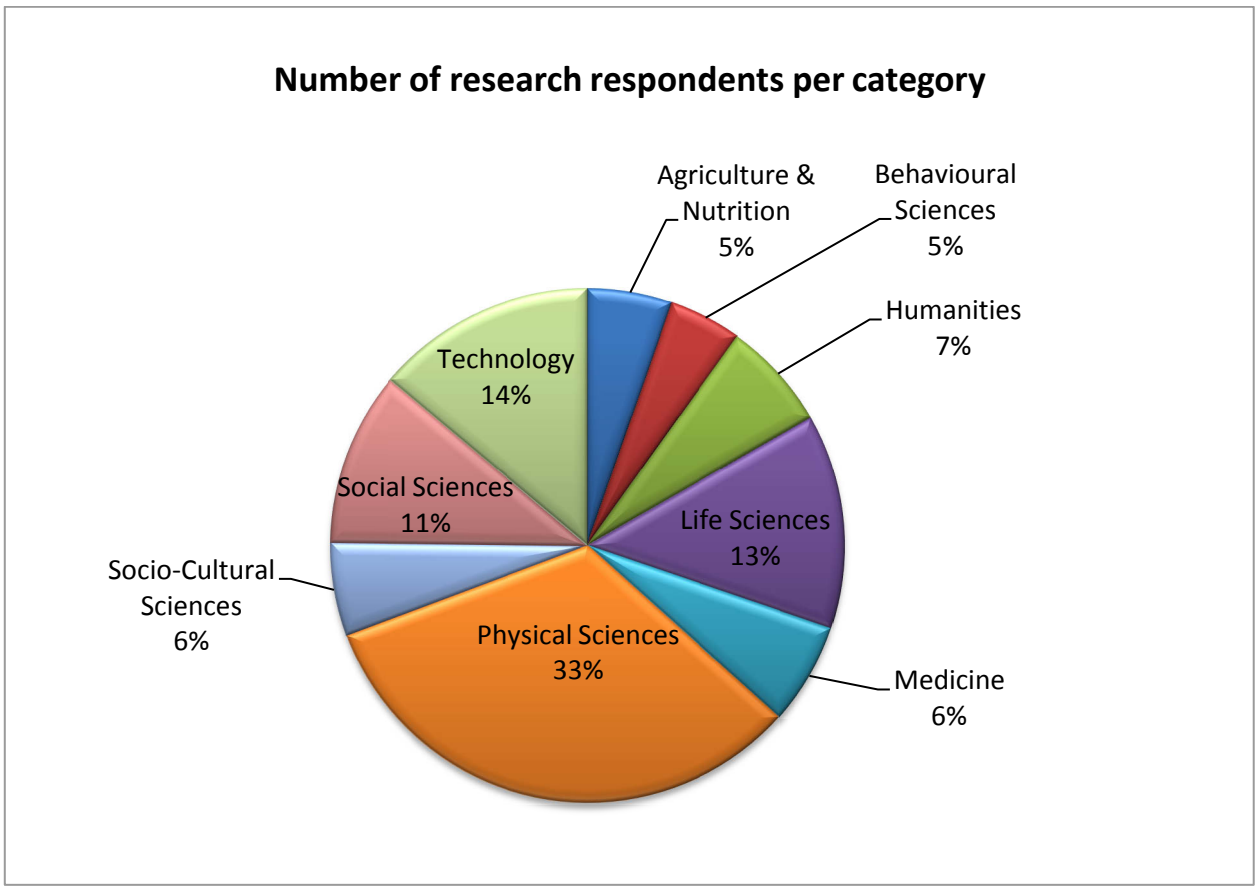


Figure 3: number of research respondents per category, n = 1387

To measure if a certain bias in research areas is present, we compared the distribution of respondents across research areas with the distribution of researchers across disciplines in Europe. The European Office for Statistics (Eurostat), offers statistical information on the number of Fulltime Employees (FTE) of researchers in the current 27 members of the European Union[15]. As Eurostat uses different categories in science, we adapted our research areas to the categories of Eurostat (see Table 4). Figures 4 and 5 show the graphs of both data sets.

Table 4: mapping of PARSE.Insight research areas to Eurostat categories

Eurostat categories	PARSE.Insight research areas
Agriculture	Agriculture & Nutrition
Engineering and technology	Technology
medical sciences	Medicine, life sciences
natural sciences	Physical sciences
Social sciences	Social sciences, behavioral sciences, socio-cultural sciences
Humanities	Humanities

[15] Science, technology and innovation in Europe, Eurostat Pocketbooks, 2008, ISSN 1830-754X, pg 36-37

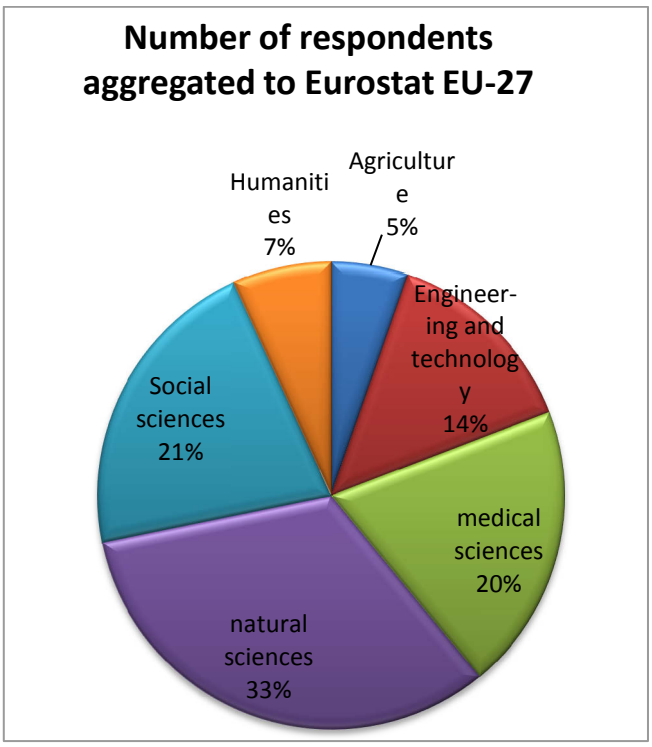


Figure 5: respondents per Eurostat category, n = 1387

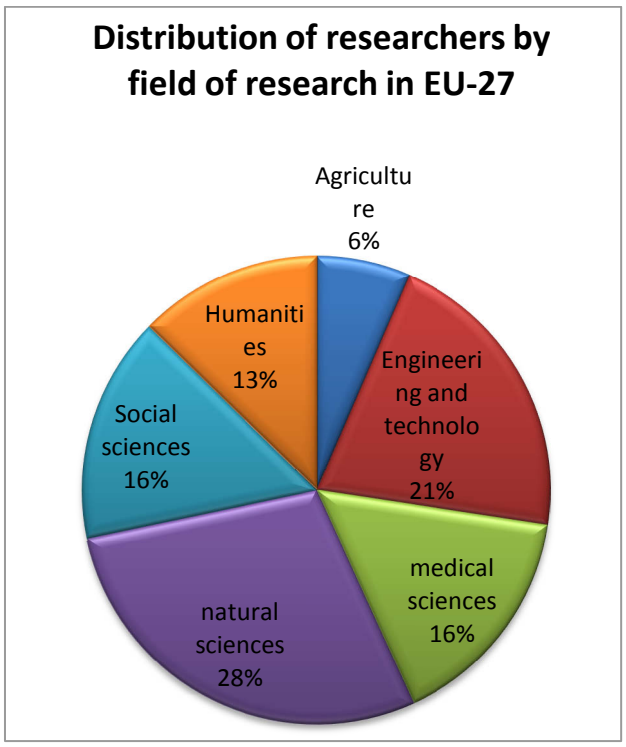


Figure 4: distribution of researchers in Europe (based on 2004 figures and in FTE), n = 625,898

The survey results show a slight overrepresentation of social sciences, natural sciences and medical sciences, while engineering & technology and humanities are less present. The strong presence of natural sciences / physical sciences (33%) seems logical as they represent the largest group of researchers in Europe. Furthermore, they might have better awareness of digital preservation as Information Technology has been a major component in the physical sciences for longer than in most other sciences and the physical sciences tend to produce far greater amounts of data than other sciences.

Overall, the results of the research survey seem to be an adequate representation of the actual distribution of research areas in Europe. Therefore, we did not apply weights to these results afterwards.

Comparison between disciplines

When it comes to similarities and differences between disciplines, we compared the largest group (physical sciences) with the other categories of disciplines. In our survey the physical sciences is an aggregation of the disciplines astronomy & astrophysics, chemistry, computer sciences, mathematics and physics. The following comparisons and differences were found:

- compared with the other respondents the respondents from physical sciences seem to deal with more data within their current research project. Not only now, but also in 2 and 5 years;
- regarding the reasons for preservation and the threats to it, physical sciences respondents do not differ from the general picture drawn of the rest of the respondents;
- respondents from physical sciences seem to be more eager to make their own data openly available. They scored about 10% higher than respondents from other disciplines;
- compared to other respondents, slightly less (about 5%) of the physical sciences respondents are interested in data outside their own discipline;

- about 10% more respondents of physical sciences stated that they use general search engines for finding new information on their research topic;
- regarding data that had already become inaccessible, physical sciences respondents stated that this is most often due to the fact that *software to interpret the data is no longer available*, while the rest of the respondents in research regard this more as the result of *hardware problems*;
- as to where to publish data, respondents from other disciplines stated that they are most willing to publish data in an archive of their own organisation, while physical sciences respondents prefer to publish data in a specific archive of their own discipline;
- however, respondents from physical sciences and other respondents seem equally unaware of available external preservation facilities such as data archives or other services;
- regarding the influence of funders, a small majority of physical sciences respondents stated that funding organisations do provide mandatory procedures for managing and preserving digital research data. Others either thought funders didn't or they didn't know.

6.1.3 Experience

The majority of respondents (70%) stated to have more than 20 years of professional experience in research (see Figure 6). This is not surprising since the Elsevier editors comprise the largest group of respondents. The majority of researchers at the editorial boards of Elsevier's prestigious journals are senior researchers. As explained earlier, we tried to make up for this by locating other distribution channels. In spite of specifically targeting the group only 10% of the respondents had less than 10 years of research experience.

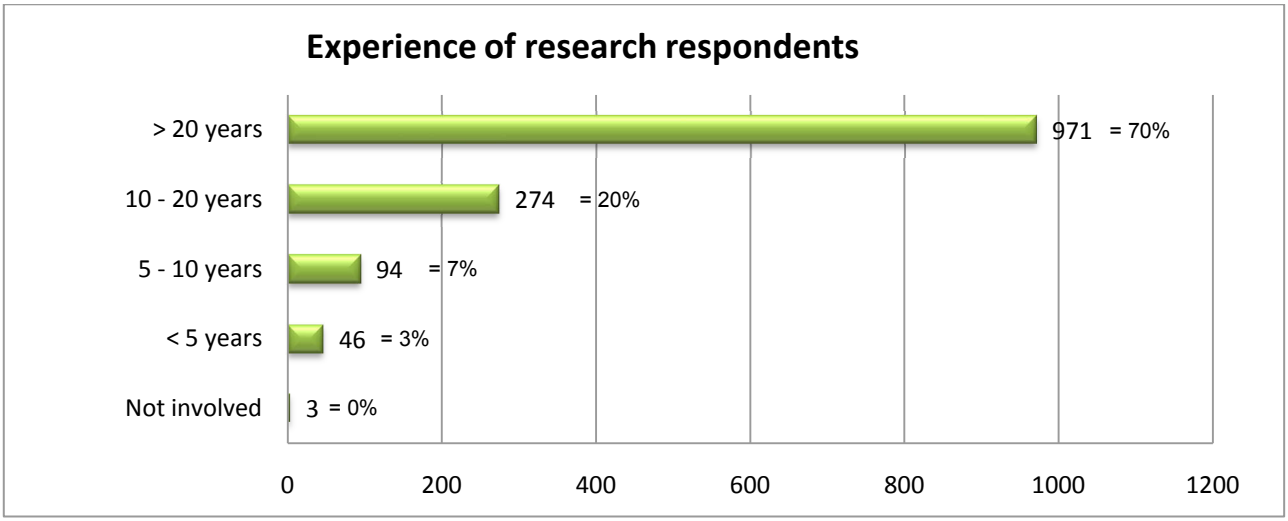


Figure 6: experience of research respondents, n = 1388

The question is if this overrepresentation of experienced researchers affects the opinion of our group of respondents as a whole. Therefore, we compared the group representing more than 20 years experience to the group of respondents that stated to have less than ten years experience in research. The comparison showed us that apart from a few small differences (less or equal to 10% discrepancy) only two bigger discrepancies (> 10% discrepancy) were found between less experienced (and potentially younger) researchers and those that have been working in research for more

than 20 years. First, novice researchers seem to be much more eager to use research output from other disciplines than experienced researchers. Secondly, although online collaboratories are not very commonly used by either of the groups of researchers, novice researchers are more in favour of having these kinds of platforms preserved as well.

Apart from these two bigger differences, the smaller ones are also interesting to note:

- experienced researchers are often research group leaders or managers;
- novice researchers are more familiar with using general search engines (such as Google) to find information for their research while experienced researchers consult their colleagues in the field. This is probably due to the fact that novice researchers are more acquainted with digital possibilities and put more trust in newer media. Experienced researchers probably already have a large network of contacts they can use;
- novice researchers are also less aware of older data that has become inaccessible;
- if they knew about data that has been lost, they considered this a result of a lack of compatible software and contextual information rather than hardware incompatibility. The experienced researchers probably have seen more changes in hardware support over time;
- a large number of novice researchers is willing to donate their research data to a digital archive within their discipline while experienced researcher are less willing to do so;
- regarding sharing of data, novice researchers clearly pointed out that they often do not share data, but that they would like to do so in the future. For experienced researchers the picture is less clear;
- novice researchers more often experience barriers in legal issues and lack of financial resources than experienced researchers;
- regarding funding, both groups stated that in the first place national governments should provide direct funding for preservation of data and publications. However, novice researchers think the European Union has a role in that as well.
- novice researchers also seem to be more optimistic about an infrastructure that can counter the threats to digital preservation;
- but regarding threats to already preserved research data novice researchers think that human errors are the most acute threat, while experienced researchers think the lack of technical support as the most important threat.

As a final check, we narrowed down the group of novice researchers to less than five years experience to see if this picture still holds. In this case, some of the differences were more pronounced than in the broader group of novice researchers, but none of them is considered significant.

6.2 Perceptions of preservation

The focus of this section is the respondents' perception of preservation issues. Preservation is a confusing term. What some, unfamiliar with intricacies of preservation, may regard as preservation, others would simply call storage. This is what we wrote in the introduction to the survey to explain the difference.

In this questionnaire we make a distinction between **storing information**, routinely, in your day-to-day practice, on your computer or a faculty server, and **preserving information**, meaning data is specifically curated to be re-usable in the long term. In the latter case not only the data itself must be archived, but also information about the data: where did the data come from? How have they been stored? Which file formats have been used? What special terminology or other information is needed to interpret and use the data? etc.

Bearing this in mind, respondents answered questions on the (perceived) reasons for preservation; they evaluated the importance of certain threats to preservation and expressed their opinion about the need for an infrastructure to counter the threats.

6.2.1 Reasons for Preserving Data

While the reasons for preserving digital research data are often regarded as self-evident by the specialists, it is good to know what researchers think about these reasons. Researchers were presented with a list of seven well-known reasons for preserving data and asked whether they regarded the reasons as *very important*, *important*, *slightly important*, or *not important*. The reasons are:

- if research is publicly funded, the results should become public property and therefore properly preserved;
- it will stimulate the advancement of science (new research can build on existing knowledge);
- it may serve validation purposes in the future;
- it allows for re-analysis of existing data;
- it may stimulate interdisciplinary collaborations;
- it potentially has economic value;
- it is unique.

Looking at the *very important* and *important* results, it is clear that researchers consider the possibility of re-analysis of existing data as the most important driver for the preservation of research data, closely followed by future validation purposes (90%), the advancement of science (89%), and public funding (87%).

Economic value is regarded as the least important reason for preservation. Only 39% of the researchers who responded perceived economic value as either an *important* or a *very important* reason for preservation. The stimulation of interdisciplinary collaborations (71%) is still regarded as rather important, while a slight majority also considers the uniqueness of the research data as either an *important* or a *very important* reason to preserve research data (see Figure 7).

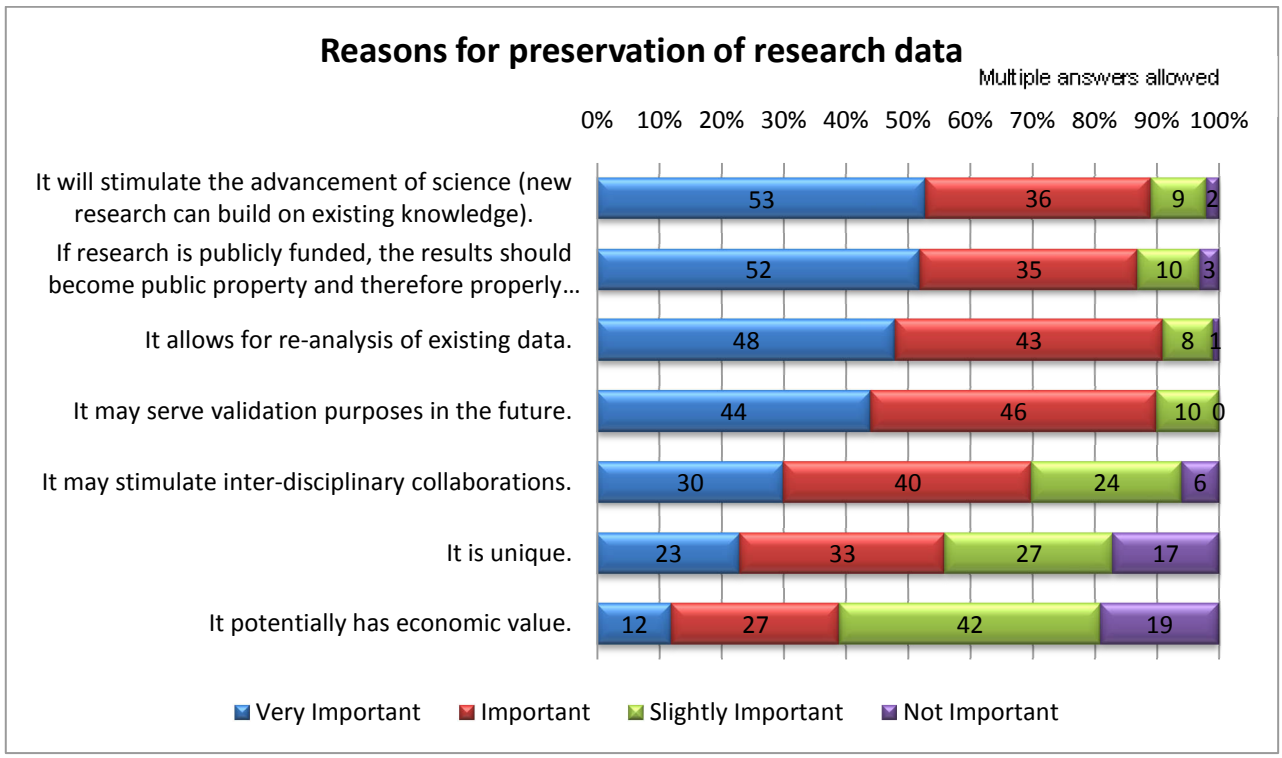


Figure 7: reasons for preservation of research data, n = 1213

Different disciplines have different needs and requirements. It is interesting to see whether these differences are reflected in the respondents' opinions on the formulated reasons for preservation. To compare them we collected the top three reasons for each discipline. We based the position of the reasons on the *very important* answers for all reasons. The reason with the highest number of *very important* responses occupied the top position; the reason with the second highest number of responses took the second position, etcetera. We did this for each of the nine main disciplines categories we identified. If the number of responses was equal for two or more reasons we looked at the *important* answers as a tie-breaker.

There is no disagreement on the three least important reasons. All disciplines consider economic value, the uniqueness of the data, and the possibility of interdisciplinary collaborations as the least important reasons for preservation. None of the disciplines had these reasons in their top three of most important reasons.

Furthermore, only behavioural scientists placed future validation purposes in their top three of most important reasons. Yet, with the exception of behavioural scientist the following reasons had the highest number of 'very important' responses for all disciplines:

- If research is publicly funded, the results should become public property and therefore properly preserved.
- It will stimulate the advancement of science (new research can build on existing knowledge).
- It allows for re-analysis of existing data.

The order of the top three of course differs for the disciplines. The disciplines Agriculture & Nutrition, Social Sciences and Technology place the highest emphasis on public funding as the prime

reason for preservation. For disciplines Humanities, Life Sciences, Physical Sciences, and Socio-Cultural Sciences, the reason that preservation may stimulate the advancement of science gained the highest number of 'very important' responses. Finally, for Behavioural Sciences and Medicine the highest number of 'very important' responses was ascribed to the possibility of re-analysis of existing data.

6.2.2 Threats to Digital Preservation

Software, hardware, organisations, and people are all important elements in the production, distribution and consumption of digital research data. Yet, at the same time, they may also be a threat to the long-term availability and usability of that data. To enable the transmission of digital data to future users it is important to tackle the threats that endanger a smooth transmission.

The survey contained two questions on the threats to digital preservation: a specific question on seven detailed threats with immediate and direct impact on all digital data and a more general question with threats which may occur but are not relevant to all data and in all situations.

For the specific threats question we formulated seven threats, similar to the threats used in the projects CASPAR[16] and SHAMAN[17]. The seven threats are:

- users may be unable to understand or use the data e.g. the semantics, format or algorithms involved;
- lack of sustainable hardware, software or support of computer environment may make the information inaccessible;
- evidence may be lost because the origin and authenticity of the data may be uncertain;
- access and use restrictions (e.g. Digital Rights Management) may not be respected in the future;
- loss of ability to identify the location of data;
- the current custodian of the data, whether an organisation or project, may cease to exist at some point in the future;
- the ones we trust to look after the digital holdings may let us down.

For each of these threats respondents were asked to indicate their importance. The choices available were *very important*, *important*, *slightly important*, *not important*, or *don't know*. An important point to notice is that there would be a temptation for respondents to pick the middle box i.e. *slightly important*.

[16] EU FP6 project CASPAR: http://www.casparpreserves.eu/
[17] EU FP7 project SHAMAN: http://shaman-ip.eu/shaman/

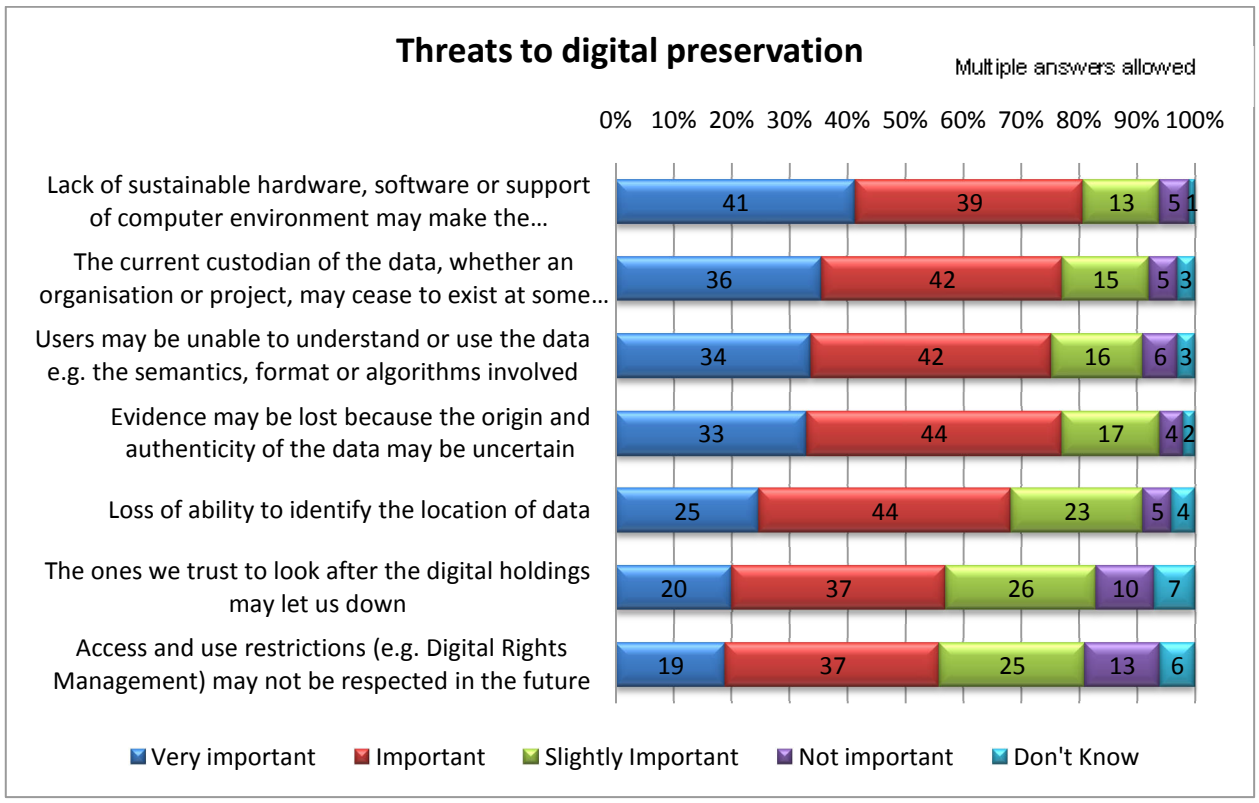


Figure 8: threats to digital preservation, n = 1209

Figure 8 shows that there is high degree of awareness on the major threats to long-term preservation of digital research data. Between 56% and 80% of the responses indicate that all threats are recognized as either *important* or *very important*. Access and use restrictions is regarded as the least important threat to preservation, while the lack of sustainable hardware, software or support is recognized as the most important threat to preservation.

Similar to the question on the reasons for preservation, we made a top three for all disciplines on the threats to digital preservation. There is no disagreement on the three least important threats. All disciplines consider the following threats as the least important threats to preservation:

- Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future
- Loss of ability to identify the location of data
- The ones we trust to look after the digital holdings may let us down

If we look at the threats which are considered most important by the disciplines, a diverse picture emerges. Most disciplines agree that the influence the lack of sustainable hardware and software or support may have on preservation is considerable. The humanities researchers seem mostly concerned with the threat that future users may be unable to understand the data. Researchers from the agriculture & nutrition disciplines and medicine disciplines are most concerned with the loss of evidence due to uncertain origin and authenticity of the data. Sustainability is also a major concern among the researchers. Many—especially socio-cultural and social sciences researchers—consider the possibility that organisations or projects may cease to exist a major threat to the preservation of digital research data.

In addition to their opinions on the above-mentioned detailed threats to preservation, respondents were asked to attach importance to several more general threats. Respondents were presented with a list of several general threats and asked whether they thought these were important threats to their current digitally stored data. The choices available were:

- Lack of structural funding
- Lack of technical support
- Natural disasters
- Political instability
- Continuity of organisation
- Human errors
- Don't Know
- Other

There is some overlap with the detailed threats question here, and it is perhaps not surprising that, in general, the lack of technical support was checked most often as one of the main threats. 59% of the respondents chose the lack of support option (see Figure 9). We may recall that the lack of support, together with lack of sustainable hardware and software, was counted as the most important threat to preservation for the detailed threats question.

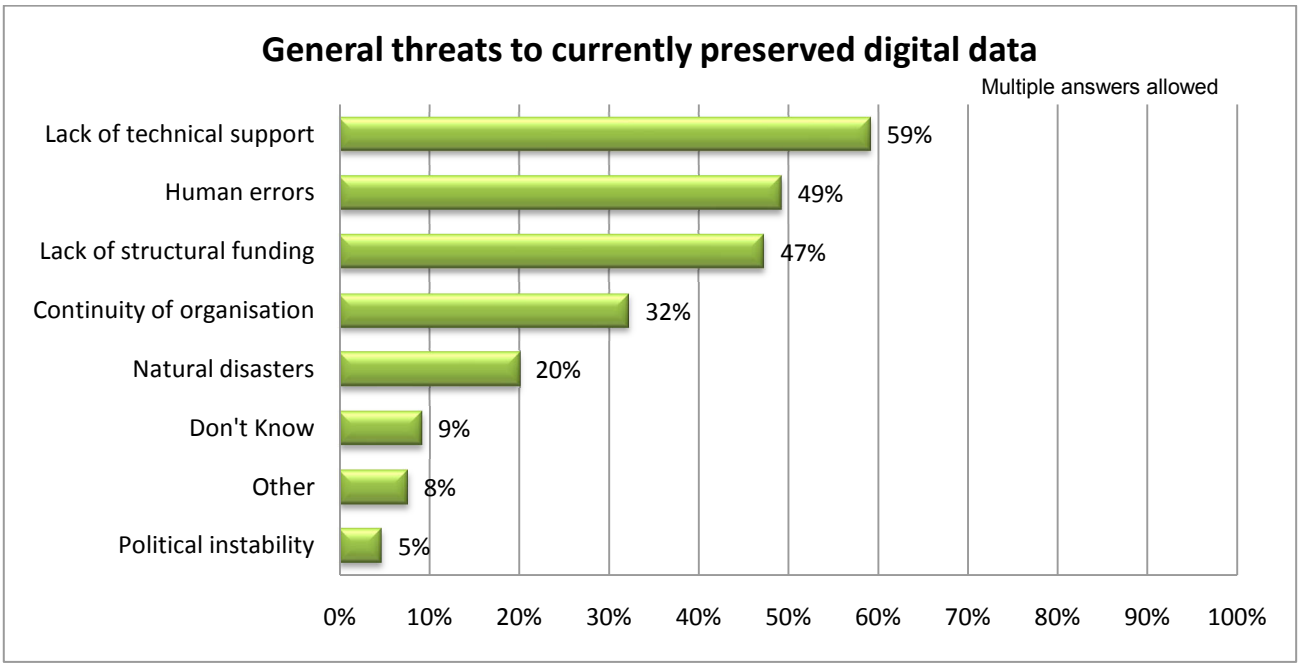


Figure 9: general threats to digital preservation, n = 1190

The two other major threats of the list of general threats are human errors and the lack of structural funding. The second one here is perhaps not very surprising and ties in with other sustainability issues as support for hardware and software. Funding and human errors are both elements that are not tackled by technique. In addition to the predefined threats people also felt that security issues—viruses, hacking, theft, vandalism—and personnel issues—change of jobs, retirement, death—may be threats to preservation.

6.2.3 The need for an Infrastructure

PARSE.Insight is based on the premise that an e-science infrastructure will deal with many of the threats to preservation of digital research data. We wanted to know the opinion of researchers on the e-science infrastructure as a solution to preservation (see Figure 10). 58% of research respondents believe that some kind of international infrastructure for data preservation and access should indeed be built to help guard against some of the above-mentioned threats.

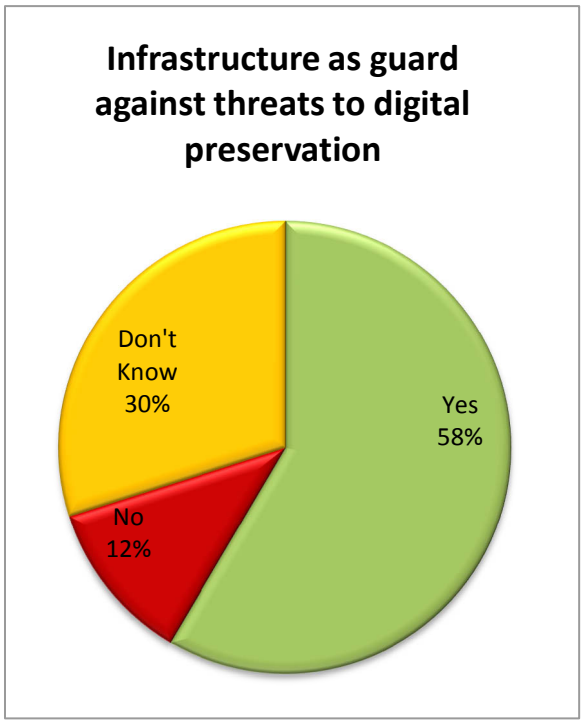


Figure 10: the need for an infrastructure, n = 1207

If we break this down to discipline-level, what strikes most is the significantly higher than average percentage of humanities researchers (75%) who feel that there is a need for an e-science infrastructure to counter the threats to digital preservation. It may be a sign of the advance technology has made in the humanities, but more research would be needed to know whether this could explain the percentage. [18]

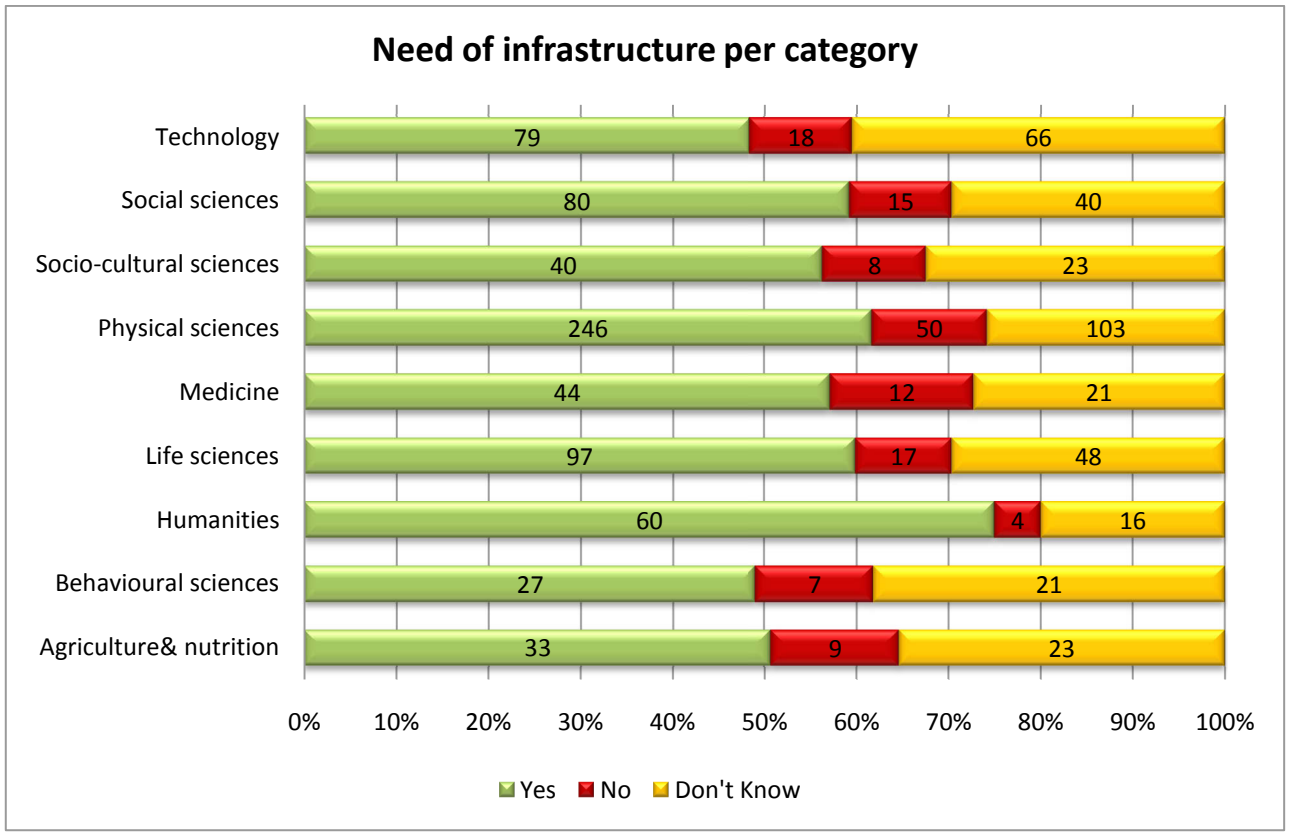


Figure 11: need for infrastructure per research category, n = 1207

[18] See also deliverable 3.3 Case Studies Report (Jan 2010), which includes the humanities case study on Book Studies.

When asked what such an infrastructure should look like, the answers were anything but uniform. A tag cloud in figure 12 shows that as well. Many admitted not to know, but the idea of building some sort of central repository/database for researchers was espoused several times. In contrast, others mentioned a distributed network of safe places across Europe. Countering the threats to digital preservation encompasses more than dealing with the technical aspects.

A large part of the problem is also psychological. And there is ample proof for that in the survey as well. Making an e-science infrastructure available does not necessarily mean researchers will make use of the tools and sources available.

Figure 12: tag cloud of answers on what an infrastructure should look like (created with Wordle.net)

Apart from an infrastructure, we also formulated additional needs and requirements we believed to be necessary to guarantee that valuable digital research data is preserved for access and use in the future:

- Training
- More expertise
- More resources
- More digital repositories

Respondents noted that all other needs are of (almost) equal importance (see Figure 13).

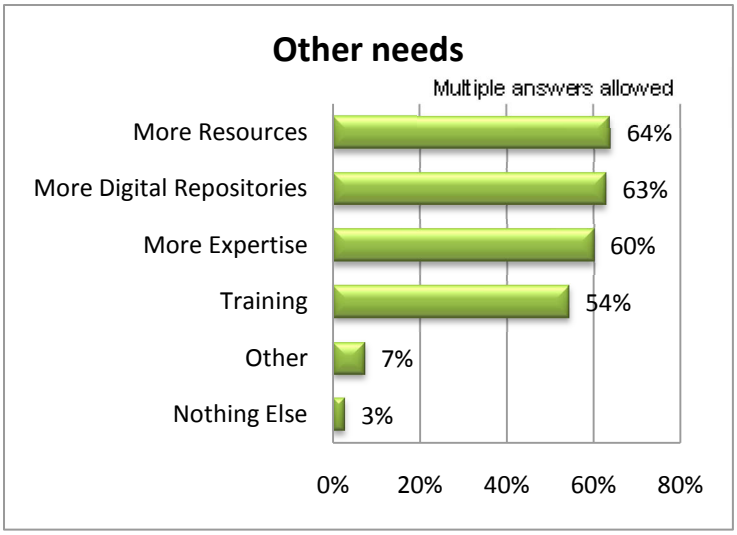


Figure 13: needs apart from infrastructure, n = 1202

6.2.4 Initiatives to raise the level of knowledge

Digital preservation is still a relatively new field of study. While progress is being made and the awareness and knowledge surrounding digital preservation is spreading, most respondents agree that more knowledge/expertise is necessary. What can be done to raise the level of knowledge regarding digital preservation? Training in one form or another seems an obvious method here. But one could also develop expert workshops. Forums (online or physical) for the exchange of knowledge may be useful, as may the development of guidelines and manuals that describe how to preserve digital data. The respondents believe that especially the guidelines and manuals would be useful here, but in general all above-mentioned measures are considered to be useful for raising the level of knowledge (see Figure 14).

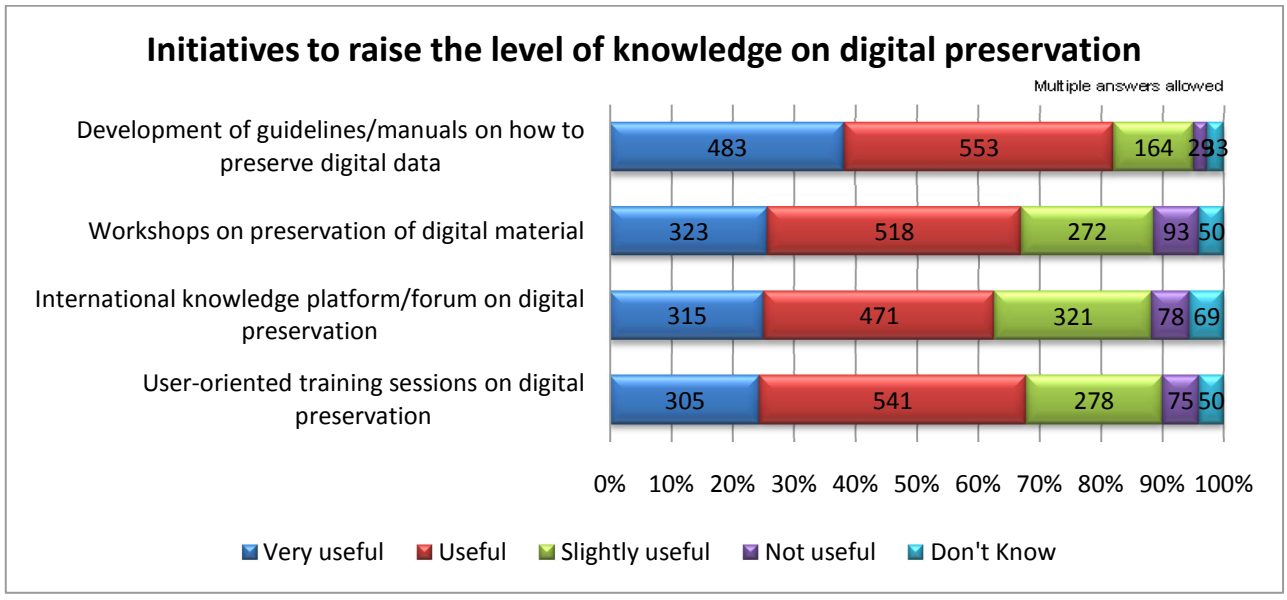


Figure 14: initiatives to raise level of knowledge on digital preservation, n = 1249

6.3 Preservation – the state of affairs

To be able to determine what is needed for the preservation of research data, we need to know more about the day-to-day work practice of researchers. What kind of digital data do they have? How much data they do produce and/or use? How do researchers store their data? What do they store? And how much do they store? These are the questions which this section deals with.

6.3.1 Data types

Each type of data or data format has its own set of characteristics and therefore requires a distinctive preservation strategy. To determine the kinds of data currently used by researchers a list of data types was formulated. Respondents were asked to check the data types they use. Not surprisingly, office documents are most often used by the respondents (see Figure 15). What is a bit surprising perhaps is that still 6% of the respondents do not use office documents. The other two of the top three most used data types are: network-based data (web sites, e-mail, chat history, etc.) and images (such as JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc). For both data types 79% of the respondents claimed to use them.

What is rather more surprising is that almost half of respondents have source code, software applications, raw data and databases. It is likely that these forms of digital objects offer significant challenges in terms of usability and understandability, beyond those of documents and images.

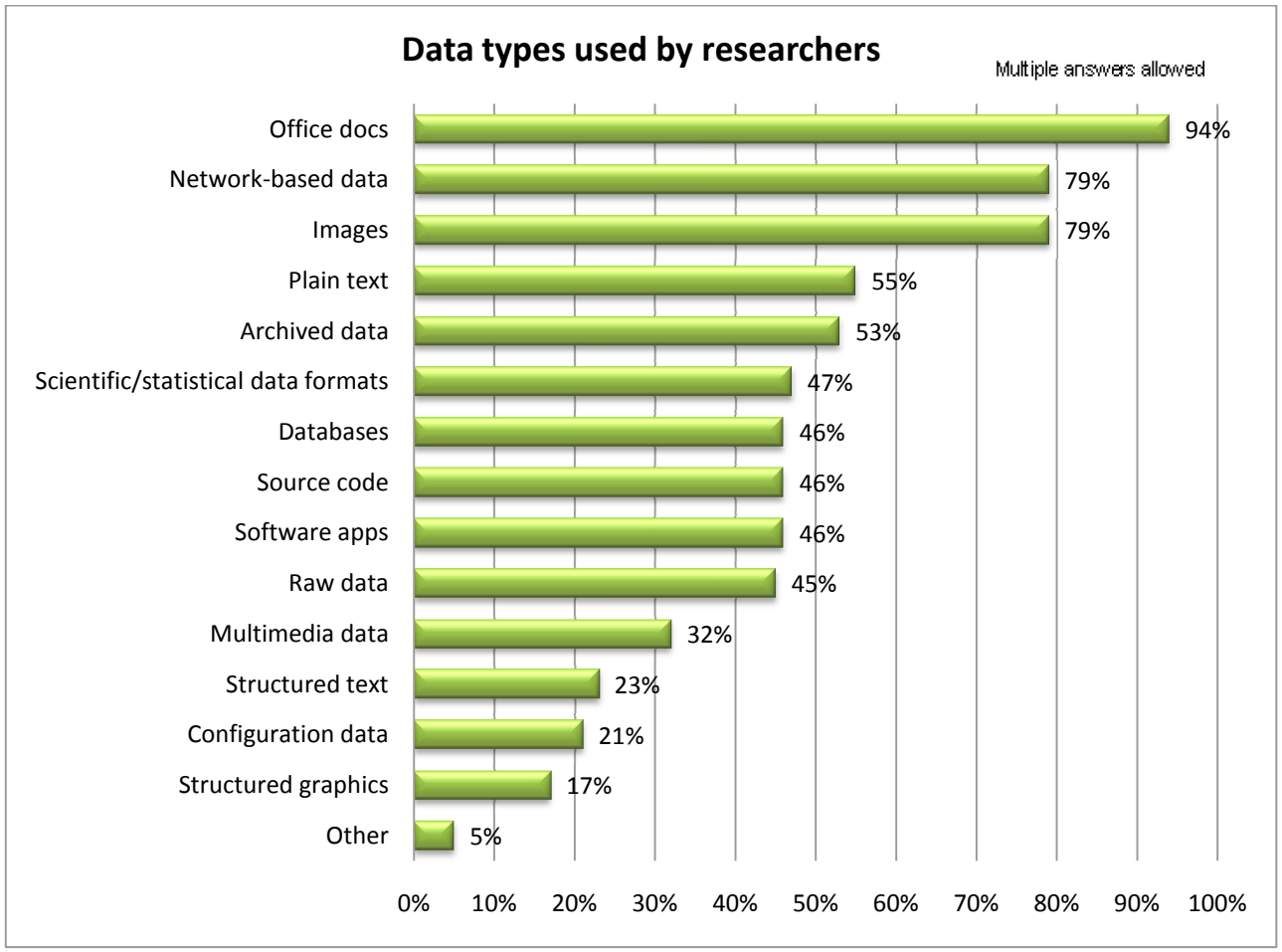


Figure 15: data types used by researchers, n = 1366

6.3.2 Amount of Data

Besides the kind of data types researchers use, they were also asked to provide an estimate of the data they currently store as well as the amount of data they think they will store in two and five years respectively (see Figure 16). This proved to be a difficult question. About 10% of the respondents really had no idea about the amount of data they currently store, a percentage which grew to 17% for the estimate of stored data in five years.

The break seems to appear around 1TB. With a surprising exception for 0MB, the chart shows that the percentages for the data ranges up to 1TB are declining with the years, while those larger than 1TB are increasing with the years. Within five years only 7% of the respondents estimate to store 1 PB or more, while 76% of the respondents estimate that their stored data amounts to less than 1PB—17% doesn't know.

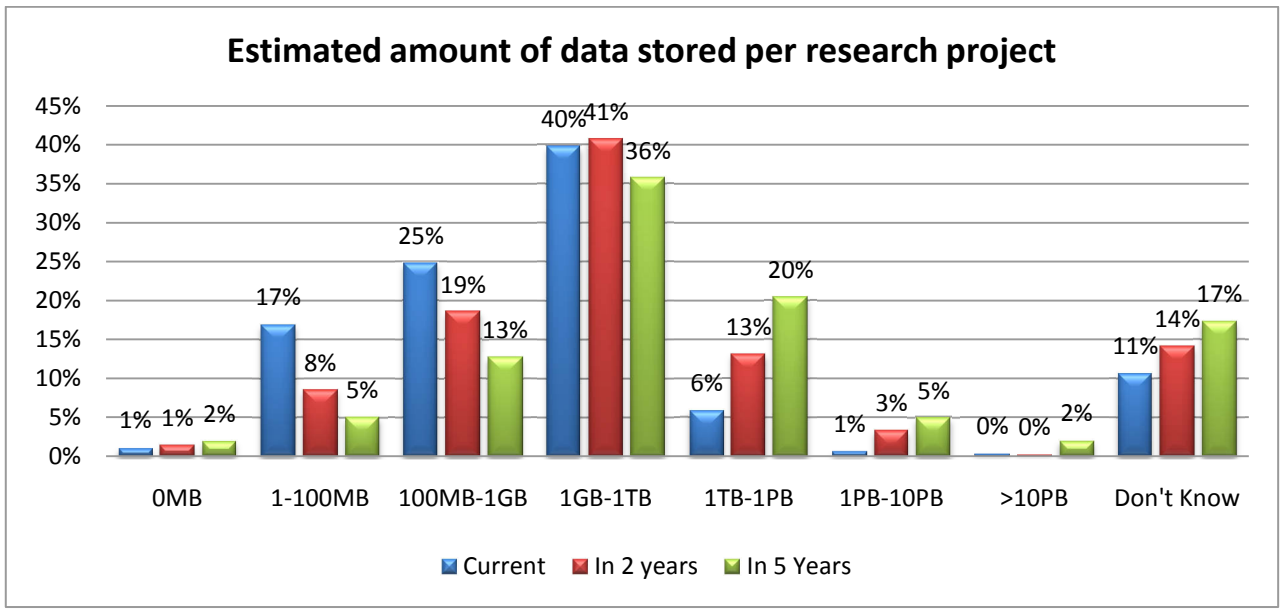


Figure 16: estimated amount of data stored per research project, n = 1296

For preservation purposes it is good to know that a significant number of respondents also assign additional information to their digital research data. 39% of the respondents claim to assign administrative information (e.g. creator, date of creation, filename, provenance) to their data, while 30% claim to record technical information (e.g. encoding type, description, file format, settings, software utilities); it is, however, also concerning that 70% do not record such vital details.

6.3.3 Where data resides…

When asked where researchers store their research data, the most important locations, in order of the number of responses, are: personal computer at work (81%), portable storage carrier (66%), organizational server (59%), and computer at home (51%). Of the 41% of the respondents who do not store data on organisational servers the majority stores their data on a local directory on their computer at work, on portable storage carriers, or on the computer at home (see Figure 17).

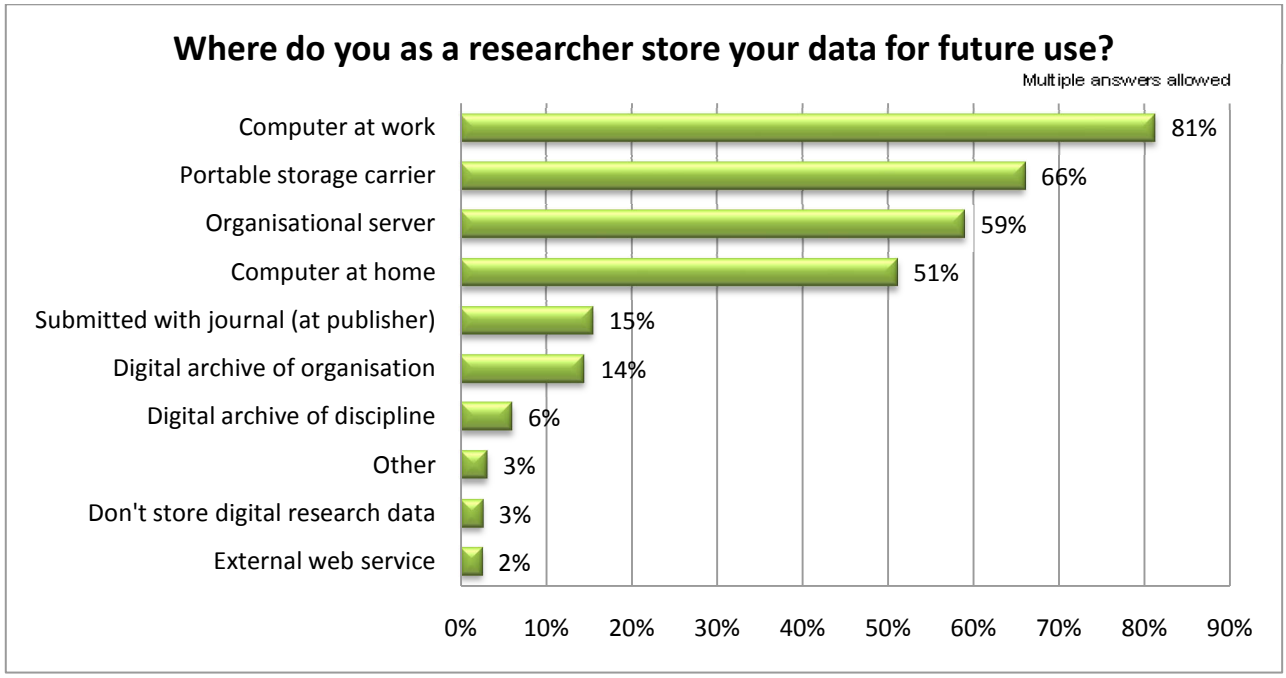


Figure 17: where researchers keep their data for future use, n = 1202

Only 20 % of the respondents submit data to a digital archive. This is a telling figure which is more meaningful in the context of the questions we asked about sharing data (see section 6.6 cross-disciplinary use of data).

6.4 Preservation – the outlook

The amount of data is growing, but currently not many researchers store their data in digital archives. This can partly be explained by a lack of trust in those digital archives, but it may also be that researchers are unfamiliar with existing digital archives or that there are simply not enough archives. When asked whether researchers knew if there are any plans to build digital archives in the (near) future, most responded not to know (84%), while roughly 10% believed one to be created within the next three years (see Figure 18).

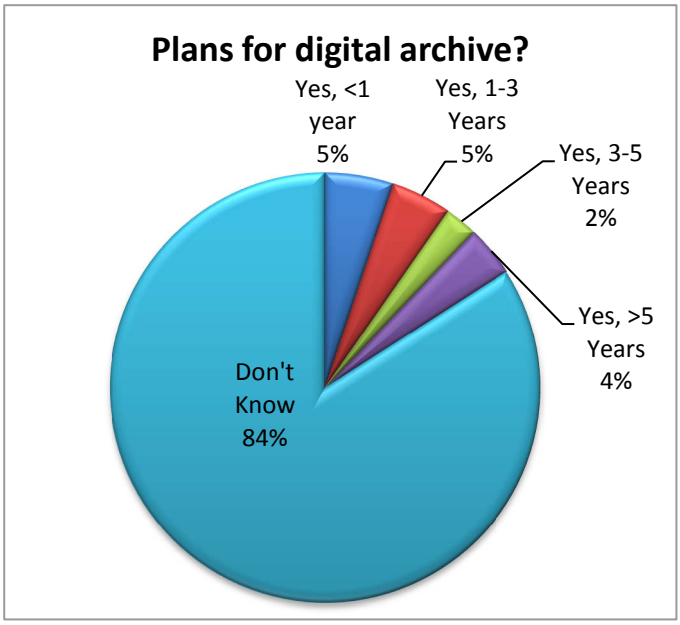


Figure 18: plans for a digital archive, n = 1200

6.5 The cross-disciplinary use of research data

As it turns researchers are not so eager to share their research data with others. Only 25% of the respondents state their research is openly available to everyone (see Figure 19). For the others there is some barrier or restriction. Some do not make any data available while others make it only available to researchers with whom they closely work together. Only 11% of the respondents make their data available for researchers within their research discipline. The majority of respondents do make their data available to researchers within their research collaborations and groups, but even 58% may not be considered a very high figure.

While the percentages of the respondents who share data are small, sharing does take place. However, the sharing of these data does not seem to take place through established digital archives, not even when they are specific to the discipline. The obvious conclusion would be that researchers want some sort of control over their data and they see many problems surrounding the sharing of data (see Figure 20). The major problems researchers foresee in sharing their data through digital archives are legal issues (41%), misuse of data (41%), and incompatible data types (33%). Based on the responses, it looks like there still is a lot of distrust in the capability of digital archives to properly handle research data.

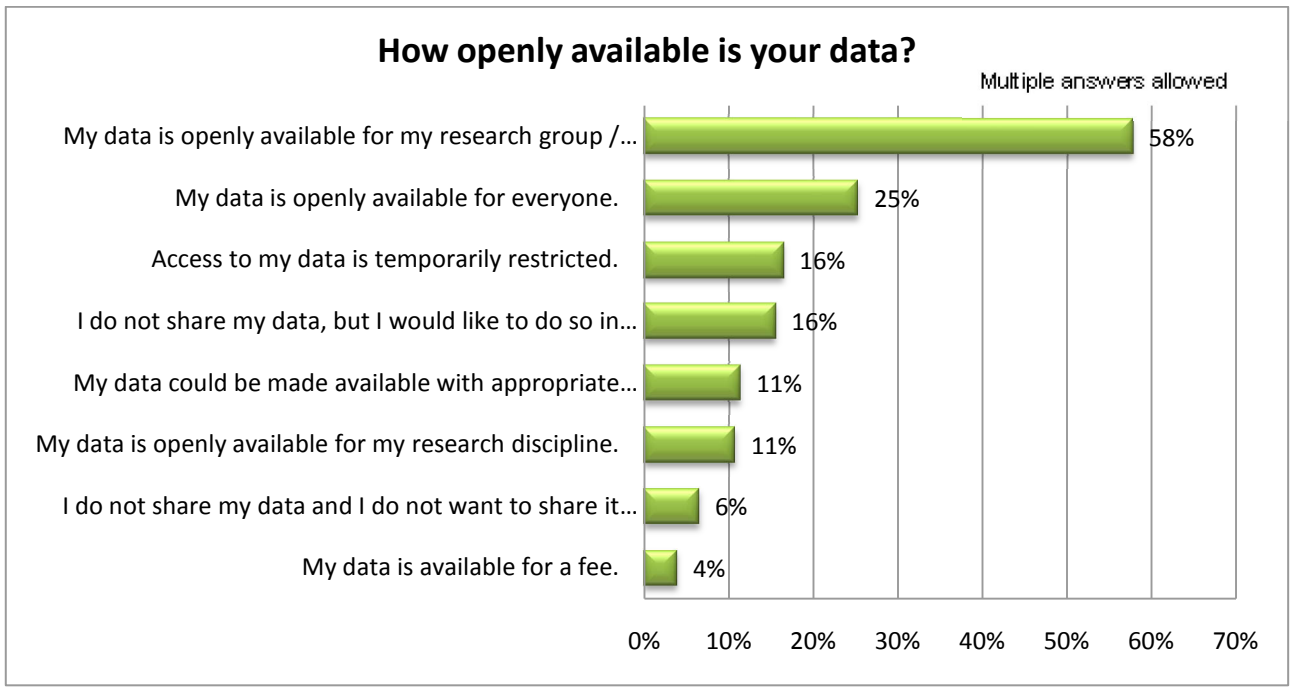


Figure 19: how openly available is your data? n = 1270

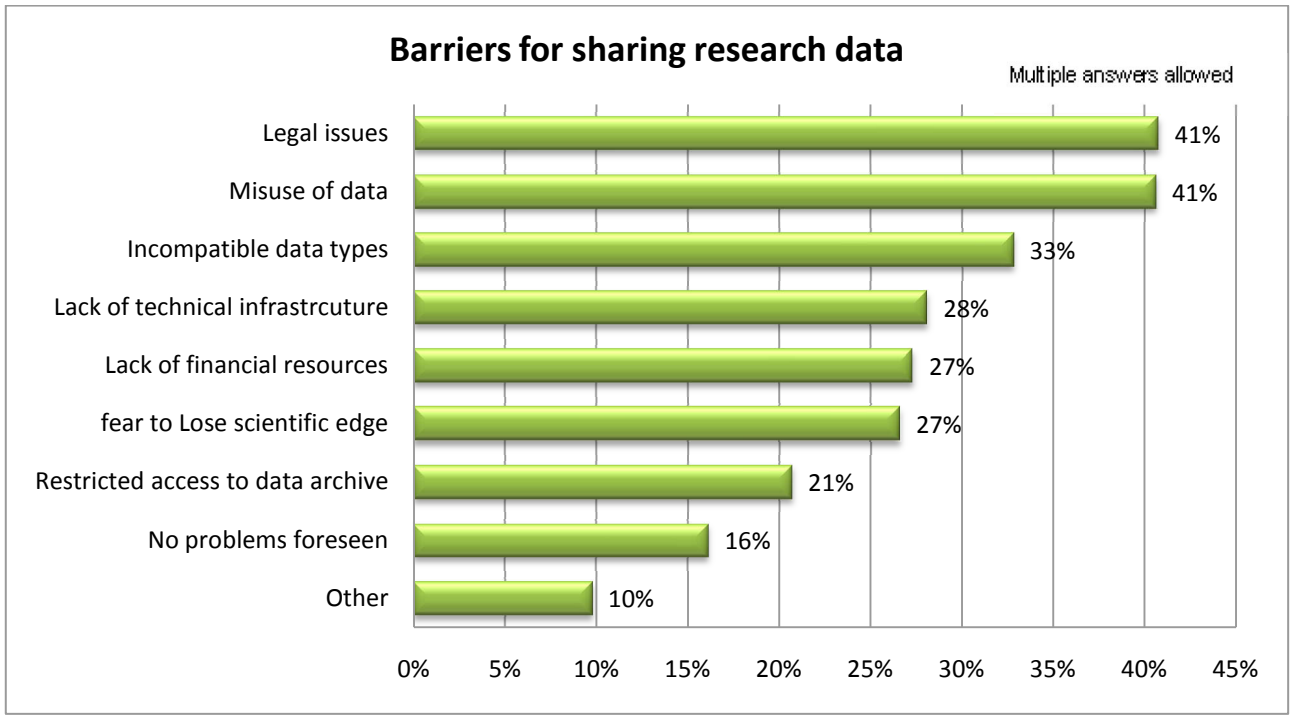


Figure 20: barriers for sharing research data, n = 1270

The current practice is not be explained by a disinterest of researchers in other people's data. 63% of the researchers, who do not currently make use of other researchers' data within their discipline, would like to do so in the (near) future — 40% for data from other disciplines.

When asked whether they ever truly needed digital research data by other researchers that was, for whatever reason, not available, 53% of the respondents answered yes (see Figure 21).

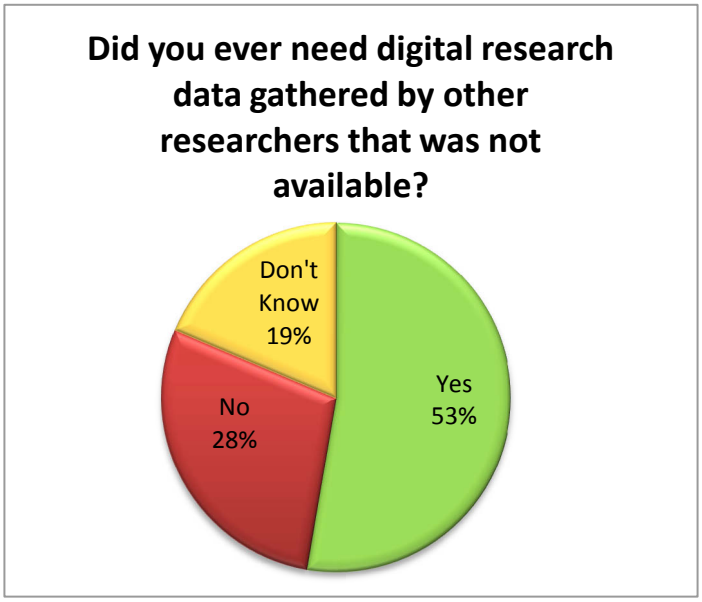


Figure 21: need for data from other researchers, n = 1252

6.6 Funding

Besides the technical and human elements digital preservation involves costs. Respondents had to provide their views on who is responsible for the preservation of research data and publications. A majority of the respondents believed that their national government should pay the bill for the preservation of research data (61%) and publications (57%) (see Figure 22 and 23).

The top three for research data is completed by the researchers' organisations (41%) and the European Union (36%).

For publications the picture looks a bit different. More than for research data, many respondents believe that the brunt of the costs for the preservation of publications should be borne by publishers (42%) or the research community (35%).

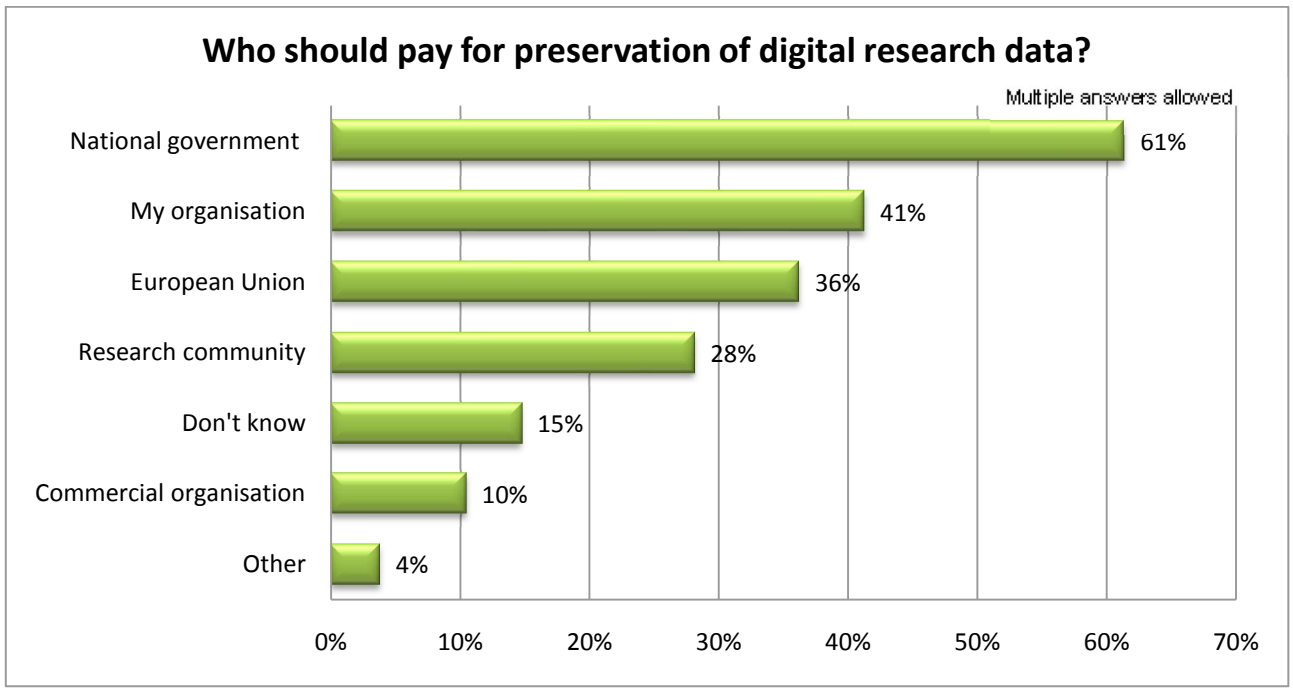


Figure 22: who should pay for preservation of digital research data? n = 1188

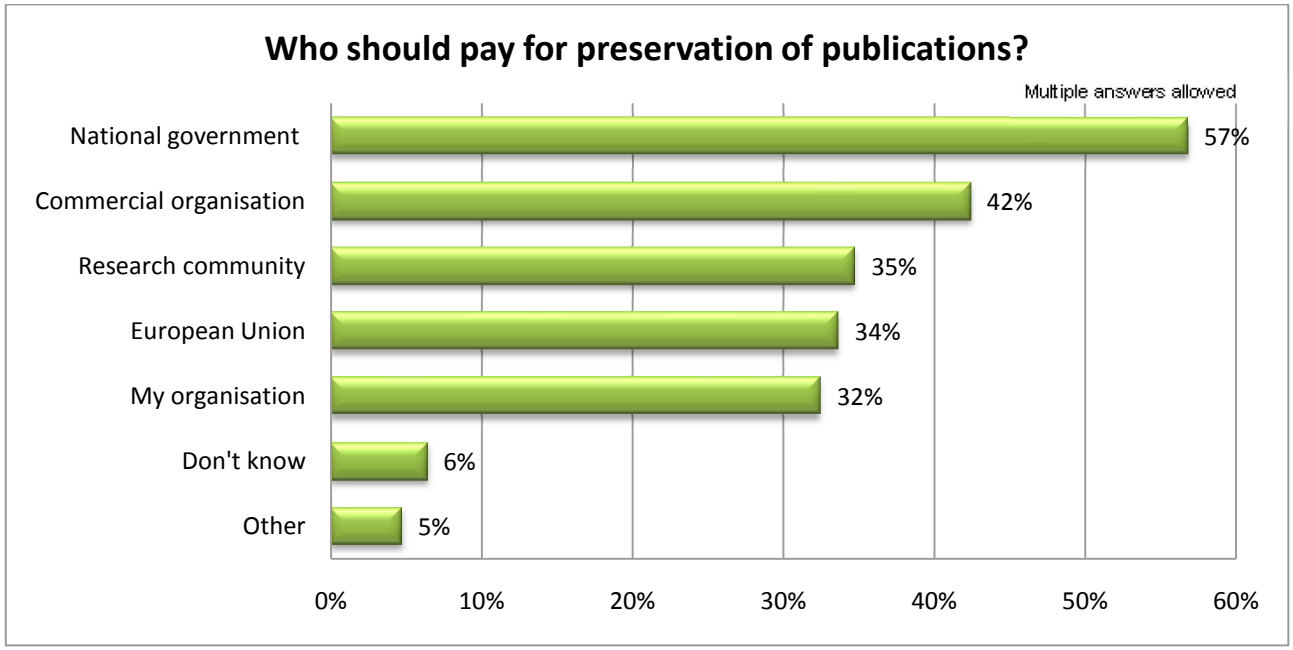


Figure 23: who should pay for preservation of publications? n = 1188

7 Data Managers

7.1 Introduction

Good data management is essential for research and a benefit to all. It involves many actors and starts at the creation of data. The better researchers manage their research data the easier that data can be shared and preserved for future access and use. In fact, the researcher is also a data manager, but for this report our definition is less broad.

In this report data managers refer to professionals with a clear responsibility for the preservation of research data and publications. As we have seen in the prior chapter, researchers' perceptions, needs and requirements are dealt with separately. Data managers here then refer to research libraries, data centres, archives, and other data management organisations (see Figure 24).

We did not analyse the data individually for all kinds of organisations present in the survey, but we did look at the kinds of organisations that are represented in the data managers' survey. It is based on the unique number of organisations in the respondents list of the survey—possible doublings are accounted for. The largest group of respondents are research libraries (64%), which is understandable since the surveys we distributed though the LIBER survey has the largest number of responses.

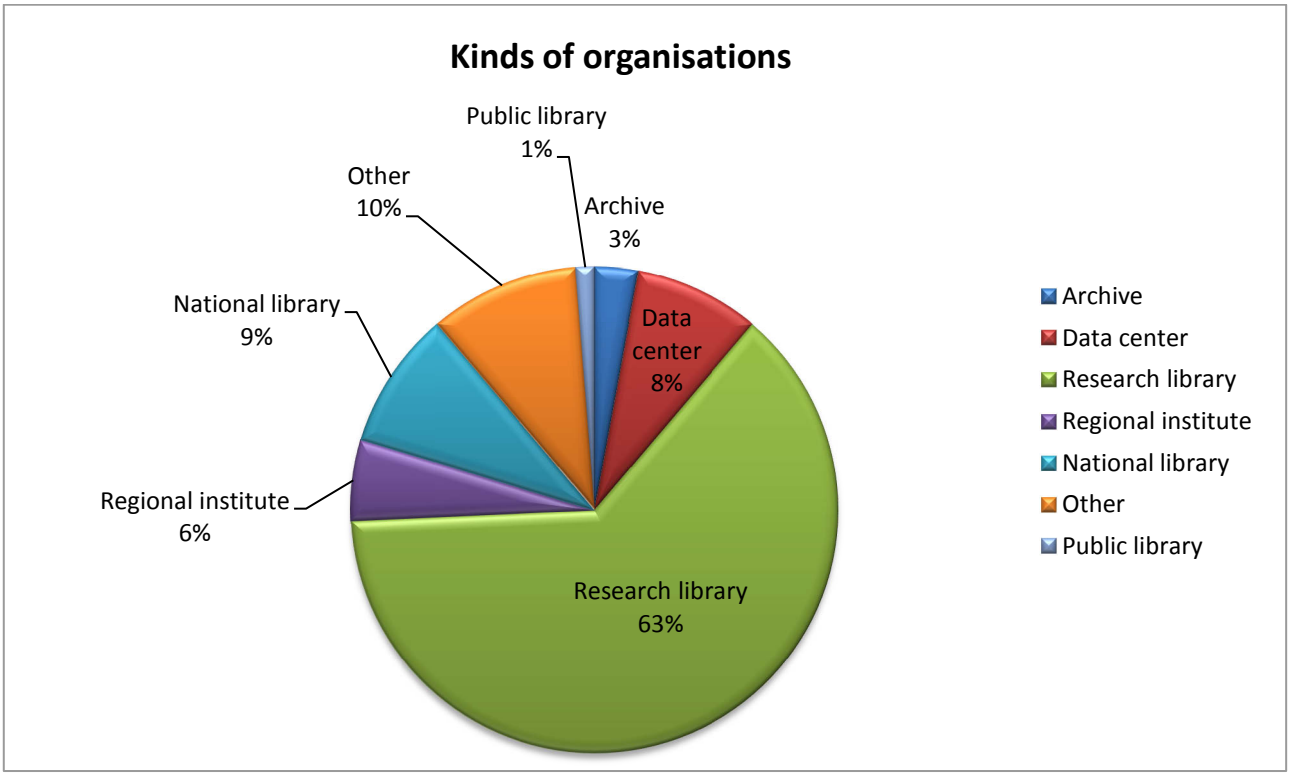


Figure 24: kinds of data management organisations , n = 241

The distribution channels of the Research and Data Management surveys partly overlap. Merged surveys were sent to mailing lists which could attract responses from any of the stakeholder roles. For the Data Management survey the following mailing lists were the most important:

- Association of European Research Libraries (LIBER)
- WePreserve
- Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)
- Digital Curation Centre (DCC)
- Alliance for Permanent Access (APA)

Again, since we do not know the total number of members for each list it is impossible to calculate the exact response rate. In total the Data Managers surveys yielded 273 responses of which 146 respondents completed the survey.

7.1.1 Country of Respondents

The majority of responses came from Europe (see table below). Unlike the other surveys, the Data Management survey used a specific European distribution channel, LIBER. Most responses came from the LIBER distribution channel (45%), with DCC (42%) as the second large source of responses.

Table 5: geographic spread of data management respondents

Country/Region	Numbers of respondents	Percentage
EU	207	76%
USA	38	14%
Other	28	10%
Total	273	100%

7.2 Perceptions of preservation

The questions of this section deal with the respondents' perception of preservation issues. Similar to the researchers' survey, respondents to the data managers' survey answered questions on the (perceived) reasons for preservation; they evaluated the importance of certain threats to preservation and expressed their opinion about the need for an infrastructure to counter the threats.

Unlike researchers, whose main concern is the availability of the data for individual research purposes, preservation issues are at the heart of the data managers' daily activities. So it is interesting to see if and in what way the data managers' perception differs from the researchers and what implications these possible differences have on the Roadmap. The latter issue will be dealt in chapter 6.

7.2.1 Reasons for Preservation

Data Managers were presented with the same list of seven well-known reasons for preserving data as the one for researchers and asked whether they regarded the reasons as *very important*, *important*, *slightly important*, or *not important*. Again, the reasons were:

- If research is publicly funded, the results should become public property and therefore properly preserved
- It will stimulate the advancement of science (new research can build on existing knowledge)
- It may serve validation purposes in the future
- It allows for re-analysis of existing data
- It may stimulate interdisciplinary collaborations
- It potentially has economic value
- It is unique

Six of the seven reasons formulated were regarded as either *important* or *very important* by 76% to 98% of the respondents (see Figure 25). Even more than researchers, data managers believe public funding (98%) to be either an *important* or *very important* reason to preserve research data. The other two major reasons are the way in which preservation will stimulate the advancement of science (96%) and the fact that preservation makes re-analyses of existing data (95%) possible. Only the potential economic value of research data is regarded as a bit less important reason. Still, 62% of the respondents regard economic value as either an *important* or *very important* reason.

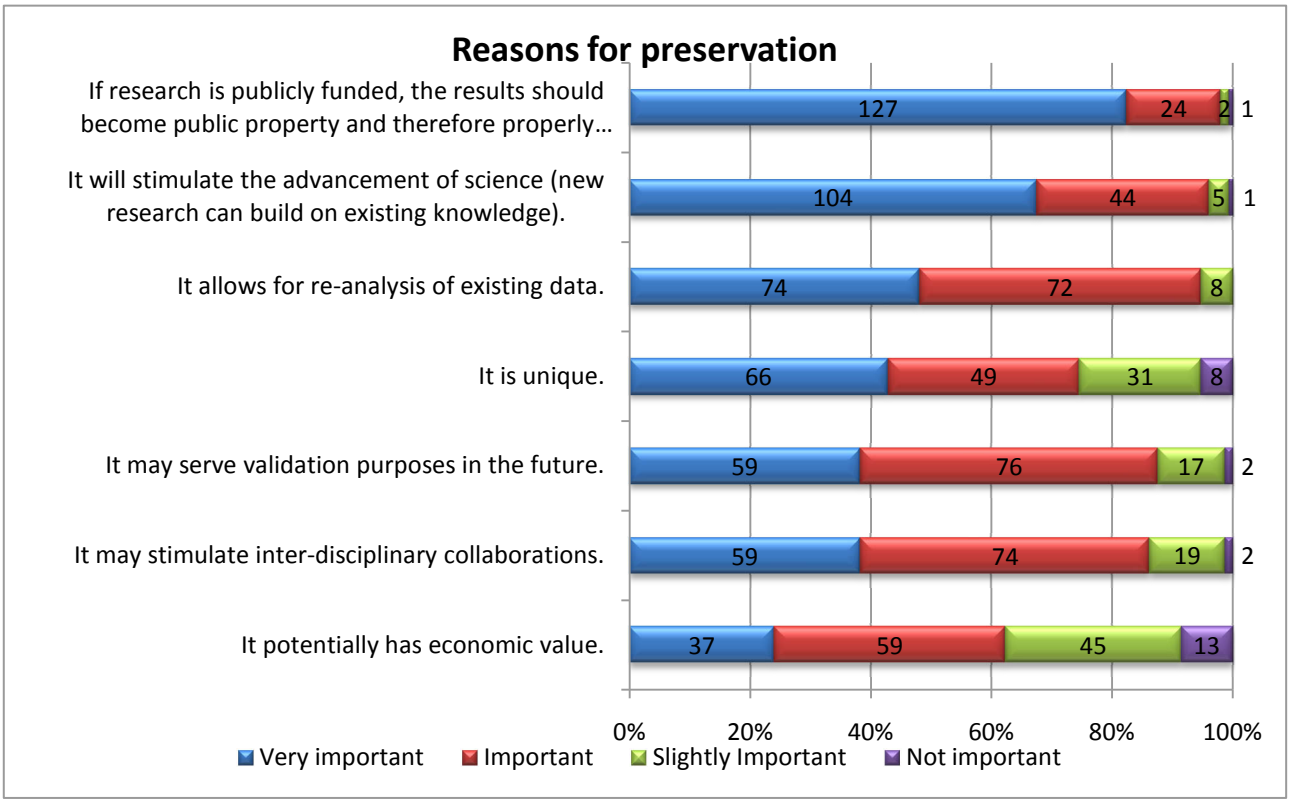


Figure 25: reasons for preservation, n = 154

7.2.2 Threats to Preservation: The View of Data Managers

We also asked data managers about the threats to digital preservation. As stated earlier, seven threats were formulated for this question, similar to the threats used in the projects CASPAR[19] and SHAMAN[20]. The seven threats are:

- Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
- Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
- Evidence may be lost because the origin and authenticity of the data may be uncertain.
- Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future.
- Loss of ability to identify the location of data.
- The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
- The ones we trust to look after the digital holdings may let us down.

For each of these threats respondents were asked to indicate their importance. The choices available were *very important*, *important*, *slightly important*, *not important*, or *don't know*.

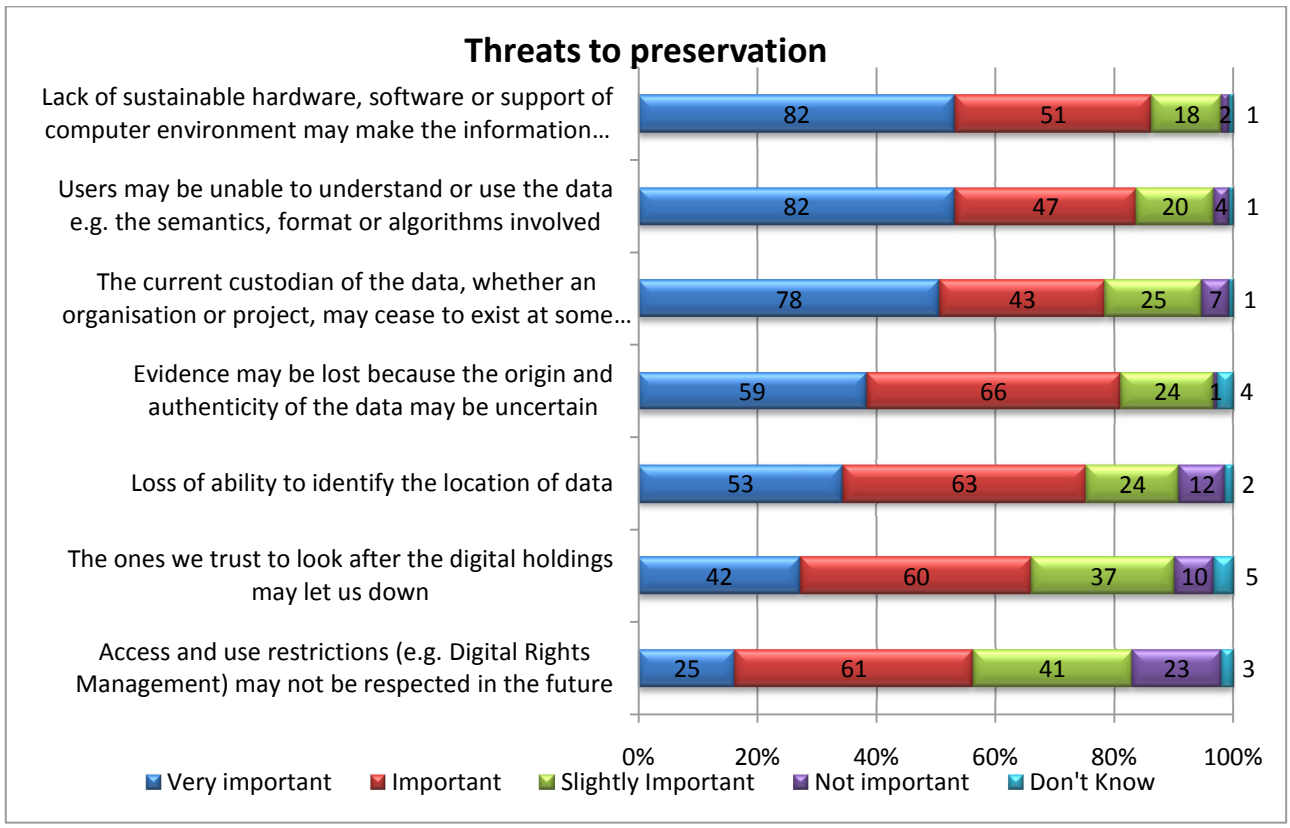


Figure 26: threats to preservation, n=154

[19] EU FP6 project CASPAR: http://www.casparpreserves.eu/
[20] EU FP7 project SHAMAN: http://shaman-ip.eu/shaman/

The difference with the researchers on this question is minor. Data managers are also highly aware of the major threats to long-term preservation of digital research data. The top three of threats which the respondents regarded either *important* or *very important* are: lack of sustainable hardware, software or support (86%); problems with understanding the semantics, formats or algorithms of data (83%); uncertain origin and authenticity (81%); Of all formulated threats access and use restrictions (56%) are regarded as the least important threat to preservation.

7.2.3 The Need for an Infrastructure

60% of respondents believe that an international infrastructure for data preservation and access should indeed be built to help guard against some of the above-mentioned threats. When asked what such an infrastructure should look like, data managers are also not always sure. Analysing the open text answers showed us that data managers seem to agree with researchers on the idea of building some sort of central international repository/database for researchers. Contrary to researchers though, a larger proportion of respondents believe in a national solution. Figure 28 shows a tag cloud based on the open text answers.

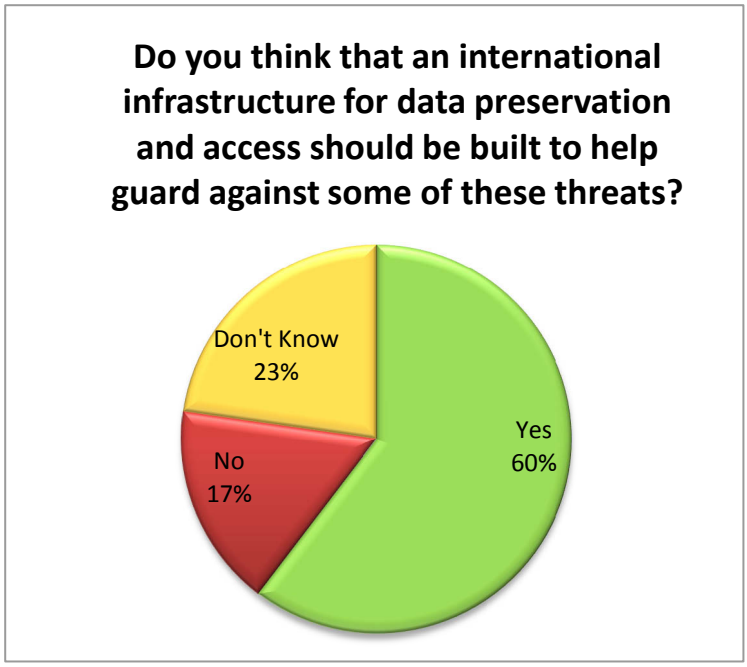


Figure 27: need for an infrastructure, n = 154

Figure 28: what should an infrastructure look like? n = 154

Apart from an infrastructure data managers think that more resources (86%) and more knowledge (82%) is necessary to guarantee long-term access and usability of research data. In addition training is also considered to be *important* (68%) (see Figure 29).

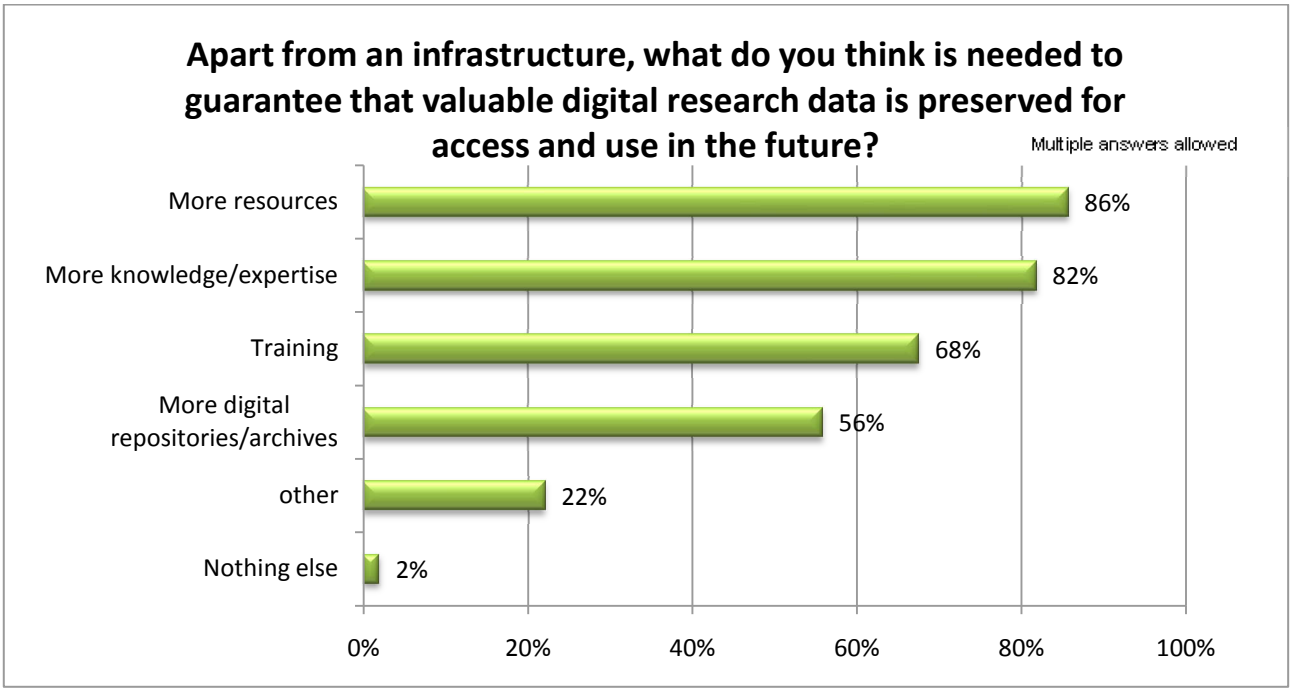


Figure 29: other needs to preserve research data, n = 154

7.3 Preservation – the state of affairs

To be able to determine what is needed for the preservation of research data, we asked researchers about their day-to-day work practice. It is just as important to look at it from the side of the data managers. What are their practices currently? What kind of digital data do they curate? How much is stored/preserved/curated currently? What policies do they have regarding digital research data? These are the questions which this section deals with.

7.3.1 Kind of digital material

We asked data managers which kind of digital objects are stored at their organisation (Figure 30). Of the options we offered in this multiple choice question, theses (69%) was checked most often, closely followed by journals and e-journal publications (68%) and illustrative material (62%). More complicated materials such as auxiliary material (27%) and data sets (44%) were chosen less often.

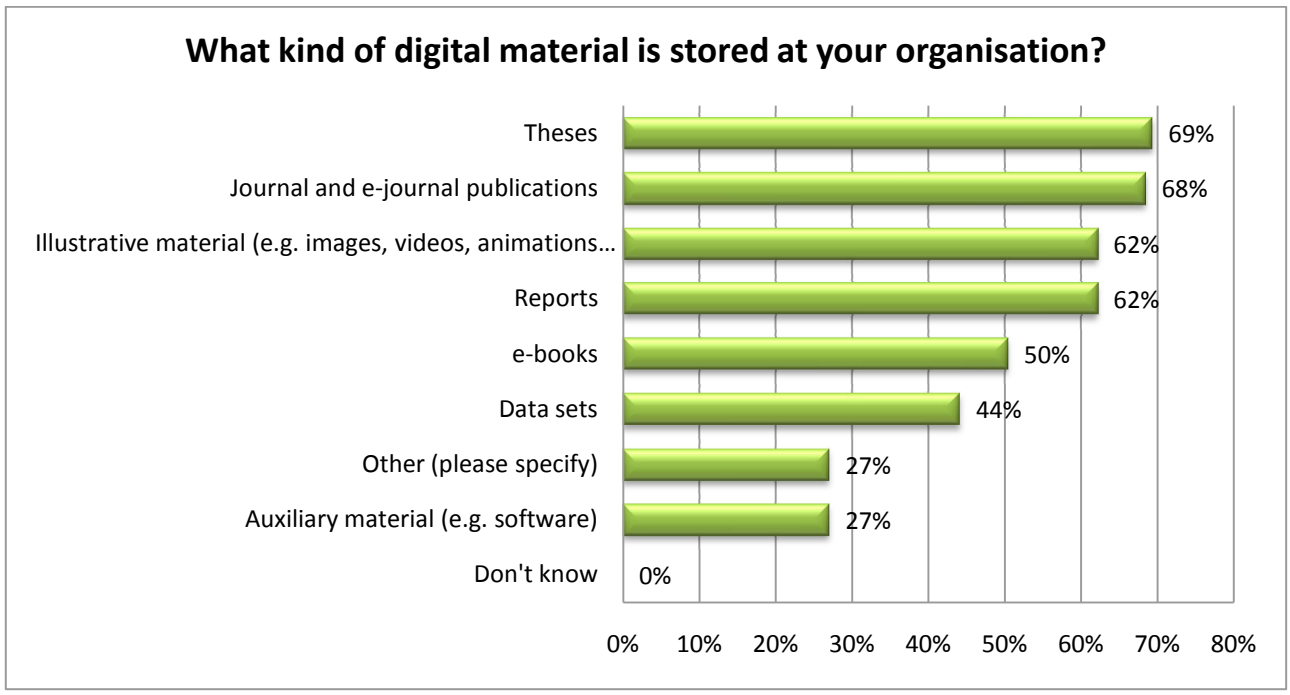


Figure 30: what kind of digital material is stored at your organisation? n = 111

It is one thing to know which kinds of objects are stored at these organisations, but another to know which formats are stored. So, we also asked respondents about the data formats that are currently stored at their organisation. The top three choices are images (81%), office documents (74%), and audiovisual materials (46%) (see Figure 31). As it turns out then there is dissimilarity between what researchers use and data managers store (see chapter 9).

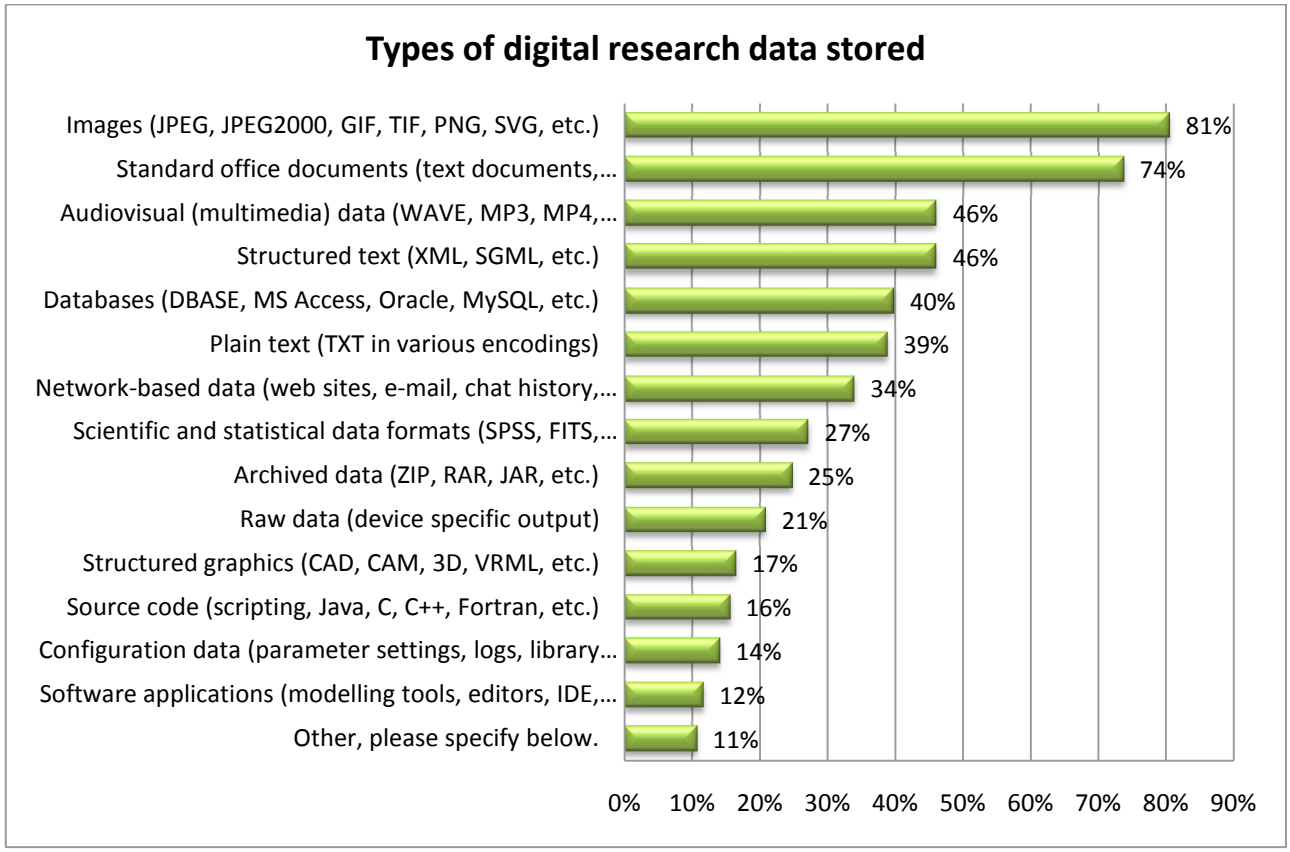


Figure 31: types of digital research data stored, n = 206

Besides the kind of data formats data managers store and manage, we asked them to provide an estimate of the data they currently store and manage as well as the amount of data they think they will store and manage in two and five years respectively. Since data managers are very specifically dealing with data storage and management, we expected that data managers would have a better idea of the amount of data they currently store and manage. Yet, the percentage of respondents to the data managers' survey who don't know how much their organisation currently stores is 17%, which is higher than the percentage of researchers who don't know (10%). The percentage grew to 21% for the estimate of stored data in two years and to 28% for the estimate of stored data in 5 years.

For data managers the break seems to appear around 1TB, in that more respondents estimate they are going to store between 1TB and 1PB of data in two years than in five years. Yet the chart also shows that the percentages for the data ranges up to 1TB are declining with the years, while those larger than 1TB are increasing with the years. Within five years 24% of the respondents estimate to store 1 PB or more, while 48% of the respondents estimate that their stored data amounts to less than 1PB—28% don't know.

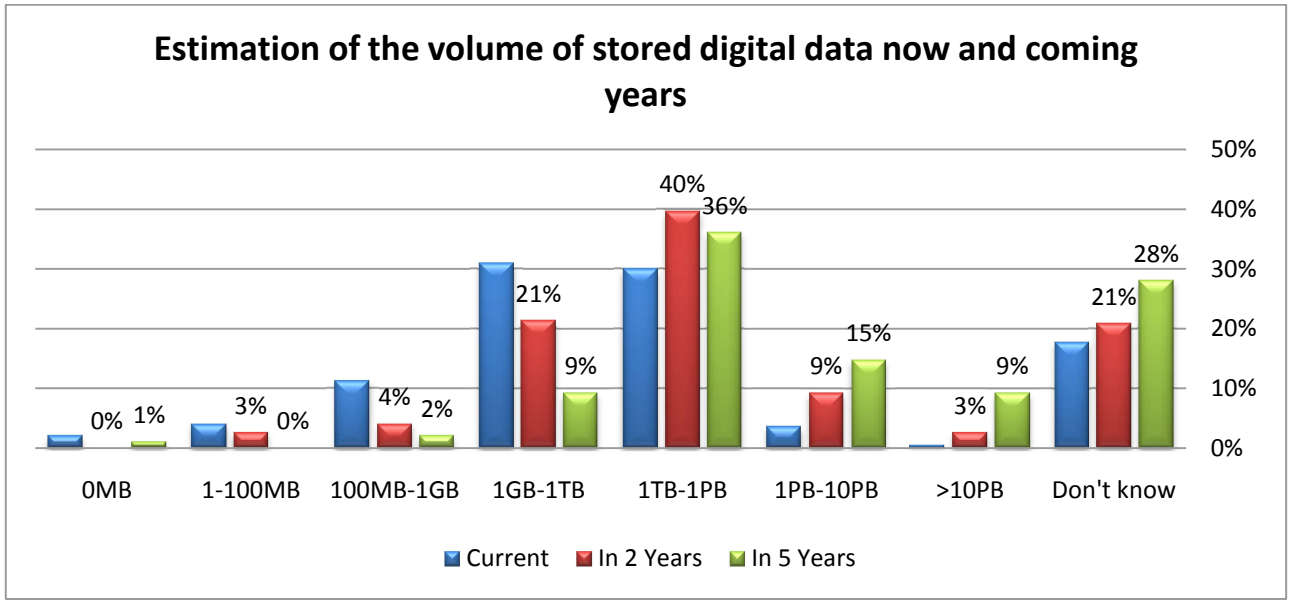


Figure 32: estimation of the volume of stored digital data now and over the next years, n = 197

7.3.2 Policies and Procedures

To know what and how much is stored by the data managers who responded to our survey tells us little about how data is stored. What procedures do data managers follow? Have their organisations formulated policies for the curation and storage of data? To find out more about these issues, the respondents were asked to confirm whether their organization has policies for a number of issues related to the storage and management of data.

When asked whether their organisations have policies and procedures in place which determine what kinds of data is accepted for storage/preservation and how and when it needs to be submitted,

64% of the respondents answered affirmative. Nevertheless this means that 32% stated that they don't have such policies.

Digging a little deeper, we also asked respondents what these policies entail. We specifically wanted to know whether the policies and procedures included:

- Selection criteria regarding what to submit/accept
- Requirements regarding standard formats
- Information about copyrights of data submitted
- The way in which data is submitted
- Responsibilities for data storage and management
- Liability when data is lost or affected

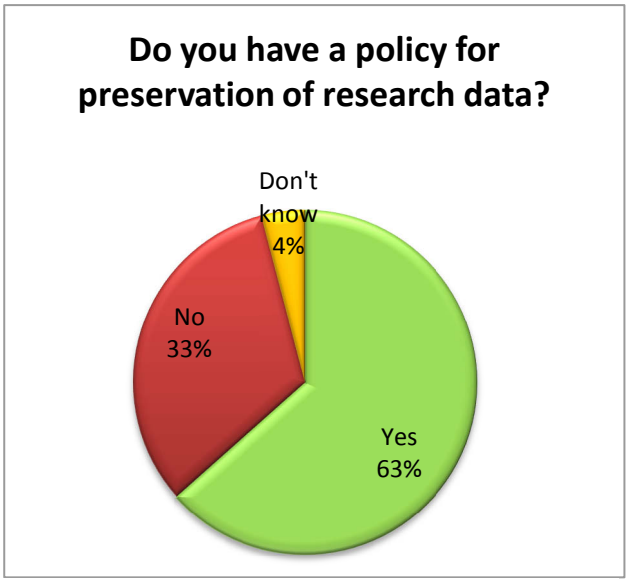


Figure 33: do you have a policy for preservation of research data? n = 197

The majority of respondents have policies for submitting data which include most of the above-mentioned criteria. The only real exception is liability. Only 34% of the respondents stated that their organisation has policies which include arrangements for liability when data is lost or affected (see Figure 34). From a preservation point of view the 23% of respondents who answered that their organisation does not include in their policies the way in which data is submitted is still quite large. This may be concerning if one believes that, just as researchers should be conscious of preservation of data when the data is created, so data centres and archives should be aware of requirements for preservation when data is submitted.

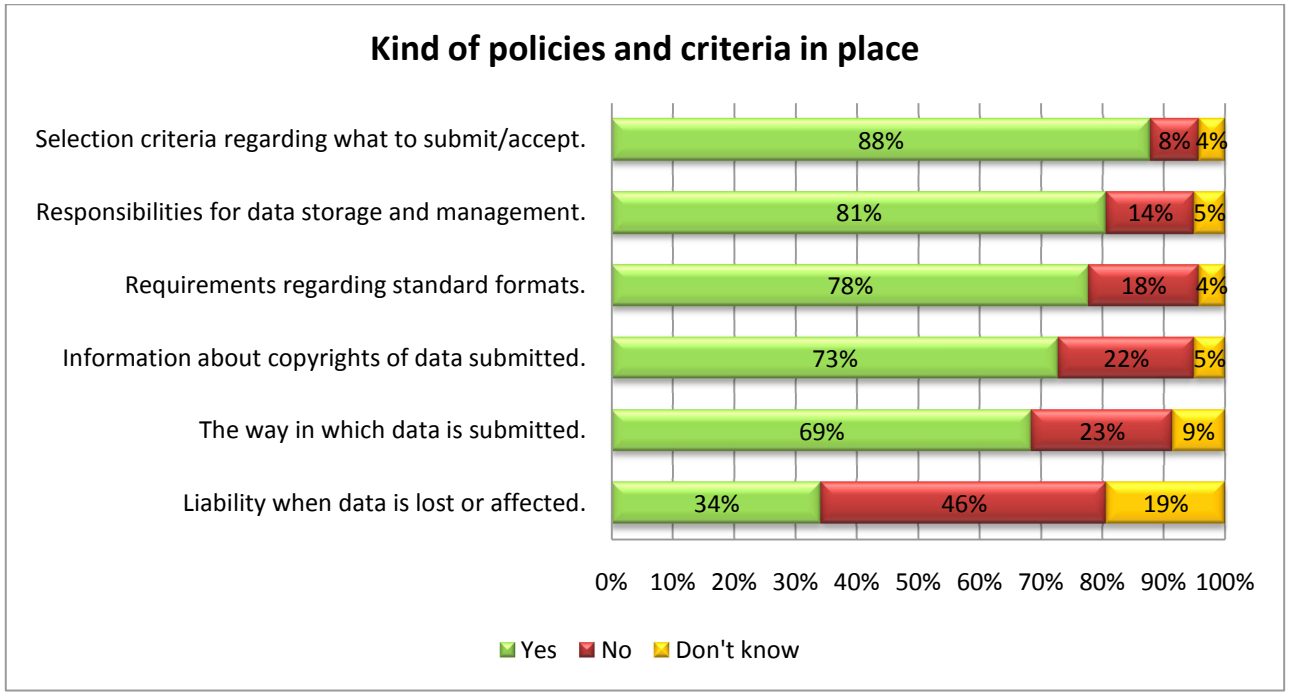


Figure 34: kind of policies and criteria in place, n = 140

When asked whether the organizations of the respondents have policies and an infrastructure to guarantee that data are properly managed and maintained to ensure continued access and usability, 62% answered yes.

Regarding the protection of the data's authenticity, the vast majority (73%) of the respondents stated that they do not have policies in place which require those who submit data to show who has previously enhanced, annotated or had access to the data (Figure 35).

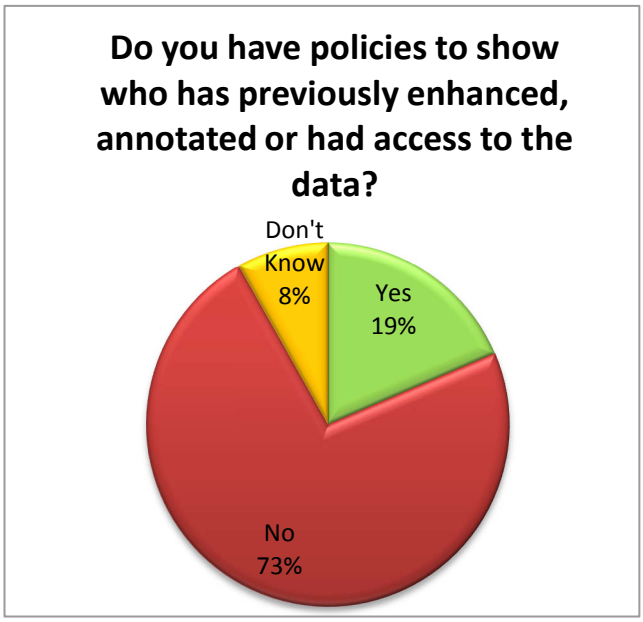


Figure 36: policies for keeping track of changes to data, n = 172

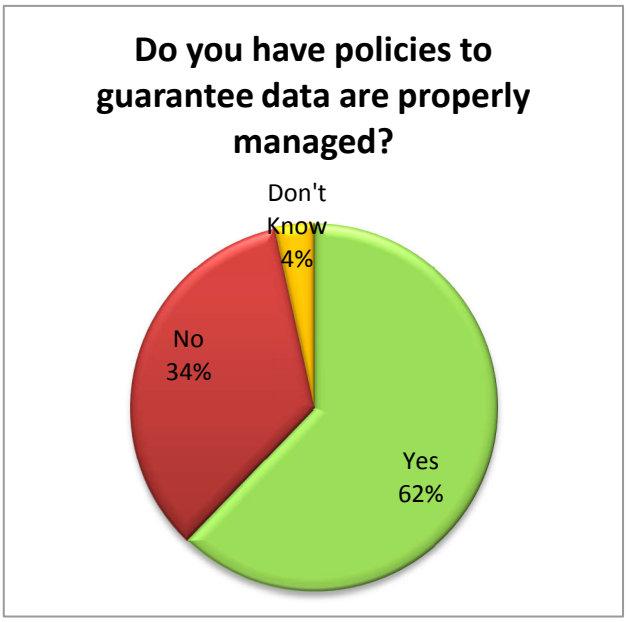


Figure 35: policies to guarantee data are properly managed, n = 172

Once the data has been submitted, however, most data management organisations (72%) do have security protocols that protect stored data from unauthorized modification, damage or deletion (see Figure 36).

7.3.3 Data linking

The context in which research data are accessed and used also determine the data's meaning. In this respect it is important to be able to access data that has been used for a specific journal article. Therefore, we asked data managers whether it is possible for users of the data stored at their organisation to link to that data when referencing it in a journal (see Figure 37). A small majority of 54% states that it is indeed possible within their organisations.

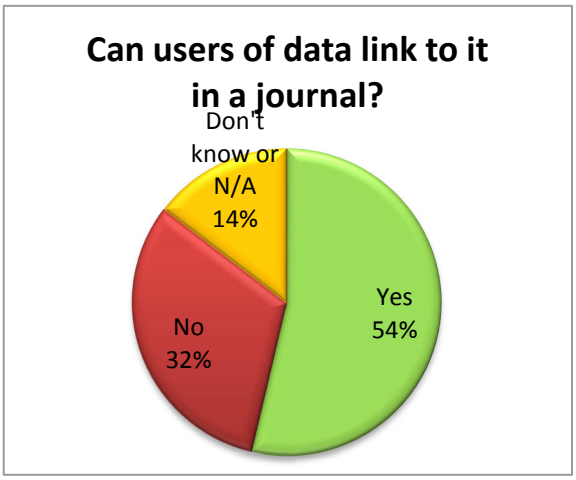


Figure 37: can users of data link to it in a journal? n = 172

7.4 Preservation – the outlook

Looking towards the future, it becomes clear that not all organisations are confident they are prepared for the future preservation needs and requirements. For instance, a majority of 59% of the respondents to our data management survey don't think that the tools and infrastructure available to them suffice for the digital preservation objectives they have to achieve (see Figure 38).

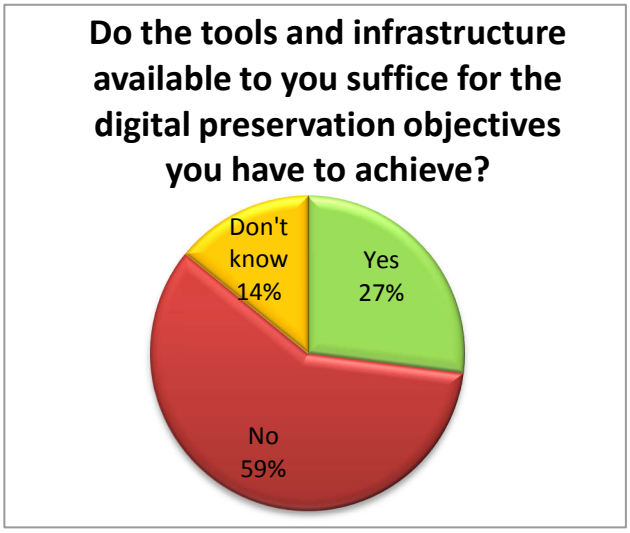


Figure 39: do the tools and infrastructure suffice for the preservation objectives you have to achieve? n = 164

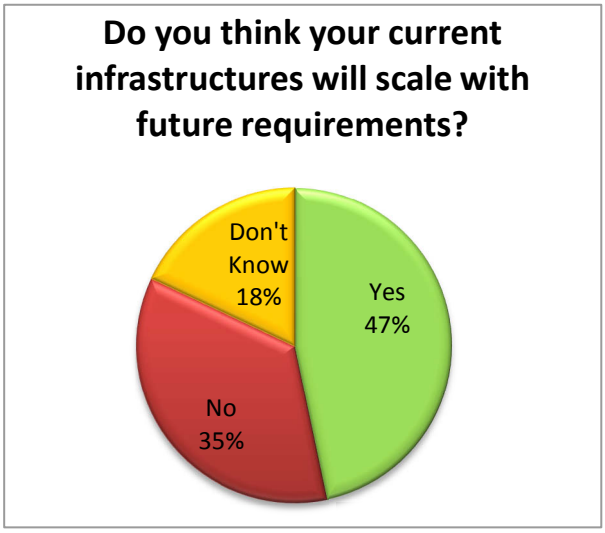


Figure 38: do you think your current infrastructure will scale with future requirements? n = 167

When asked whether they think if their current infrastructures will scale with future requirements, 47% responded yes, while 35% stated that they don't believe it will—18% said they don' t know (see figure 39).

7.5 The cross-disciplinary use of research data

Just as researchers may share data for their research, data management organisations can share data or offer combined services to users. Only 46% of the respondents claim their organisations share data or services with other organisations. No less than 50% of the respondents answered that they don't share infrastructures with other organisations.

The current practice may not lean towards sharing, but respondents do believe that there will be a need for sharing resources. When asked a vast majority (89%) state there will indeed be a need for sharing. Only 5% don't think there will be a need, while 6% of the respondents don't know.

7.6 Funding

Currently most digital preservation work is done in short term projects (<5 years), meaning that there is an emphasis on rather short term funding of preservation activities. This may explain why less people are sure whether funding will become an issue for them in 10 or more years (see Figure 40).

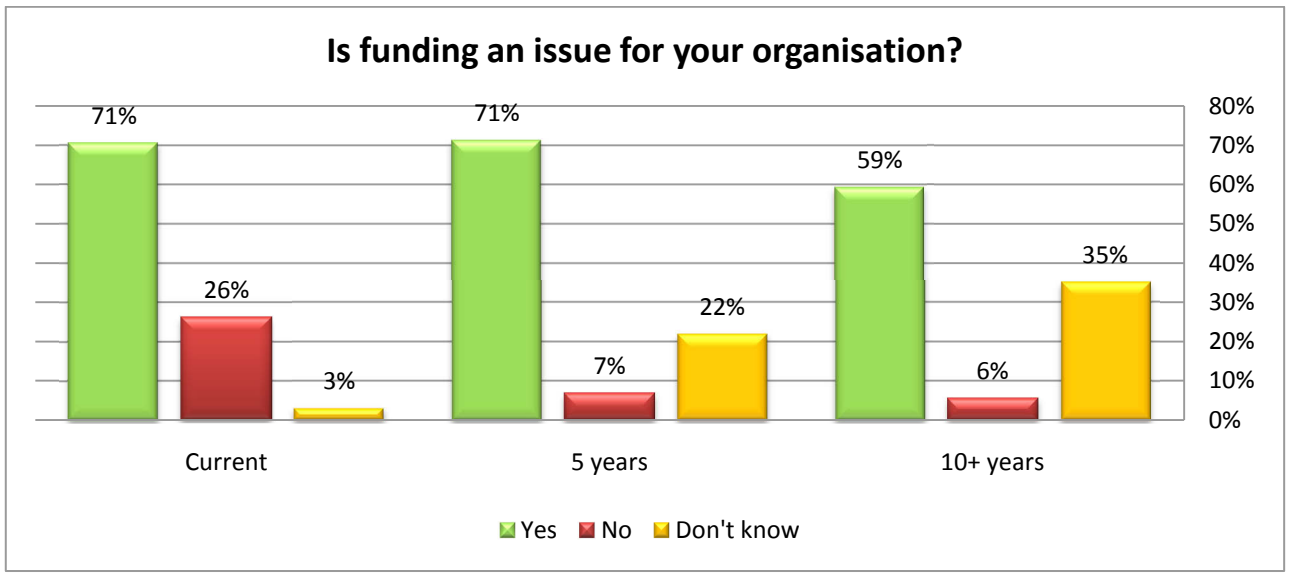


Figure 40: is funding an issue for your organisation? n = 160

7.7 Roles and responsibilities

Digital preservation involves responsibilities and costs. Just like the researchers, data managers had to provide their views on who should be responsible for the preservation of research data and who should pay for it (Figures 41 and 42). According to the data managers of our survey the National Library (71%) is the well-chosen organisation to take on responsibility for preservation of research data. The two other most chosen options are the researcher's institute (60%) and research libraries (56%).

Carrying responsibility for preservation does not immediately imply having to take care of the bill. When asked who should pay for the preservation of these data, data managers agree that public bodies such as the government should fund the preservation of digital research data. As shown in Figure 42, the top three choices are government (77%), research funders (51%), and the EU (42%).

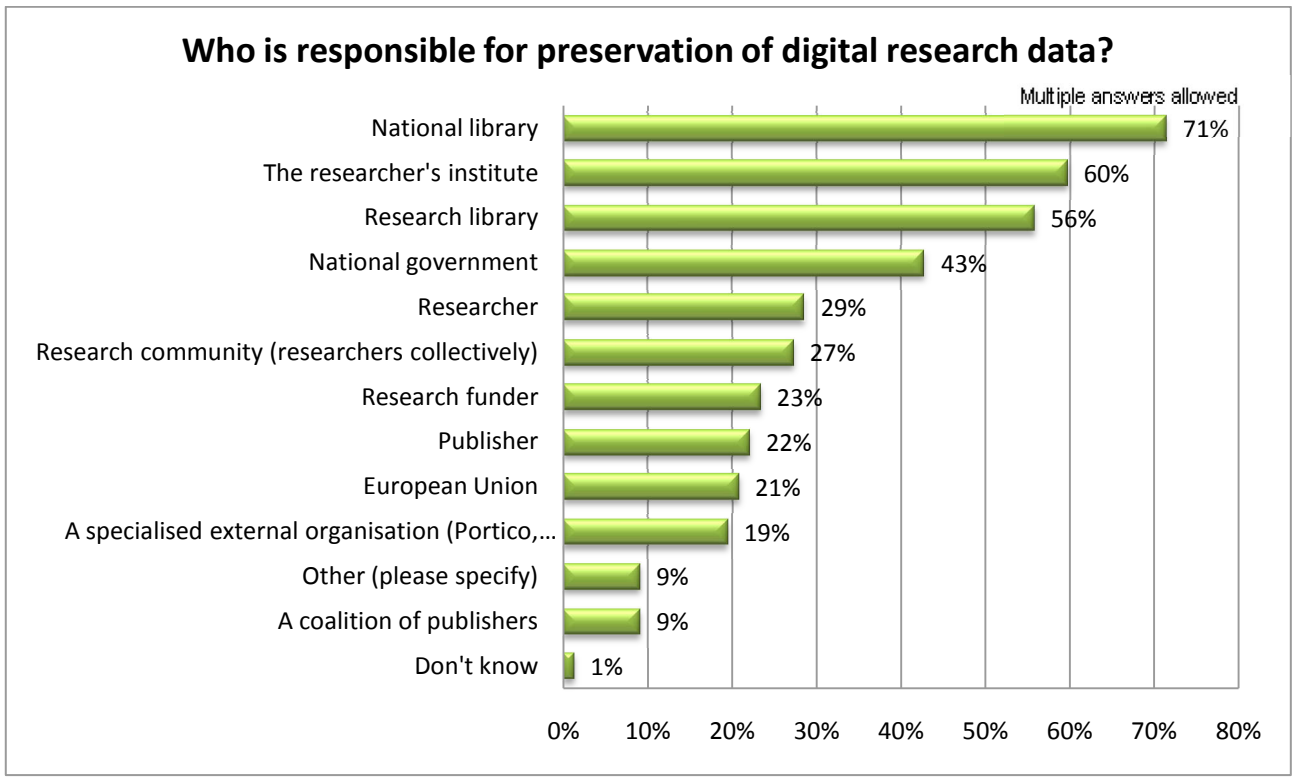


Figure 41: who is responsible for preservation of digital research data? n = 77

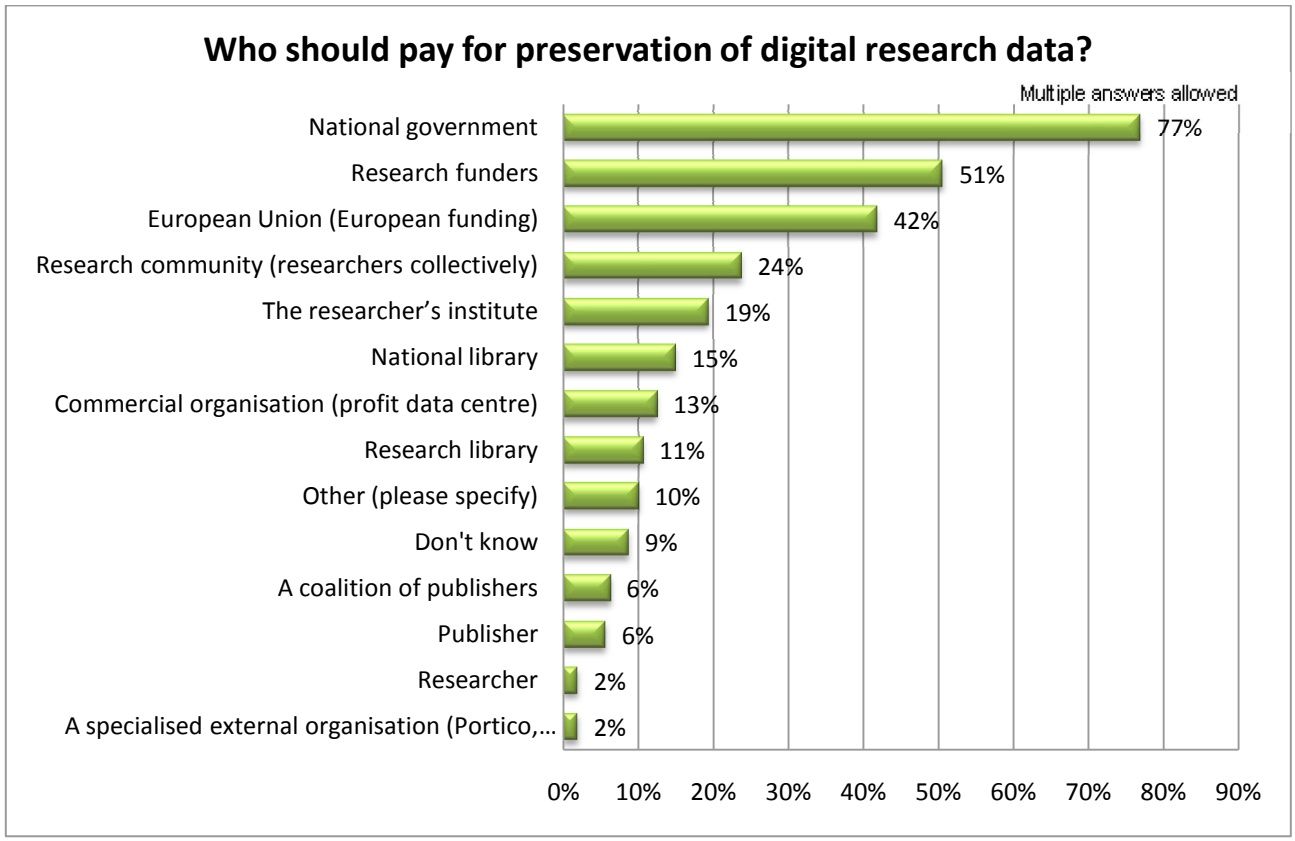


Figure 42: who should pay for preservation of digital research data? n = 160

8 Publishers

8.1 Introduction

The two most important distribution channels for the Publishers survey were the International Association of Scientific, Technological and Medical Publishers (STM) and the Directory of Open Access Journals (DOAJ). In addition—like with the other stakeholder surveys—this survey was sent as part of the merged survey to several mailing lists.

A selective number of 290 individuals of the STM mailing list received an invitation to respond. The STM part elicited 59 responses of which 40 completed the survey and 19 filled out the survey partially. Yet some organisations responded more than once. Based on the help from the STM association we selected the responses which best represented the company in question. To make sure we did not count responses more than once we filtered out the remaining responses. This left us with a total list of 43 unique responses. The respondents represent 43 different publishing companies, which is over 40 % of the organisational members of the STM list and more than 50% of the member-publishers of the STM Association. In total, these publishers collectively publish approximately 8,800 journals, or roughly 97% of all journals covered in the survey.

The DOAJ mailing consists of approximately 2,000 small open access publishers. 127 people responded to our general call, representing an equal number of open access publishers. 97 respondents completed the survey. The total number of journals published by these Open Access publishers is approximately 250, or roughly 3% of the number of journals covered in this survey.[21]

The total number of publishers who responded is a bit larger, because in addition to the STM and DOAJ surveys the merged surveys elicited another 8 responses; however only 1 respondent completed the questionnaire.

Thus, when removing the double responses in the STM list, the total number of publishers who responded is 178. Of these people 138 respondents completed the survey. Again, since we do not know the total number of people who received an invitation through the different mailing lists, it is difficult to give an exact total response rate.

But more important for the significance of the response rates is that the responses to this survey cover a significant portion of all peer-reviewed scholarly journals being published worldwide. The overall estimate is that at present approximately 25,400 peer-reviewed scholarly journals[22] are be-

[21] We do not know the exact number of journals published by each individual publisher. Since the DOAJ has roughly 2,000 publishers in its collection and a little bit over 4,000 journals, we estimate the average number of journals for the open access publisher represented in this survey at 2 journals per publisher.

[22] http://www.stm-assoc.org/about.php?PHPSESSID=5a0ce8c1d23246500dd5a6fc3042ea99. Carol Tenopir, renowned scholar on scholarly publishing, estimates the total number of scholarly journals (not only peer-reviewed) at just under 50,000. http://www.libraryjournal.com/index.asp?layout=articlePrint&articleID=CA374956 . It is a little bit more diffi-

ing published. The survey results include responses from all major publishers (Elsevier, Springer, Wiley Blackwell, etc.). Most major publishers have filled out the survey. These responses represent roughly 8,800 of the peer-reviewed scholarly journals, or 35% of the market.

In addition to the peer-reviewed journals of the major publishers, the survey also included responses from the open access publishers, collected in the Directory of Open Access Journals. As said, these publishers roughly represent 250 journals, but it is unknown whether these are all peer-reviewed journals. Mark Ware estimates that about 10% (app. 2,540 journals) of all peer-reviewed journals are open access. Even if we assume that all these journals are peer-reviewed, it is safe to say that the DOAJ journals in the survey responses represent less than 10% of the market of peer-reviewed open access journals.[23]

8.1.1 Country of Respondents

As with the researchers and data managers, we asked publishers for the country of their organisations. The question here is probably less useful as a basis for further analysis. Most STM publishers are large globally operating companies with global policies regarding the preservation of journal articles. This may be different for the publishers operating solely on an open access business model. Compared to the large STM organisations, these open access publishers are usually very small and not operating under global policies[24] despite often being members of larger representative bodies such as the Directory of Open Access Journals and the Open Access Scholarly Publishers Organisation (OASPA).

When all surveys (STM, DOAJ and merged) are combined the majority of responses for the publishers came from Europe. The same goes when we limit our focus to the DOAJ respondents. Yet when we look at the top five of countries from which DOAJ publishers responded only two European countries are represented.

Tables 5 through 7 consecutively present the figures for all publishers who responded (including the respondents from the merged surveys), the figures for the DOAJ publishers who responded, and the top five countries of DOAJ respondents.

---

cult to determine the number of open access journals, especially the peer-reviewed. Mark Ware estimates that about 10% of all peer reviews journals are open access.

[23] http://www.stm-assoc.org/about.php?PHPSESSID=5a0ce8c1d23246500dd5a6fc3042ea99. 10% would only be accurate if all journals counted in the responses are indeed peer-reviewed; they probably are not, but we simply were unable to find out.

[24] Many commercial STM Publishers provide open access services for their journals. These are not included here. When mentioning open access journals publishers we refer to publishers who only use the open access model for their operation.

Table 6: geographic spread of respondents amongst country/region (total)

Country/Region	Numbers of respondents	Percentage
EU	97	50%
USA	36	19%
Other	60	31%

Table 7: geographic spread of respondents amongst country/region (DOAJ)

Country/Region	Numbers of respondents	Percentage
EU	55	44%
USA	18	14%
Other	53	42%

Table 8: top 5 DOAJ respondents

Country	Numbers of respondents	Percentage
USA	18	14%
Spain	11	9%
Canada	10	8%
Brazil	6	5%
Croatia	5	4%

8.1.2 Number of Journals Covered by the Survey

When analysing the publishers' responses, it is important to keep the number of journals published per publisher in mind. As stated earlier, some 25,400 peer-reviewed journals are published world-wide, by approximately 2,000 different publishers. But the top-5 of publishers jointly account for more than 6,700 journals, or roughly 25 % of the total. At the other end of the spectrum, there are approximately several thousands of small open access publishers with only one or a few titles on their list.

If we were to only count the publishers, the results get skewed. The response of a publisher like Elsevier, which represents over 2,000 journals, should weigh more in this respect than a small open access publisher with only 2 titles on the list. For example, if Elsevier states to have a preservation policy in place, it means that more than 2,000 journals are covered. If the open access publishers states to have such a policy in place, it may only cover two journals. We made this distinction for the important questions in our survey.

The 10 largest STM publishers[25] publish approximately one third of all journal titles, whereas the remaining publishers are of a much smaller size – with a long tail of several thousands of publishers who publish one or two journal titles only. In the results of this survey we have made the distinction between publishers who publish less than 50 journals and those who publish more. The distinction

[25] In total, we counted 7640 journals in the top 10 of largest publishers.

is based on one of the survey's questions. To distinguish the large publishers from the small publishers we asked respondents for the number of journals they publish.[26]

8.2 Perceptions of preservation

The questions of this section deal with the respondents' perception of preservation issues. Respondents answered questions about what should be preserved and on the (perceived) reasons for preservation; they evaluated the importance of certain threats to preservation and expressed their opinion about the need for an infrastructure to counter the threats. For the publishers the questions were more geared towards publications, but not exclusively. In the evolving market of (e-)publications there is a growing demand for multimedia publishing with more diverse supplementary material, including data. This makes it interesting to have publishers reflect on the role they currently play and what they think their roles will or should be in the future.

8.2.1 What Kind of Materials Should Be Preserved?

We asked publishers which types of digital publications should be preserved in their opinion. The differences between small and large publishers are minor. A vast majority of the respondents believe that research articles should be preserved. For books the difference is larger, but we should keep in mind that not all small publishers publish books.

Publishers are keenly aware that in the digital world the form of the journal is changing into more multimedia formats, so it is perhaps not surprising that 62% of the small publishers and 48% of the large publishers believe it to be important that illustrative materials (e.g. sound, images, video, etc.) are preserved (see Figure 43). These kinds of materials can still be seen as an integral part of the publication. But what about research data and data sets?

It is interesting to notice here that 54% of the small publishers and 44% of the large publishers also think that data sets and auxiliary material should be preserved. While it is easy to see that publishers have a stake in the illustrative materials of their publications, from the survey results it is less clear if publishers think they have a stake in the data as well. It is important then to see what role publishers think they should fulfil in the digital world of publications and data.[27] We will return to these issues in section 8.4.

[26] We also asked publishers for the number of books they publish annually, but since many small (open access) publishers do not publish books, we took the number of journals as the basis for our comparison between large and small publishers. We define small publishers as publishers who publish less than 50 titles.

[27] In the Brussels Declaration STM publishers expressed their belief that "raw research data should be made freely available to all researchers." "Publishers encourage the public posting of the raw data outputs of research. Sets or subsets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars."
http://www.stm-assoc.org/2007_11_01_Brussels_Declaration.pdf.

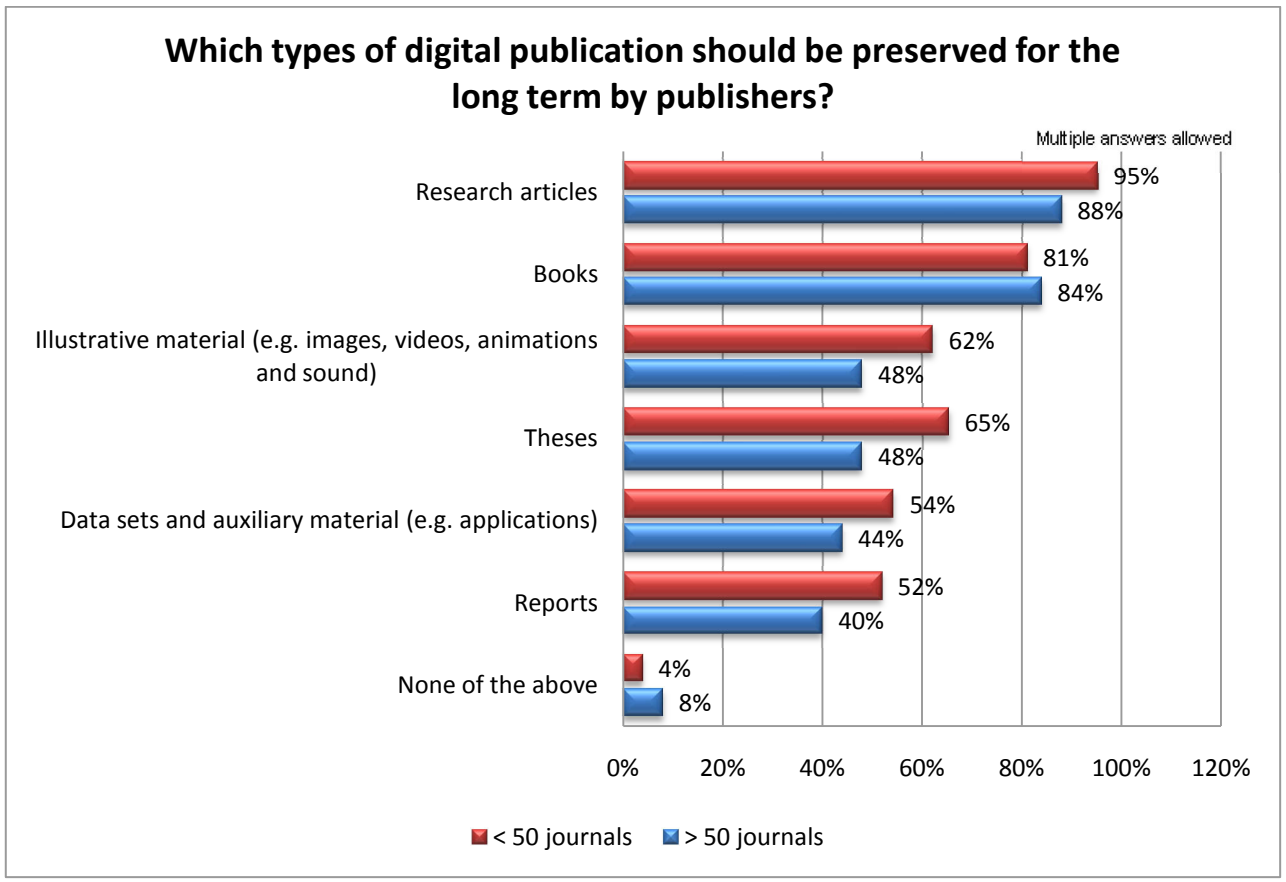


Figure 43: which types of digitals publication should be preserved by publishers? n1 = 127, n2 = 25

8.2.2 What Journal Article Versions Should Be Preserved?

The way research articles are created, published and accessed is rapidly changing. This has an impact on the stages a manuscript passes as well as the versions of a manuscript/article that are created. It was felt that in many respects the old terminology for manuscript/article versions no longer suits the needs of current practices. Therefore, in 2008 the National Information Standards Organization (NISO) in cooperation with Association of Learned and Professional Society Publishers (ALPSP) published a list of recommendation regarding journal article versions.[28]

NISO suggests a list of seven different versions:

- Author's Original (AO)
- Submitted Manuscript under Review (SMUR)
- Accepted Manuscript (AM)
- Proof (P)
- Version of Record (VoR)
- Corrected Version of Record (CVoR)
- Enhanced Version of Record (EVoR)

We adopted the terminology of the NISO recommendations and asked publishers which versions they think should be preserved for the long term. For this question the differences in opinion between small and large publishers is significant.

[28] http://www.niso.org/publications/rp/RP-8-2008.pdf

69% of the large publishers think the *Version of Record (VoR)*, a fixed formally published version of an article, should be preserved. Next to the *Corrected Version of Record (CVoR)* (77%) and the *Enhanced Version of Record (EVoR)* (77%), it belongs to the top three of most important versions to preserve. When looking at the small publishers, it is clear that the percentages are lower than for larger publishers.

Respondents could mark multiple versions and the difference seems to suggest that large publishers were more inclined to select more options than small publishers. Small publishers also regard the VoR as a less important than the large publishers. 33% of the small publishers chose this option. This does not put it in the top three. The most important version to preserve in the eyes of the small publishers is the accepted manuscript. 46% of the small publishers chose this version, followed by EVoR (39%) and the CVoR (35%).
An explanation here could be that the other stages such as VoR and P are often used by large publishers only. Small (DOAJ) publishers might add less additional information to the author's manuscript and therefore do not see any major interest in preserving the other versions.

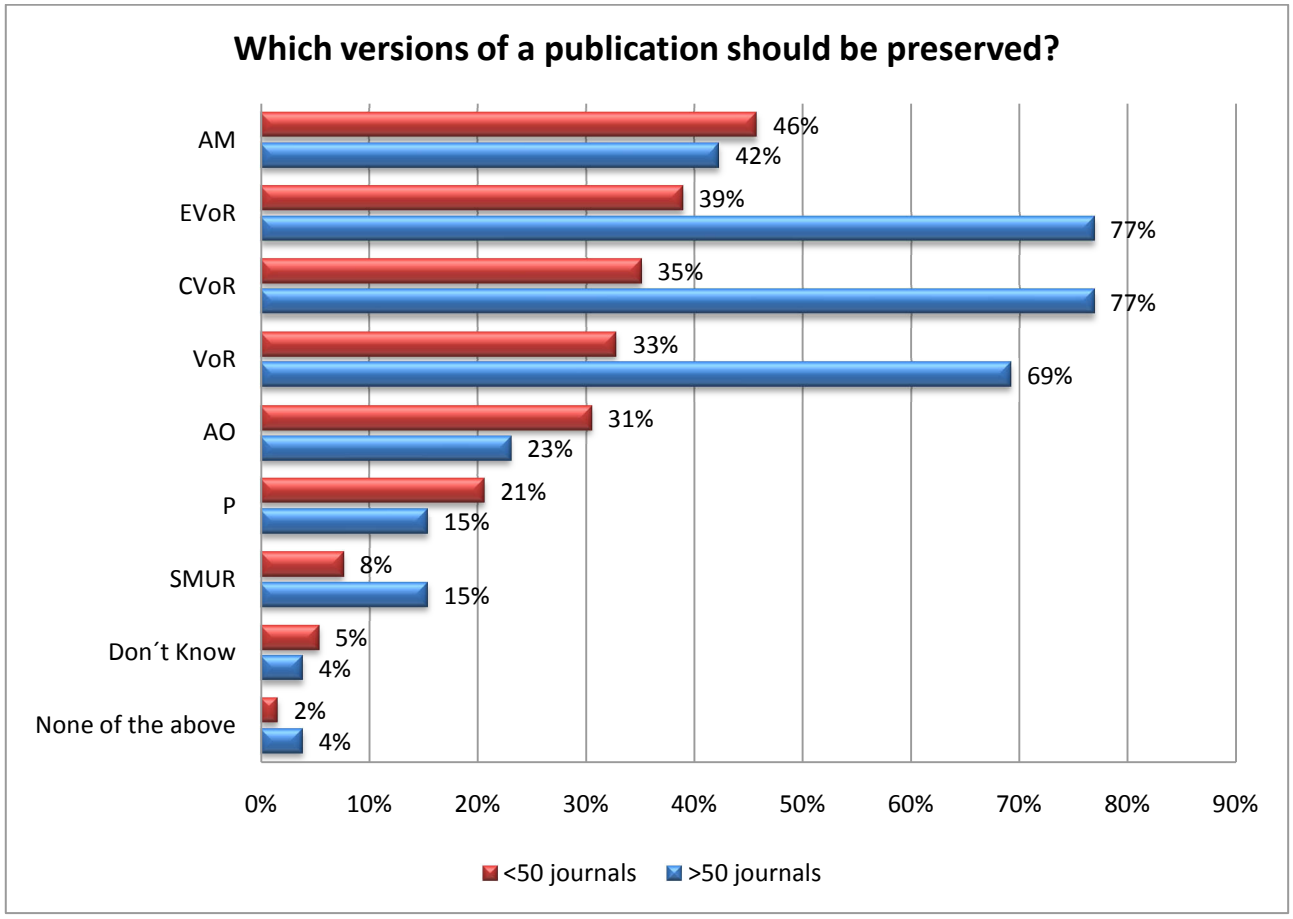


Figure 44: which versions of a publication should be preserved? n1 = 131, n2 = 26

8.2.3 Reasons for Preserving Data

Publishers were presented with the same list as the researchers and data managers and asked whether they regarded the reasons as *very important*, *important*, *slightly important*, or *not important*. The reasons were:

- If research is publicly funded, the results should become public property and therefore properly preserved
- It will stimulate the advancement of science (new research can build on existing knowledge)
- It may serve validation purposes in the future
- It allows for re-analysis of existing data
- It may stimulate interdisciplinary collaborations
- It potentially has economic value
- It is unique

Five out of the seven reasons are regarded as either *important* or *very important* by 76% to 96% by the small publishers (see Figure 45). The most important reasons marked by small publishers are preservation as stimulation for the advancement of science. 96% of the small publishers regarded this either *important* or *very important*. The top three of most important reasons for the small publishers is completed by future validation purposes (92%) and the possibility of re-analysis of existing data (92%).

The larger publishers agree with the smaller publishers on the top three of most important reasons, although the percentages differ. A vast majority of the larger publishers consider the following reasons to be either *very important* or an *important* reasons for preservation (see Figure 46).

- It will stimulate the advancement of science (96%)
- It may serve validation purposes (88%)
- It allows for re-analysis of existing data (96%)

Similarly publishers also agree on the least important reason for preservation. Only 19% of the small publishers and 17% of the large publishers consider economic value as a *very important* preservation reason.

There is therefore little disagreement on the most important reasons for preservation and the least important reason. Yet there are some significant differences between large and small publishers. Perhaps the most striking difference can be found in the publishers' opinion on public funding as a reason for preservation. 62% of the small publishers regard public funding as a *very important* reason for preservation, while only 24% of the larger publishers agree.

Perhaps this difference can be explained by the fact that many small publishers only publish research that was publicly funded, e.g. from their own university. But several disciplines (pharmacy, medicine, chemistry, engineering) are mainly privately funded – these are typical the publications

that are not published through OA. In general large publishers will preserve all publications regardless of private or public funding.

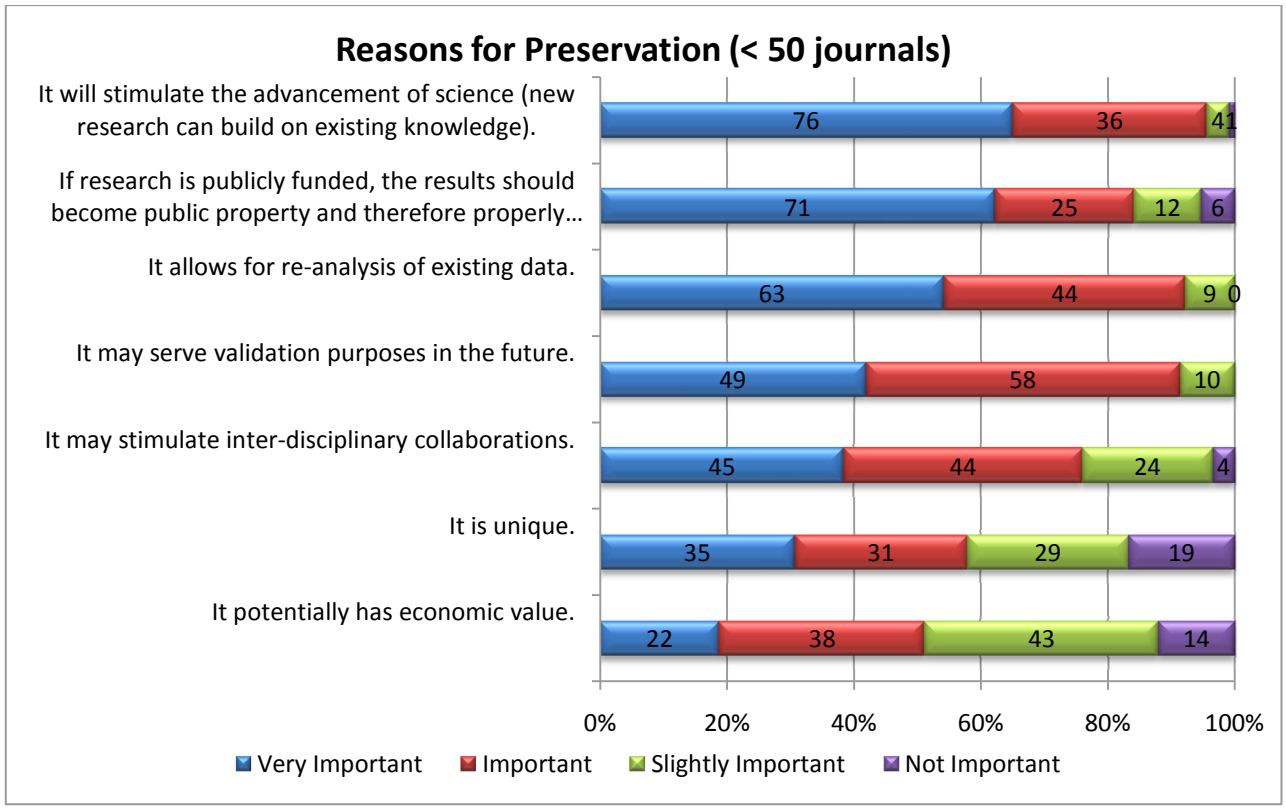


Figure 45: reasons for preservation (< 50 journals), n = 114

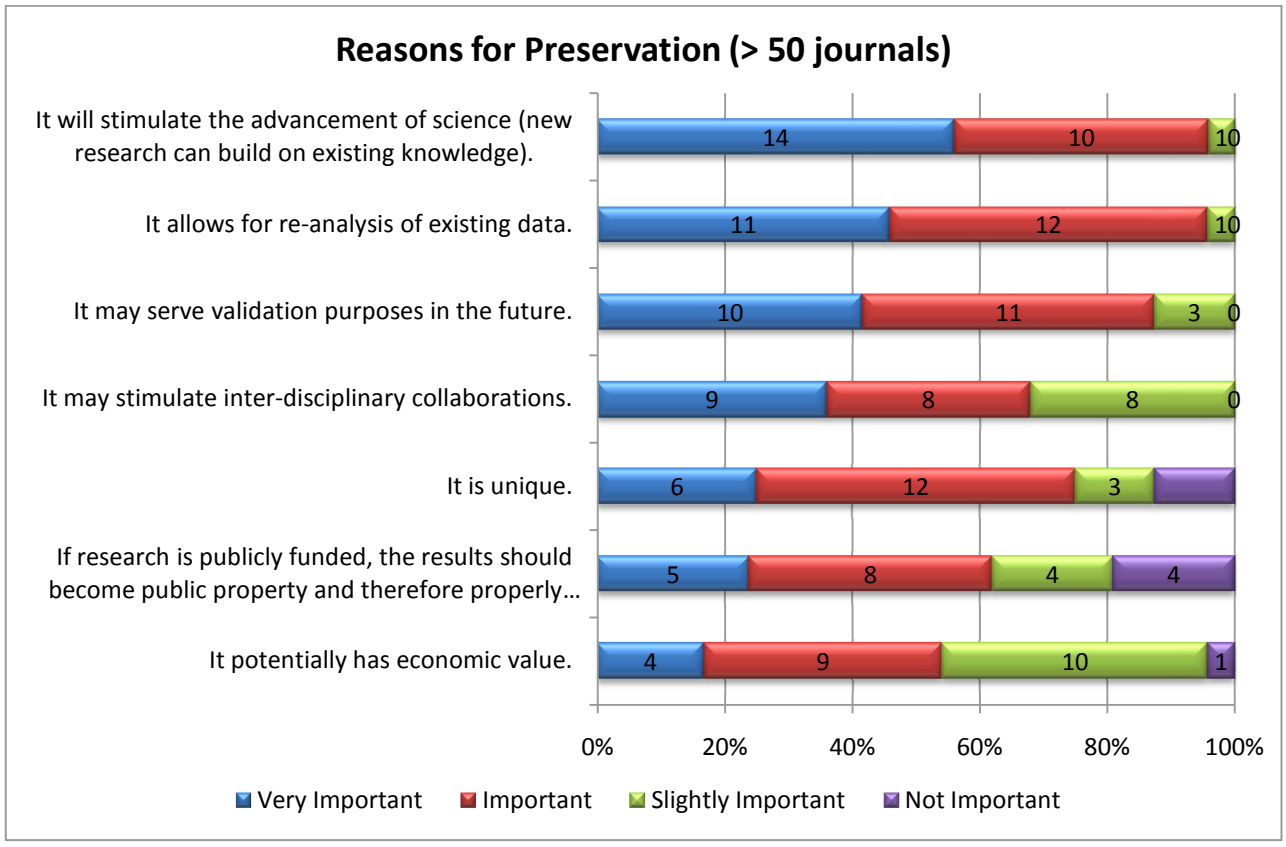


Figure 46: reasons for preservation (> 50 journals), n = 21

8.2.4 The Threats to Digital Preservation

Like researchers and data managers, we asked publishers about the threats to digital preservation. The publishers' survey contained a specific question on seven detailed threats with immediate and direct impact on all digital data.

We formulated seven threats, similar to the threats used in the projects CASPAR[29] and SHAMAN[30]. The seven threats are:

- Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
- Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
- Evidence may be lost because the origin and authenticity of the data may be uncertain.
- Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future.
- Loss of ability to identify the location of data.
- The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
- The ones we trust to look after the digital holdings may let us down.

For each of these threats respondents were asked to indicate their importance. The choices available were *very important*, *important*, *slightly important*, *not important*, or *don't know*.

There is a little disagreement between large and small publishers on the most important threats to digital preservation (see figures 47 and 48). When looking at the data the following threats are regarded by small and large publishers alike as either *important* or *very important*. 78% of the small publishers fear the sustainability of data when the current custodian of the data ceases to exist in the future. For large publishers this percentage is even 80%. Both equally (72%) fear that the lack of sustainable hardware, software or support of computer environment may make the information inaccessible. To round off the top three of threats, 72% of the small publishers and 68% of the large publishers consider the loss of ability to identify the location of data as either an *important* or *very important* threat to digital preservation.

Small and large publishers do not reply similarly. The most noticeable difference of opinion is apparent in their response to the alleged threat access and use restrictions may pose to the digital preservation of data. 61% of the small publishers and 44% of the large publishers believe this threat to be either *important* or *very important*.

[29] EU FP6 project CASPAR: http://www.casparpreserves.eu/
[30] EU FP7 project SHAMAN: http://shaman-ip.eu/shaman/

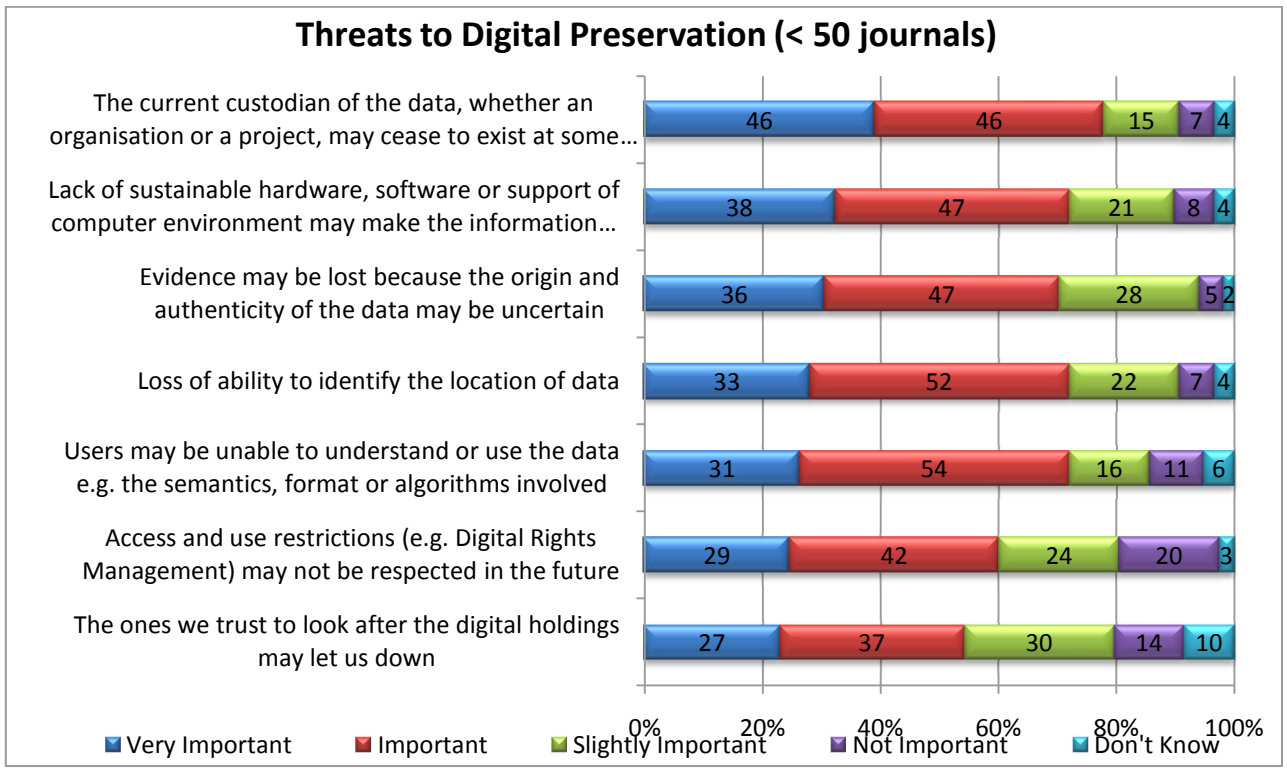


Figure 47: threats to digital preservation (< 50 journals), n = 118

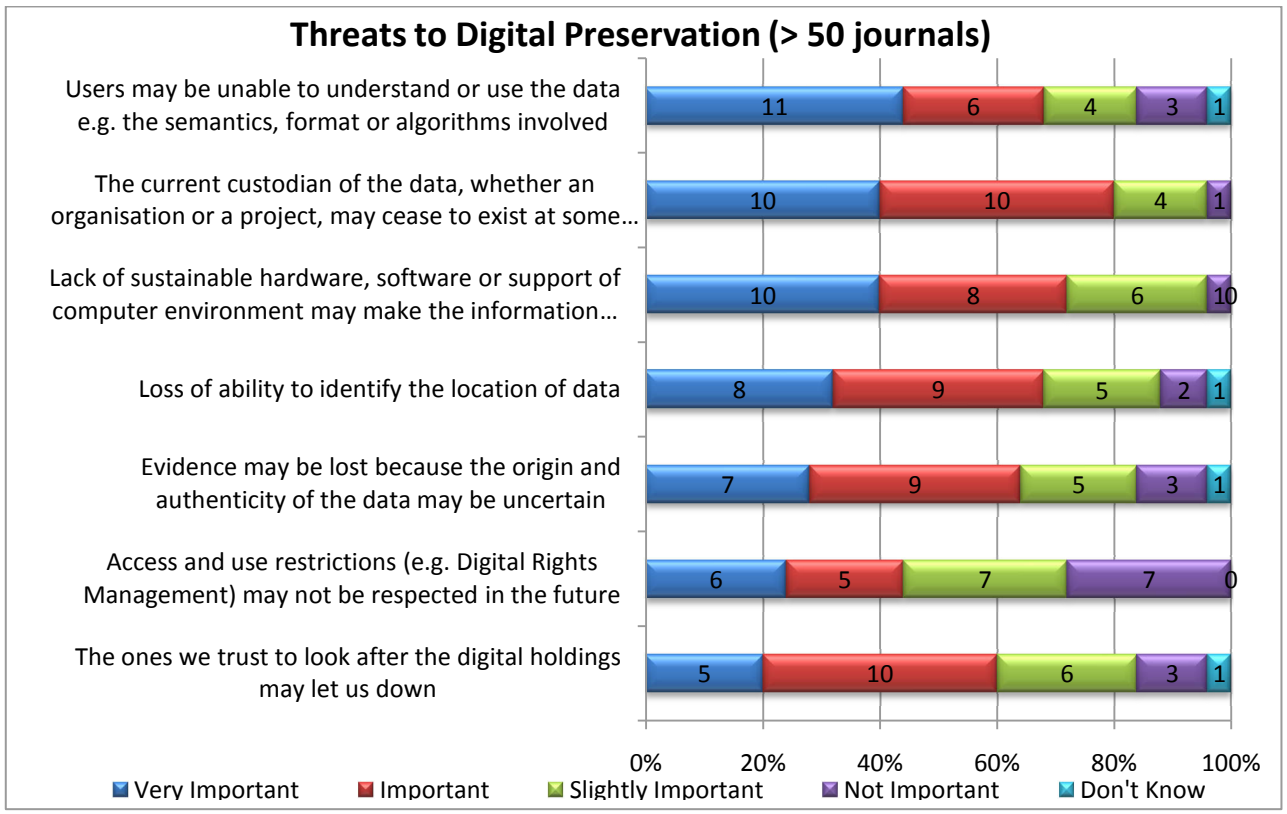


Figure 48: threats to digital preservation (> 50 journals), n = 25

8.2.5 The need for an Infrastructure

As for the other stakeholders, we wanted to know the opinion of publishers on the e-science infrastructure as a solution to preservation. More small publishers (75%) than large publishers (64%) are convinced that some kind of international infrastructure for data preservation and access should indeed be built to help guard against some of the above-mentioned threats. A significant percentage of the respondents simply don't know whether an e-infrastructure will be a guard against the threats to digital preservation.

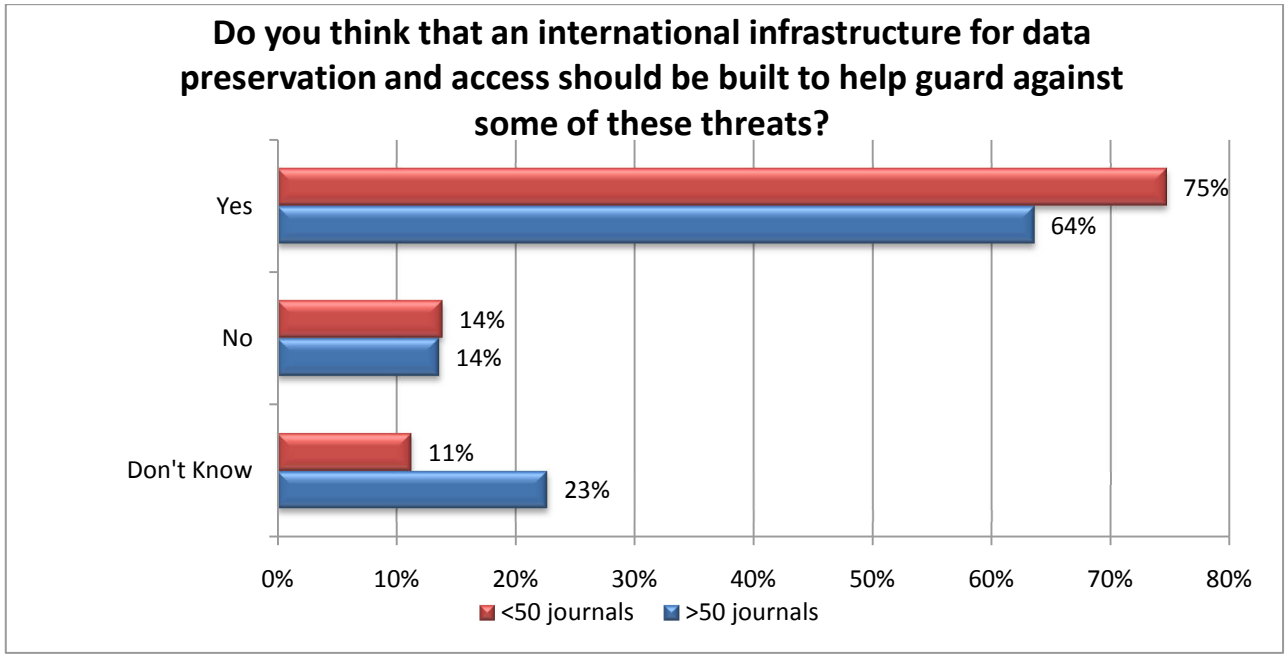


Figure 49: the need for an infrastructure, n1 = 115, n2 = 22

We asked those who answered yes if they could provide us with an idea of what such an e-infrastructure should look like. Here is a tag cloud of the free text answers from that question (Figure 50).

Figure 50: tag cloud of what an infrastructure should look like

Comparing these to the answers of data managers and researchers, it becomes clear that publishers more often mentioned “libraries” and “standards” and “access” as part of the infrastructure solution.

8.3 Preservation – the state of affairs

To be able to determine what is needed for the preservation of research data, we need to know more about the current practices of research stakeholders. This section focuses on those practices for publishers. What is the current practice of publishers regarding research data? Can authors, for instance, submit underlying research data together with their manuscripts? What kind of data do publishers actually accept? Do they have preservation policies for these data?

8.3.1 Can Authors Submit Underlying Research Data?

Before being able to determine what publishers do with research data in terms of preservation, we need to know whether publishers accept underlying research data and what kind of data they accept. So, we asked publishers whether authors can submit their underlying research data with their publication.

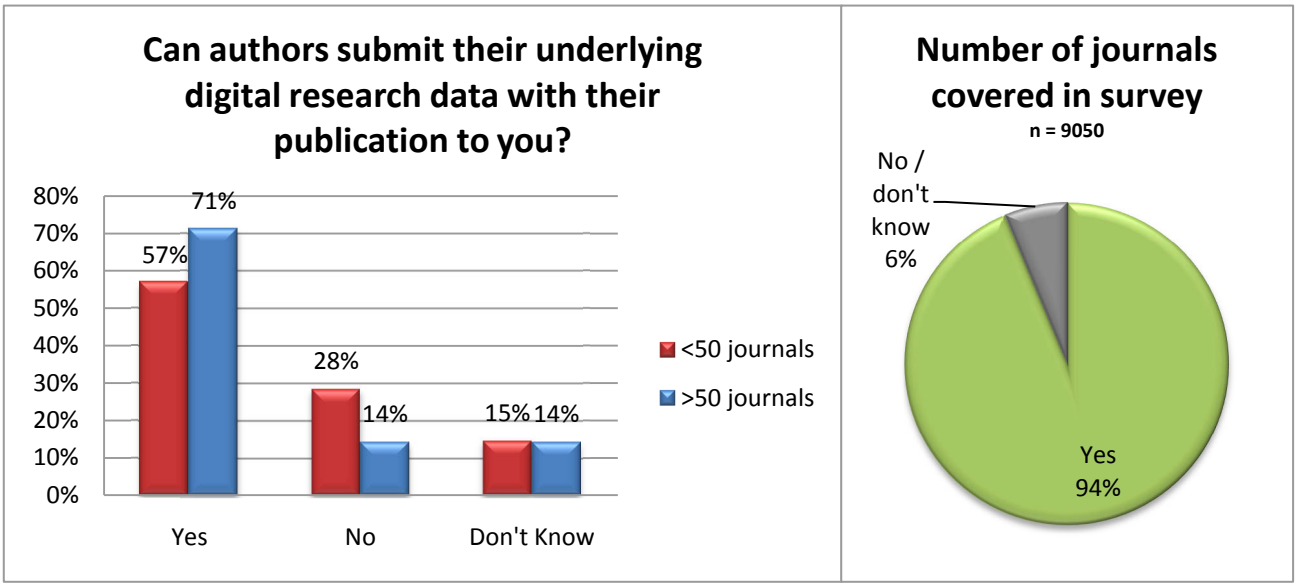


Figure 51: can authors submit their underlying digital research data with their publication to you? n1 = 137, n2 = 35

Perhaps not surprisingly the percentage of large publishers (71%) who allow authors to submit underlying research data is higher than the percentage of small publishers (57%) who do so. This may be a reflection of a difference in service levels to handle research data. In addition, roughly 20% of those who do not yet accept digital research data mentioned that they plan to do so within 5 years.

When expressing these percentages in number of journals, it follows that the large publishers publish 7,730 journals that allow researchers to submit underlying research data to the journal, while the small publishers publish 746 journals that allow for this. So, in total 8,476 journals allow sub-

mission of research data with the manuscript. This represents roughly 94% of the journals covered in the publishing survey.

8.3.2 What Kind of Data do publishers accept?

Each type of data has its own set of characteristics and therefore requires a distinctive preservation strategy. To determine the kinds of data currently accepted by publishers a list of data types was formulated. This list is identical to the list we presented to researchers. Respondents were asked to check the data types they accept (see figure 52).

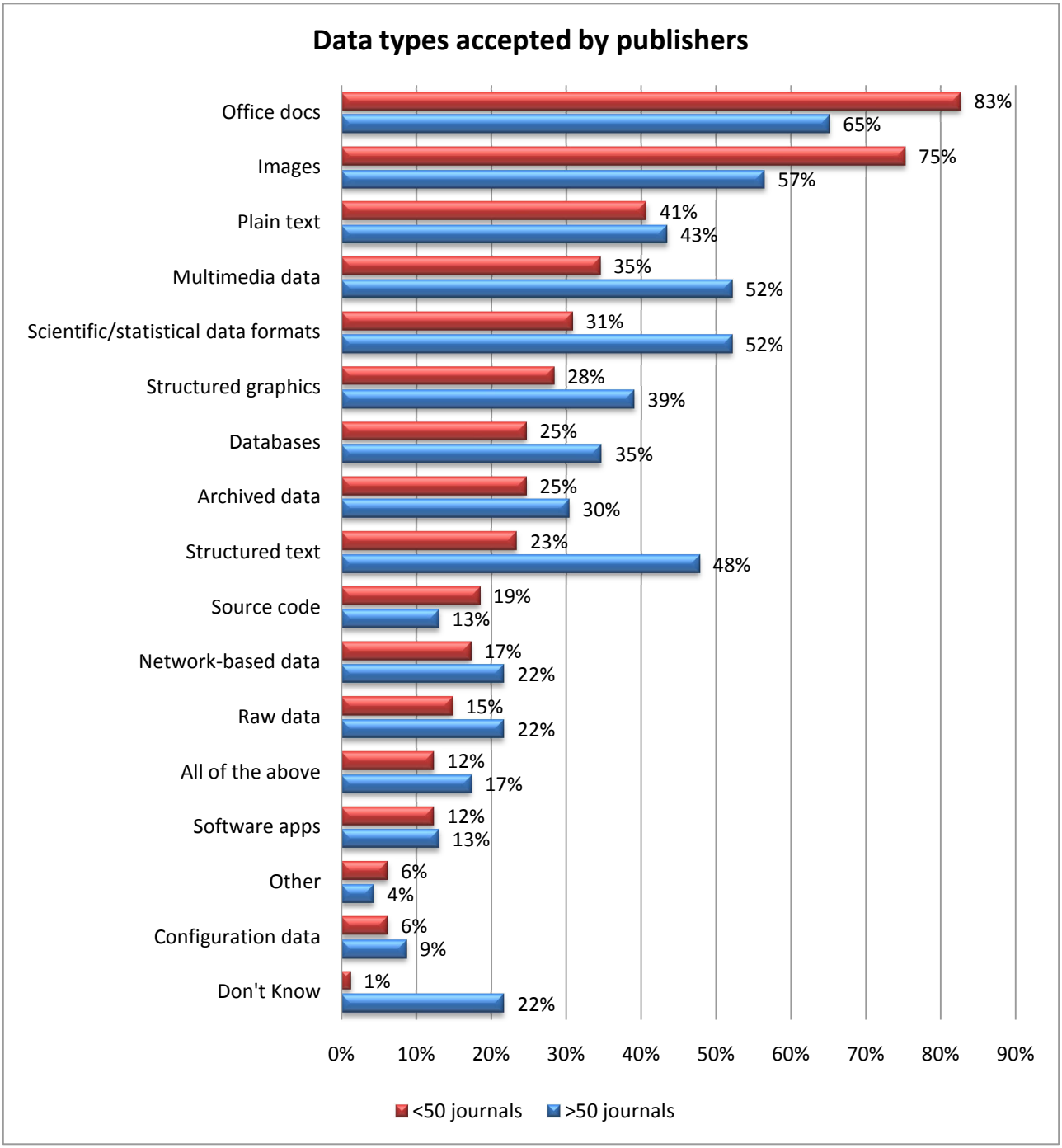


Figure 52: data types accepted by publishers, n1 = 81, n2 = 23

Office documents are accepted by a majority of both large and small publishers. The difference in percentages is worth noting though. While 65% of the large publishers accept office documents, no less than 83% of the small publishers do so. Similar differences in percentages can also be found for images. 75% of the small publishers accept images, while 57% of the large publishers do so.

Images, plain text and office documents are, one could say, the traditional data types associated with publishing. Yet nowadays publications are more than a set of made up pages. Digital publications are enhanced with video, audio, databases, etc. If we look at these other types of data it becomes clear that large publishers tend to accept these (more complicated) types of data more often than small publishers. For instance, 52% of the large publishers accept scientific/ statistical data formats and multimedia data. For small publishers these percentages are 31% and 35% respectively. Again, the difference may very well be a matter of service levels offered to authors.

8.3.3 Preservation Policies

As we have seen, quite a large number of publishers allow authors to submit underlying research data, and the data they accept is very diverse. But what happens with the data after it has been submitted and accepted? For instance, do publishers have specific preservation policies and strategies for these data? If yes, does this differ from policies with regard to journals?

Before turning to the underlying research data, publishers were asked whether they have preservation policies for their journals. 84% of the large publishers assert they have some kind of preservation policy in place. Of the small publishers only 55% claim they do (see figure 53). When expressing these percentages in number of journals, it follows that 7,698 of the journals published by large publishers are covered by a preservation policy, while 746 journals of the small publishers are covered. In total, then, 8,444 journals are covered by a preservation policy. This represents roughly 93% of the journals covered in the publishing survey.

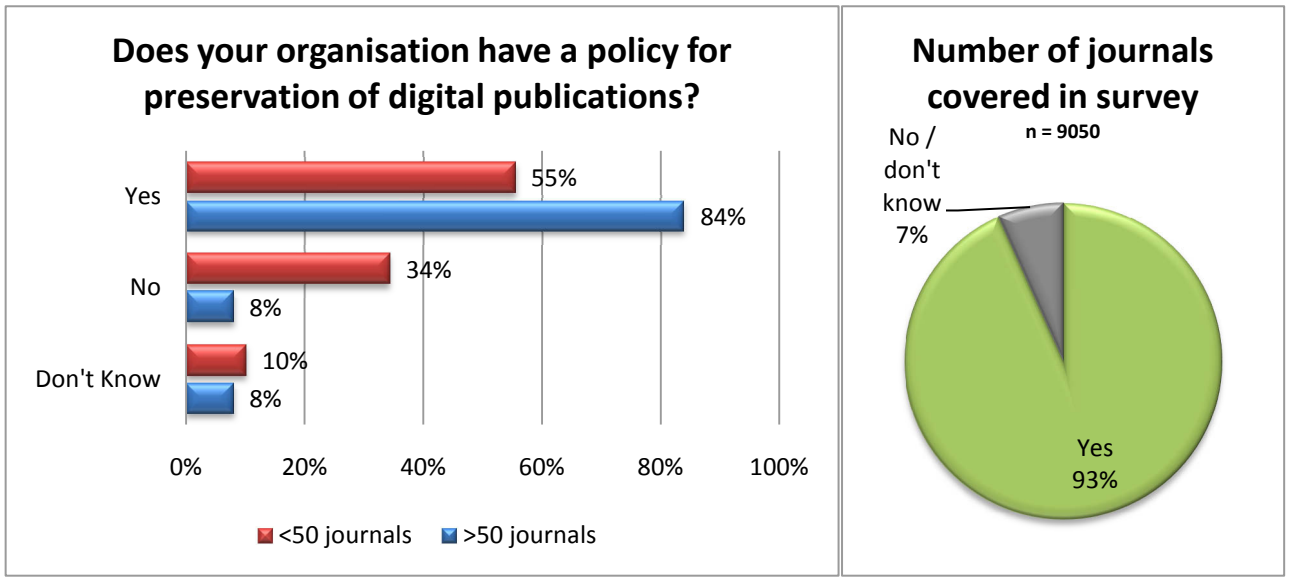


Figure 53: does your organisation have a policy for preservation of digital publications? n1 = 128, n2 = 25

What about preservation policies for the digital research data many publishers accept? When asked, it turns out that 69% of both large and small publishers do not have those arrangements (see Figure 54). Those who do have preservation arrangements in place for digital research data, mostly outsource it. It is common for publishers, as we shall see, to outsource the preservation of journal articles to other (specialised) organisations; a number of publishers have similar arrangements for digital research data. A minority of 10% of the small publishers and 3 % of the large publishers have preservation arrangements through a data archive other than for the journal articles.

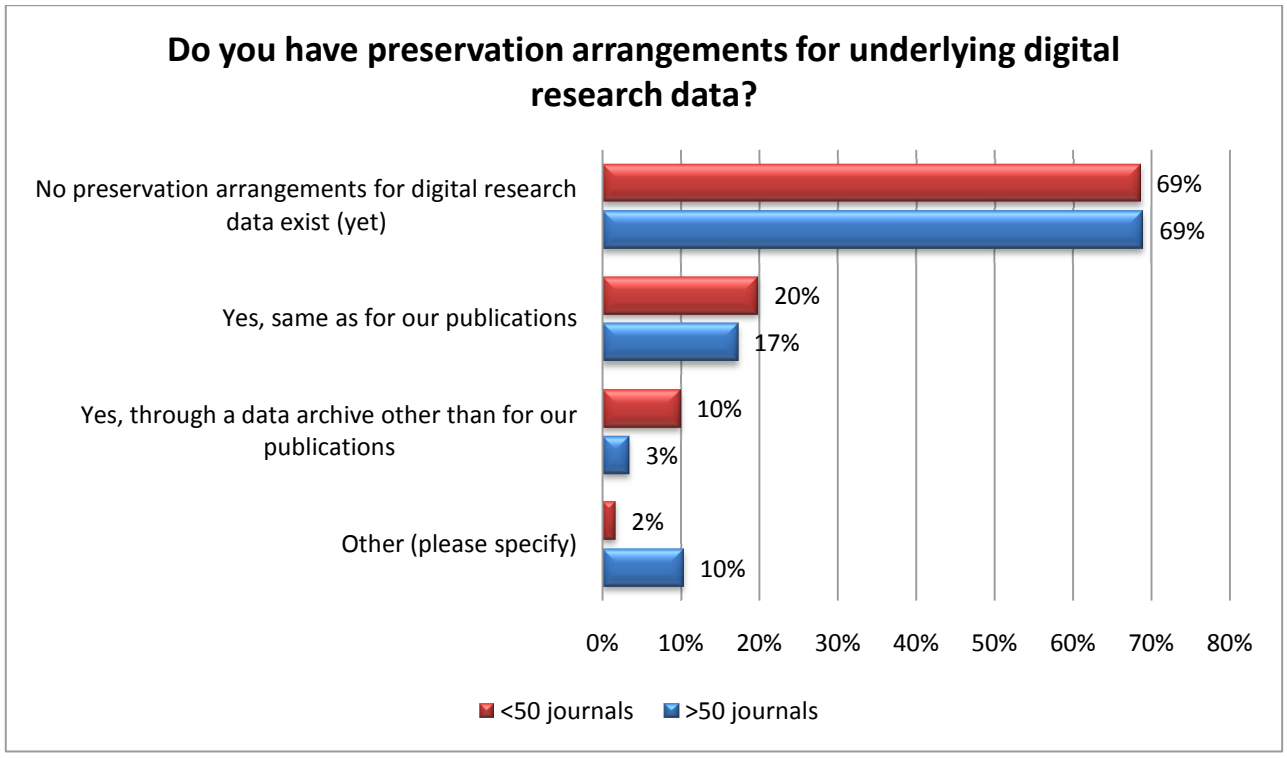


Figure 54: do you have preservation arrangements for underlying digital research data? n1 = 121, n2 = 29

When expressing these percentages in number of journals, we notice that 7,451 journals of the journals published by large publishers do not have preservation arrangements for underlying research data, while 533 of the journals published by small publishers do not have such arrangements. So, in total 7,984 journals are not covered by preservation arrangements for underlying research data. This represents roughly 88% of the journals covered in the publishing survey[31].

This, of course, tells us little about what these preservation policies exactly entail. Any good preservation strategy should at least enable an organisation to recover the data when disaster strikes. Here the percentages are better than for the preservation policies. A vast majority of the large publishers (86%) does have a disaster recovery plan and only 3% claims not have one (see Figure 55). For the smaller publishers the picture is less bright. 42% of the small publishers claim to have a disaster recovery plan, and 40% states they do not to have one.

[31] This only takes the no option into account. It is clear that underlying research data has preservation needs which differ from the needs for publications. So, we could argue that when people enter that journals preserve research data in similar fashion as they do publication, this does not really constitute a specific preservation policy for research data. If we add this option to the no option, this would mean that almost 99% of the journals covered in this survey do not have specific preservation arrangements for research data.

Expressed in numbers this means that 7,536 journals of the large publishers are covered (83% of the journals represented in the survey) by a disaster recovery plan and 759 journals (8% of the journals covered in this survey) of the small publisher are covered. For the rest of the 755 journals, published by large and small publishers, either don not have a disaster recovery plan or it is unknown whether they do.

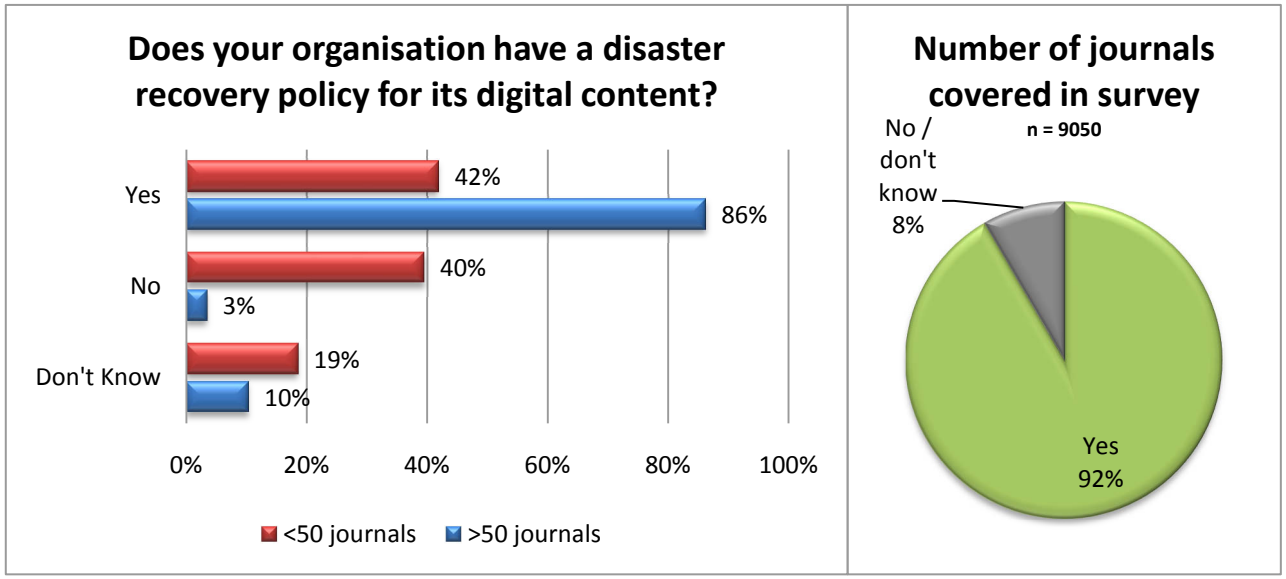


Figure 55: does your organisation have a disaster recovery policy for its digital content? n1 = 124, n2 = 29

Of course a disaster recovery plan is not an all-encompassing preservation strategy, but part of a larger set of solutions/strategies. So we also asked publishers about specific preservation strategies. The strategies we formulated and which publishers could choose from were:

- Migration
- Normalisation
- Emulation
- Outsourced to third-party service
- No preservation strategy
- Don't know
- Other

While we did not specifically state it in our survey, we can assume that the majority of respondents had journals in mind when answering this question, because, when asked, a vast majority claimed not to have preservation strategies for research data (see figure 56).

A majority of the large publishers outsource preservation to third parties (52%). Of the small publishers only 23% claim to do so.[32] In addition the top three for large publishers is completed by

[32] This has an awareness dimension though. Most small publishers were connected through the DOAJ mailing list. These small publishers are members of DOAJ and while they may not know it, at present DOAJ and e-Depot carry out a pilot project aimed at setting up a workflow for processing open access journals listed with DOAJ. In the pilot a lim-

normalisation (44%) and migration (28%). Next to 'no preservation strategy', the small publishers chose normalisation (25%) and outsourcing (23%) as the most important preservation strategies they apply. Of all options, emulation is the least chosen preservation strategies. This goes for both large publishers (8%) and small publishers (9%).[33]

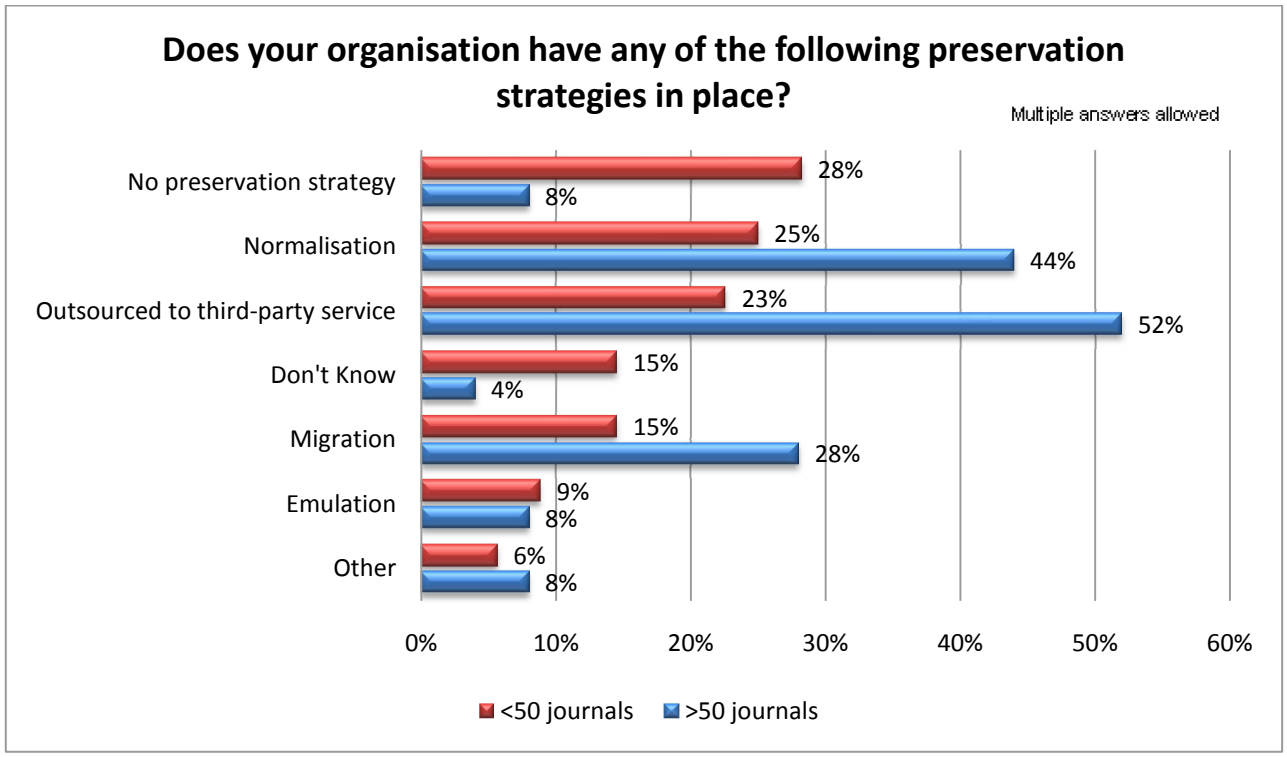


Figure 56: does your organisation have any of the following preservation strategies in place? n1 = 124, n2 =25

Multiple answers were possible, so it is possible for publishers to both outsource preservation and use in-house preservation strategies. Yet 17 out of 44 STM publishing organisations (39%) in our survey only outsource preservation.

The DOAJ publishers who stated not to have a preservation policy in place may not realize that in fact their aggregator DOAJ does have a preservation agreement for the journals of their directory with the e-depot of the national library of the Netherlands. In other words these journals are covered. It is, however, also worth noting that several small (open access) publishers also use in-house strategies for preservation.

Most large publishers who outsource the preservation of their e-journals make use of Portico (30%). CLOCKSS/LOCKSS (13%) and the national library of the Netherlands' e-depot (7%) are the two other external parties. For the small publishers these percentages are 5%, 11%, and 0%.[34]

---

ited number of open access journals will be subject to long term preservation. These activities will be scaled up shortly and long term archiving of the journals listed in the DOAJ at KB's e-Depot will become an integral part of the service provided by the DOAJ. In other words preservation arrangements are underway.

[33] In addition, one publishers mentioned website, an on-demand archiving system for web references aimed at ensuring that cited web material will remain available to readers in the future, as their arrangement.

[34] On July the 1st 2009 the DOAJ has signed a contract with the Koninklijke Bibliotheek, the national library of the Netherlands to archive all open access titles from its members in the Dutch e-Depot. However, as the publishing survey was conducted before that time, this is not taken into account.

8.4 Roles and Responsibilities

8.4.1 Funding

Digital preservation involves costs. Just like the researchers and data managers, publishers were asked to provide their views on who should pay for the preservation of research data and publications. The differences between large and small publishers are not large. A majority of the respondents believe that preservation should be paid for with public money. 63% of the large publishers and 56% of the small publishers opted for the national government (see figure 57). 41% of the small publishers and 46% of the large publishers believed that the national library should carry the financial burden of digital preservation. Yet it is also apparent that publishers think they also carry a financial responsibility and should chip in. 15% of the small publishers and 21% of the large publishers believe so.

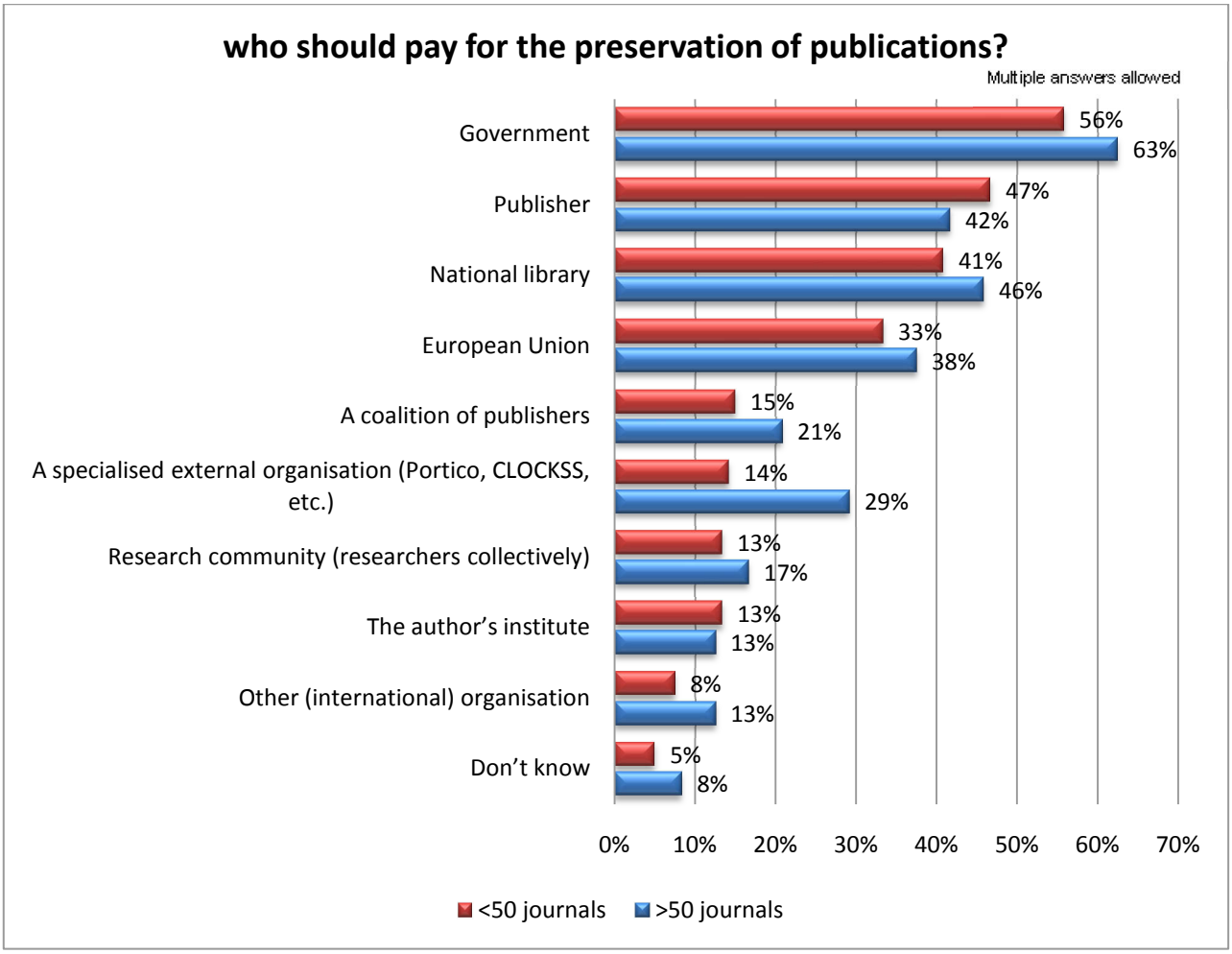


Figure 57: who should pay for the preservation of publications? n1 = 120, n2 = 24

But what about the underlying research data? While government is still the option with the highest percentage of responses – 46% for the large publishers and 55% for the small publishers – it is obvious that according to publishers the brunt of these costs also should be carried by other major stakeholders, such as researchers or their institutes (see Figure 58).

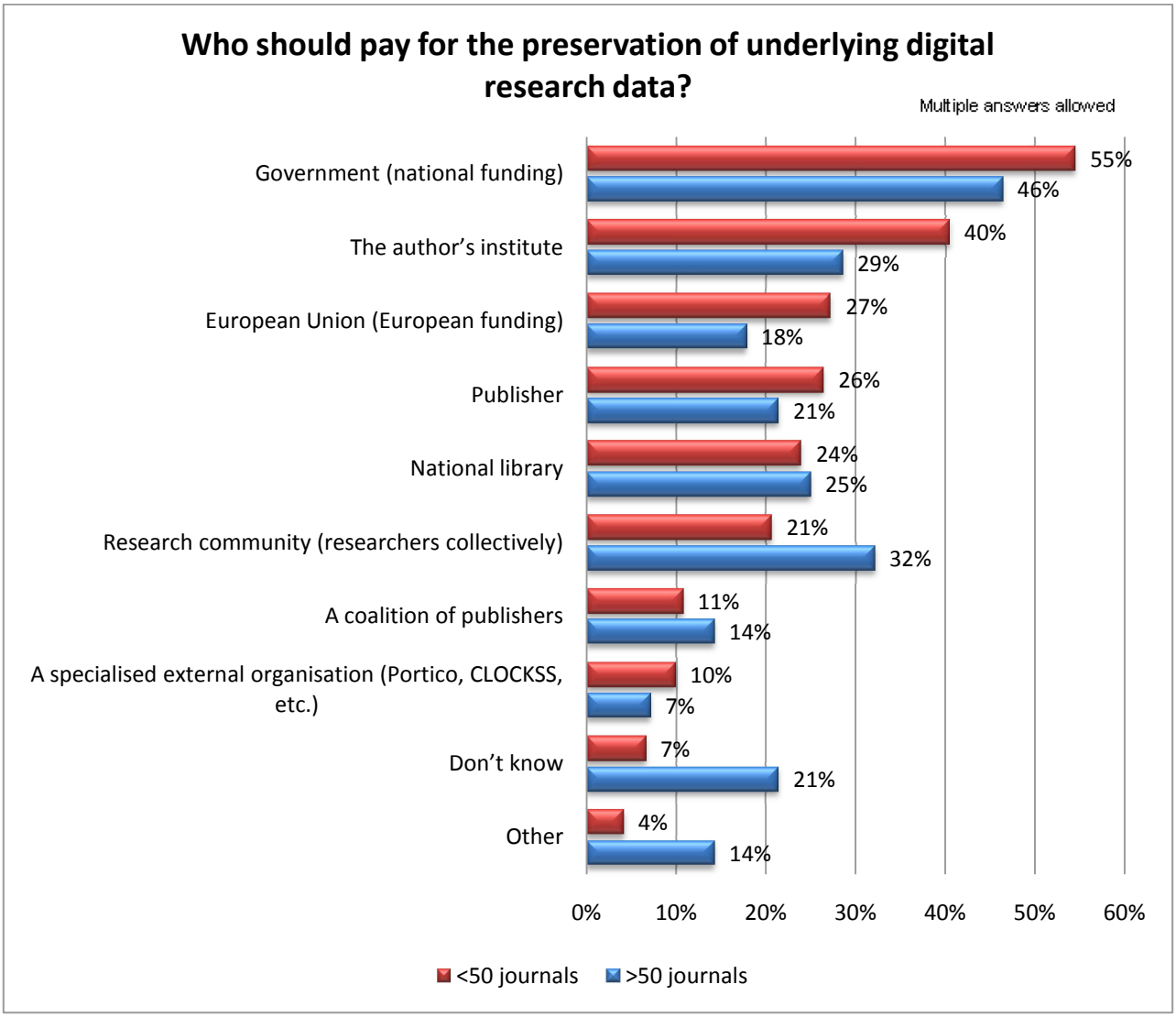


Figure 58: who should pay for the preservation of underlying digital research data? n1 = 121, n2 = 28

8.4.2 Responsibility for Preservation of Journals and Data

Preservation involves more than costs. Apart from a responsibility for financing preservation, we asked publishers – more broadly – whether they think they are (also) responsible for the proper preservation of journals and data in one form or another. Or, do publishers believe that others should carry that responsibility?

A clear majority of publishers (large and small) see themselves as the first group to carry the responsibility for the preservation of journals (69 % and 73 %). On the second place, they name the National Library (59 and 66 %). The large publishers (52 %) also rely on third party organisations like Portico and CLOCKSS (52 %), while these apparently play less a role for smaller publishers (31 %) As a second option both small publishers (66%) and large publishers (59%) agreed that the national library carries responsibility.

Large and small publishers differ in their opinion about the responsibility of public bodies for the preservation of journals. 29% of the small publishers think national governments should (also) be responsible, while only 14% of the small publishers think so. In addition, 22% of the small publish-

ers believe the EU should carry responsibility for this issue. Only 10% of the large publishers agree with this.

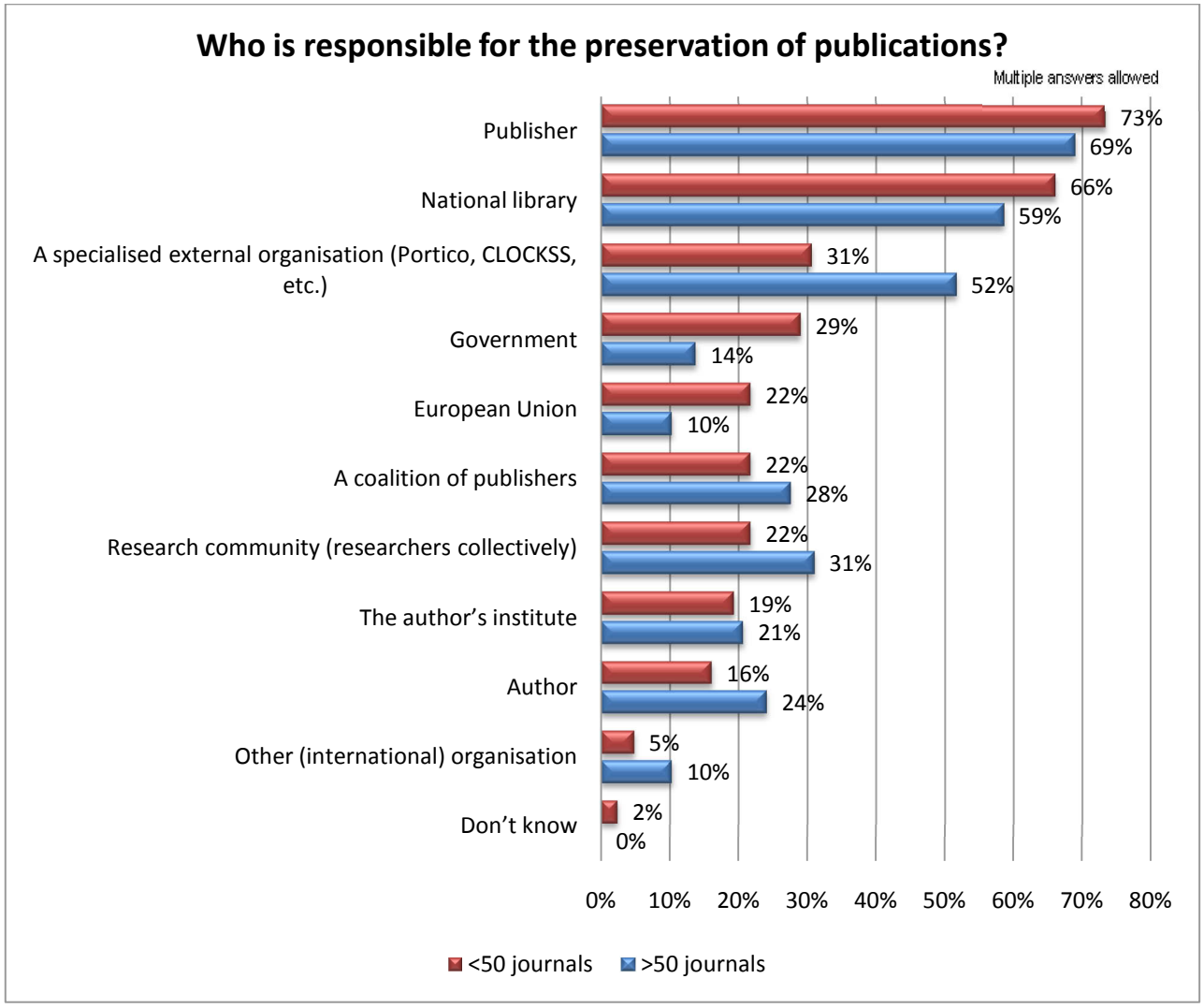


Figure 59: who is responsible for the preservation of publications? n1 = 124, n2 = 29

For the preservation of research data the figures are different from those for journals. While their responsibility for journals is undisputed, publishers see less of a role for themselves for the preservation of research data than they do for journals. 40% of the small publishers and 35% of the large publishers believe they carry responsibility for the preservation of research data (see Figure 60).

Similar to the question on who should pay for the preservation of research data, publishers think the responsibility for preservation of research data lies generally with the researchers or their institutes. 52% of the small publishers believe the author should (also) carry responsibility for preservation of research data, 43% believe the author's institute should, and 35% of the small publishers think the research community in general should. For the large publishers the figures are 48%, 43%, and 48% respectively.

It is important to note here is that the percentages are for research data are far less concrete. There is for instance only one percentage higher than 50%. It seems then save to conclude that, compared to

publications, publishers are less certain about who should carry the responsibility for the preservation of digital research data it.

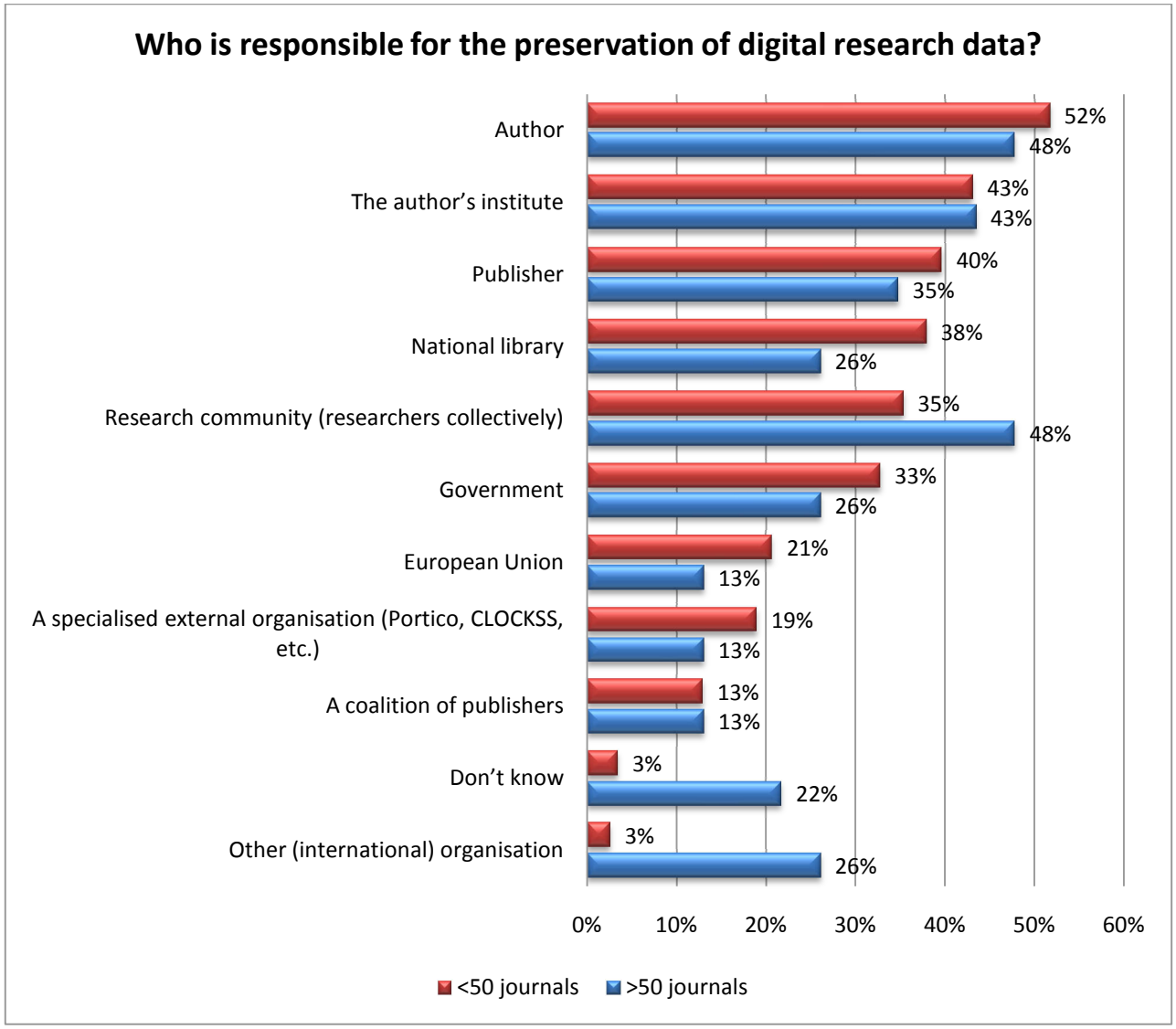


Figure 60: who is responsible for the preservation of digital research data? n1 = 116, n2 = 23

The small publishers who believe that it is (partly) the responsibility of publishers represent 490 journals (5% of the total number of journals represented in the survey). The large publishers who believe preservation of digital research data also to be the responsibility of publishers represent 271 journals (3% of the total number of journals represented in the survey).

8.4.3 Future Business models

The roles publishers will fulfil are very much tied to the way they believe the market of scholarly communication will develop. So besides asking them about their roles, we wanted to know whether publishers believe if and how under the influence of for instance digitization and open access the publishing market will change.

For this question we composed six different scenarios and asked respondents which scenario they thought were likely to happen. They could check more than one answer (see Figure 61). 61 % of the publishers believe the future will be dominated by a hybrid model, in which subscription-based and open access journals will both exist. This covers 8,150 journals. 32% of the publishers (696 journals) believe that most research results will be open access and available for free and that publishers and journals will be under strain and face difficult times. On the other hand, 35% of the publishers (7,628 journals) believe that the publishing process as such, based on journals, peer-review etc, will not change much. Because one scenario does not automatically exclude another scenario, multiple answers were possible for this question.

There are many reasons why business models may change: one of them being a change in the nature of products. What about publications? We asked what publishers think will happen by providing them four options:

- Publications will essentially not change in their function of establishing the authenticity and origin of a research result at a point in time.
- Publications will become interactive and multimedia (e.g. adding animations, sound, related web content, research data, and discussion forum).
- Publications will become living documents that are constantly updated by the research community in a wiki-like manner.
- Other

Multiple answers were possible, but what strikes is that a majority of publishers (53%) thinks that publications will essentially not change in their function. So in spite of the changes publishers believe will happen, these changes do not in the opinion of most significantly alter the function of publication.

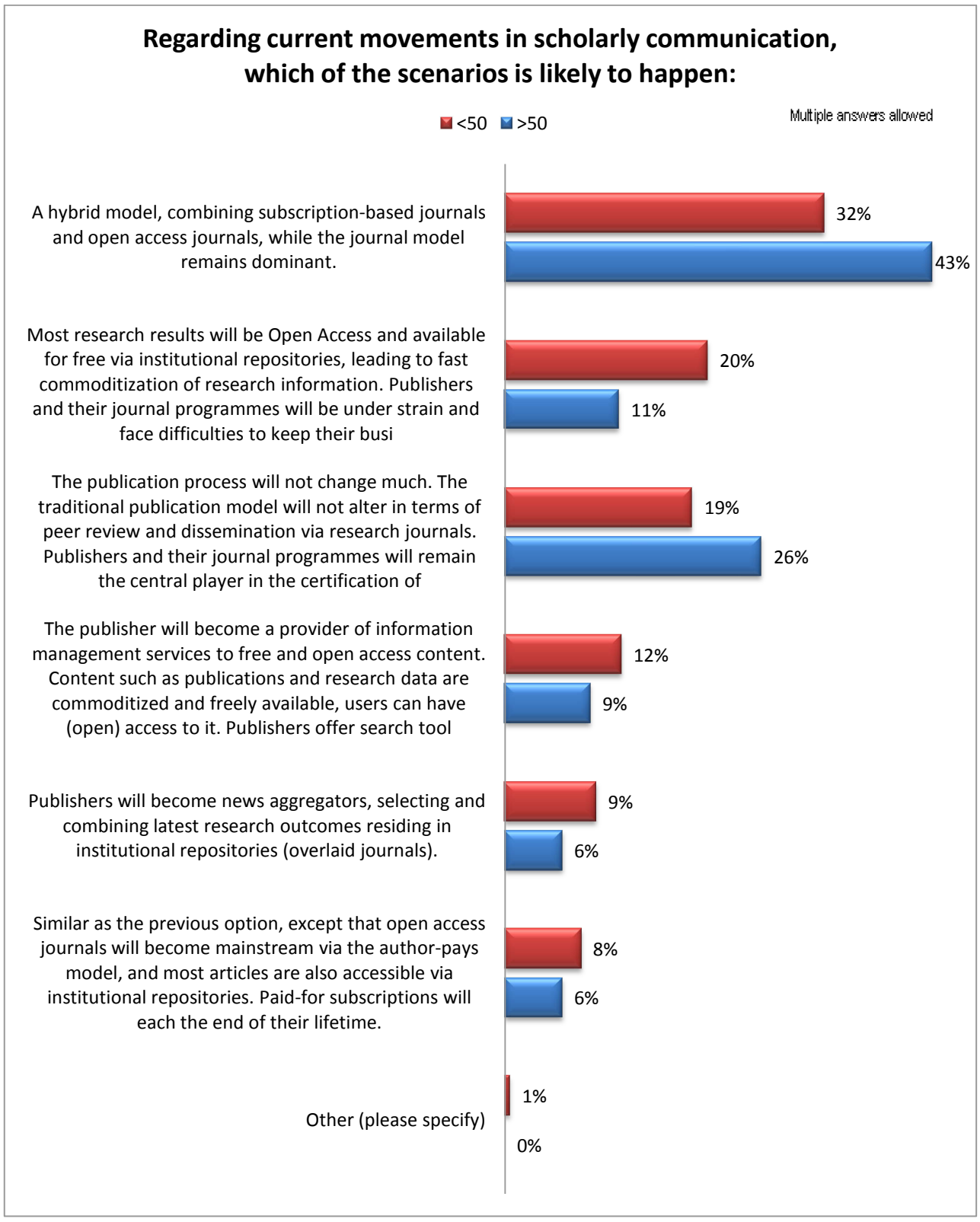


Figure 61: movements in scholarly communication, n1 = 197, n2 = 35

9 Conclusions (implications for the roadmap)

In previous sections, we presented the results of our surveys for the three stakeholders separately: researchers, publishers and data management. Although these figures are interesting already, it is even more interesting to compare them to each other. This analysis is possible because we raised several similar questions to all stakeholders. In this section we outline a cross analysis done between those stakeholders.

Note that the cross analysis is as good as the results on the individual surveys are. As the survey on data management relatively gained fewer responses than on research and publishing, conclusions in this area contain a higher uncertainty and are often indicative.

9.1 Perceptions of preservation

Each stakeholder was asked to give their opinion about why we should preserve research data. Therefore, we defined seven reasons for preservation (see previous sections for all reasons). Table 9 shows the top 3 of reasons for each stakeholder. There are no differences in top priority incentives between researchers and publishers however data managers think otherwise. While researchers and publishers clearly see a big stimulus for advancing research, data managers regard this less important. They declare uniqueness of the data as the most important incentive to keep data for the long term. A main driver for all stakeholders is that if research is publicly funded the data should be preserved as it belongs to the public as well.

Table 9: Cross analysis of top 3 reasons for preservation

TOP 3 Reasons for preservation	
Research	
1	It will stimulate the advancement of science.
2	If research is publicly funded, the results should become public property and therefore properly preserved.
3	It allows for re-analysis of existing data.
Data management	
1	It is unique.
2	It potentially has economic value.
3	If research is publicly funded, the results should become public property and therefore properly preserved.
Publishing	
1	It will stimulate the advancement of science.
2	If research is publicly funded, the results should become public property and therefore properly preserved.
3	It allows for re-analysis of existing data.

A similar analysis has been done for threats to preservation. We outlined seven threats (see previous sections for an overview of all threats) and compared the top 3 of those threats based on what each stakeholder defined as most important (see Table 10). This time, more agreement amongst the three stakeholders is found. Each of them defined technical failure and inability to understand the meaning of the data as very important. The only difference is found in whether there will be someone to look after the data in the future (research and data management) compared to publishing that stated evidential value of the data to prove results derived on that may be lost.

Table 10: Cross analysis of top 3 threats to preservation

TOP 3 Threats to preservation	
Research	
1	Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
2	The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
3	Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
Data management	
1	Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
2	Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
3	The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
Publishing	
1	The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
2	Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
3	Evidence may be lost because the origin and authenticity of the data may be uncertain.

Regarding the need for an infrastructure to counter these threats the majority of the respondents to all stakeholders agreed on the necessity of having such an infrastructure in place. Figure 62 shows this outcome for each stakeholder. Especially respondents of publishing support this kind of solution (74%). One clarification might be that publishers believe a more structured and scalable approach is needed to ensure all data (including publications and other research output) is kept safe. Current local or discipline-specific solutions that exist might be sufficient for researchers in a particular field, but do not scale enough to be used across disciplines. Sharing data and defining cross references to publications is therefore difficult to achieve right now.

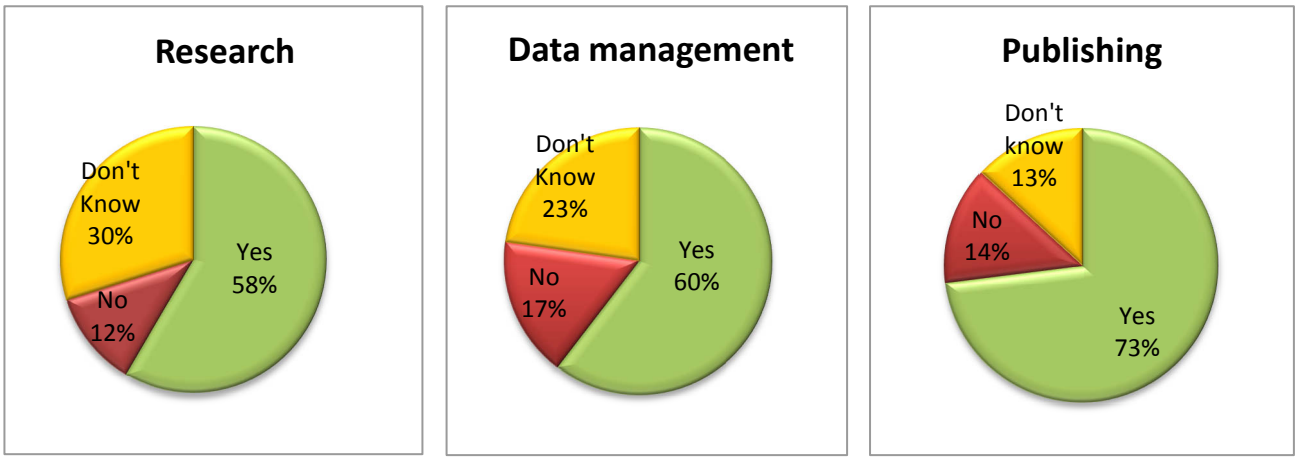


Figure 62: difference of used and preserved data types

9.2 Preservation - State of affairs

This report is about preservation of data in any kind. It is interesting to see if there is a match between what researchers use and what data managers actually preserve for the long term. Normalising the responses of this question for both research and data management shows some interesting differences (see Figure 63). Looking at what is used by researchers but not adequately preserved by data managers (see blue peaks of second graph) indicates that: network-based data (such as websites), source code, computer applications and raw data are often not preserved well.

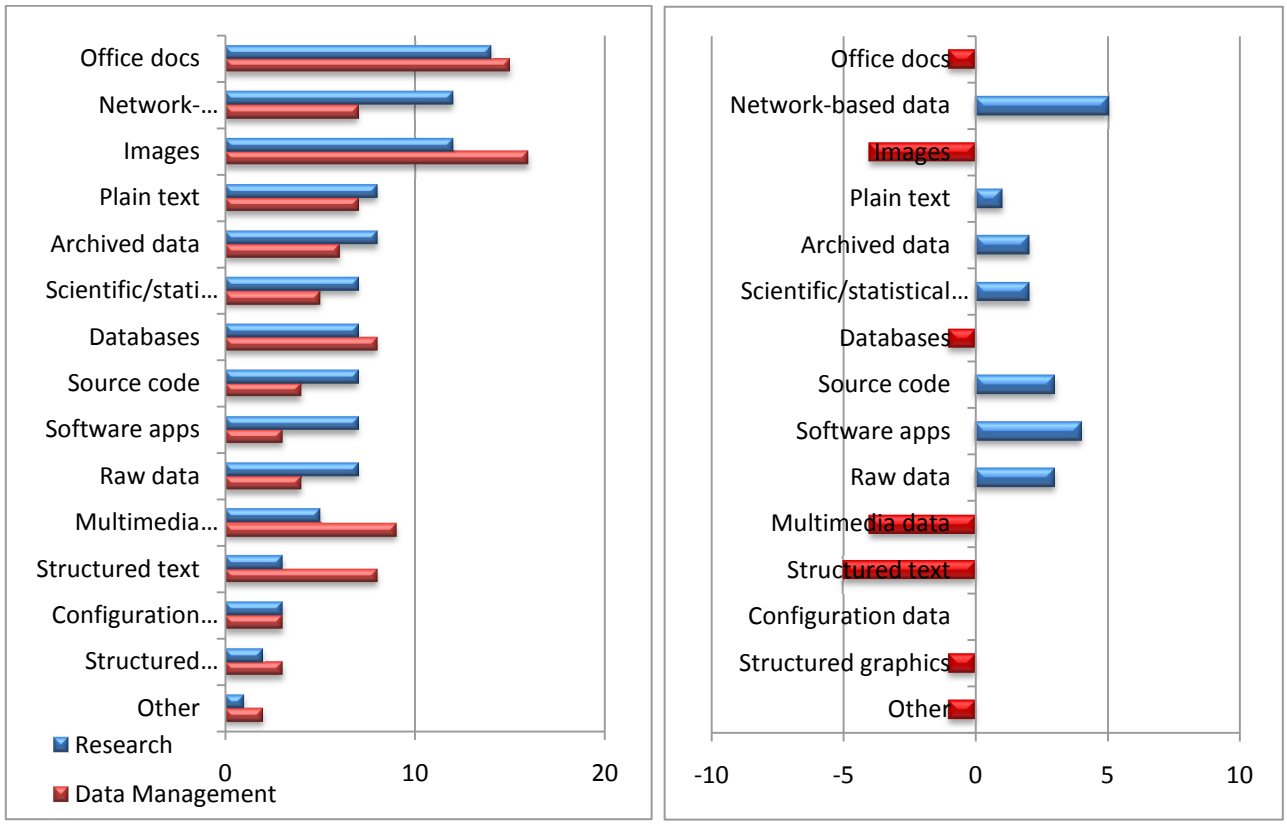


Figure 63: comparison of used and preserved data types, n1 = 1366, n2 = 206

While most publishers accept the most important data formats researchers use and store, they do not feel they are the ones who should preserve underlying data sets researchers submit with their manuscripts.

What about the amount of stored data? From the Researchers survey it is clear that researchers expect a significant growth of stored data. Likewise Data Managers too expect the amount of data they store at their facilities will grow significantly in the next five years. At the same time, however, the percentages of data managers who don't know what to expect in five years time grows to 28%. Perhaps this uncertainty can partly be explained by the current storage activities of researchers.

As became clear from the researchers survey, researchers tend not to store their digital data at external facilities. The most important storage locations for researchers were personal computer at work, portable storage carrier, organisational server, and computer at home. At the same, as we have seen, researchers are concerned about legal issues and misuse of their data when it is stored elsewhere. One conclusion we can draw from this is that there is a human, psychological dimension to the issue of preservation which has to do with trust. This cannot merely be tackled by technical solutions. It is another confirmation that a roadmap should include more than these technical solutions.

Whether the researcher's fear for misuse and legal issues is justified is open for debate. The majority of the data managers, who responded to our survey, claim to have proper data management procedures in place. There is one exception: policies which deal with liability when data is lost or affected. While we cannot claim that the researchers' fear is based on a lack of these kinds of procedures, it is obvious that lack of liability arrangements is not helpful in this respect.

9.3 Policy

As said, most data management organisations represented in our survey have preservation policies for research data in place. The same goes for publishers, be it that they tend to outsource preservation to some of the data management organisations present in our survey. For 8,444 journals, or 93% of the journals covered in the publishers survey, their publishers indicate that they have a preservation policy implemented.

Large publishers with more than 50 journals in their catalogue are better represented (84%) than smaller publishers with less than 50 journals (55%) when it concerns digital preservation. Amongst smaller publishers, a significant percentage of 34% indicates **not** to have such a policy, against only 8% among the larger publishers.[35] This means that in terms of awareness, policy implementation

[35] It should be noted here though that many of the publishers who received our survey through the DOAJ mailing list are perhaps not aware that their journals are preserved through an arrangement DOAJ made with the Dutch e-Depot at the Koninklijke Bibliotheek since the 1st of July 2009.

and e-infrastructure there may be a need for better education of the smaller publishers on this aspect.

Many publishers outsource the digital preservation of their journals. Again, this is overall better arranged for larger publishers than for smaller publishers. A few of the organisations publisher outsource to are: Portico (30%), CLOCKSS (13%) and KB (7%). An implication for the roadmap here is that in any future e-Infrastructure, these outsourcing parties should be included as important players in the integrated chain of digital preservation.

Against this relatively clear situation concerning official research publications in journals, there is a far more diffuse picture for the digital preservation of the underlying research data. Both small and large publishers accept datasets with submitted manuscripts but a significant percentage of 69% of both small and large publishers does **not** have a preservation policy in place for datasets. If they do, most of them outsource this as well.

9.4 The outlook

A majority of data managers who responded to our survey do not think that the tools and infrastructure currently available to them suffice for the digital preservation objectives they have to achieve. This is a clear justification for the infrastructure roadmap that is developed in the PARSE.Insight project. But it does not necessary follow that more archives have to be built. In fact, from our surveys there is no clear indication that many new archives are being built. If there are, at least the researchers, whom we asked, are unaware.

One of the changes data management organisations will have to cope with is the changing nature of publications. Until quite recently, online publications were little more than digital versions of paper publications. Except for a possibility of hyperlinking, they added little extra functionality. This is changing. Yet publishers believe it to be more a change in form than in function. As we have seen a majority of publishers (53%) thinks that publications will essentially not change in their function.

If this will be the case, then the implication for the infrastructure (and the roadmap) is mostly technical in nature. Preservation facilities will have to be able to handle the changing forms of publications.

9.5 Roles & Responsibilities

When publishers were asked who should be the main responsible party for the preservation of datasets, they mentioned the author, the research institute, and the research community as the important responsible parties. While publishers see themselves clearly responsible for the preservation of the publications, they think of themselves as less responsible for the preservation of datasets. About 40% of the small publishers think publishers are responsible and about 35% o the large publishers believe so.

In this context it is interesting to see that in the researchers' survey 15% of respondents submit their datasets with their manuscript to a journal and its publisher as compared to 14% who submit data to a digital archive at their organization and only 6% who store data at a digital archive of their discipline. A reason for the relatively high number of researchers who submit datasets to the publisher could well lie in the fact that a research article elaborately describes the origin of the data, the methods used, its meaning and its shortcomings. Many researchers fear that their data might be re-used out of context—accessibility of the data via their publication could help avoid that.

An important implication for the roadmap and the envisioned e-infrastructure can be that any infrastructure should ensure good linking and connectivity between research publications and the underlying data, for example via systematic depositing of such datasets in central repositories and persistent identifiers to link to and from related publications.

About the role in general for journals in the future, a majority of publishers believe that the journal will transform into a hybrid model of Open Access and subscription-based. Only 32% believe that Open Access publishing will become the mainstream. Yet 35% believe that the traditional role of the research journal for peer reviewed articles will not change much.

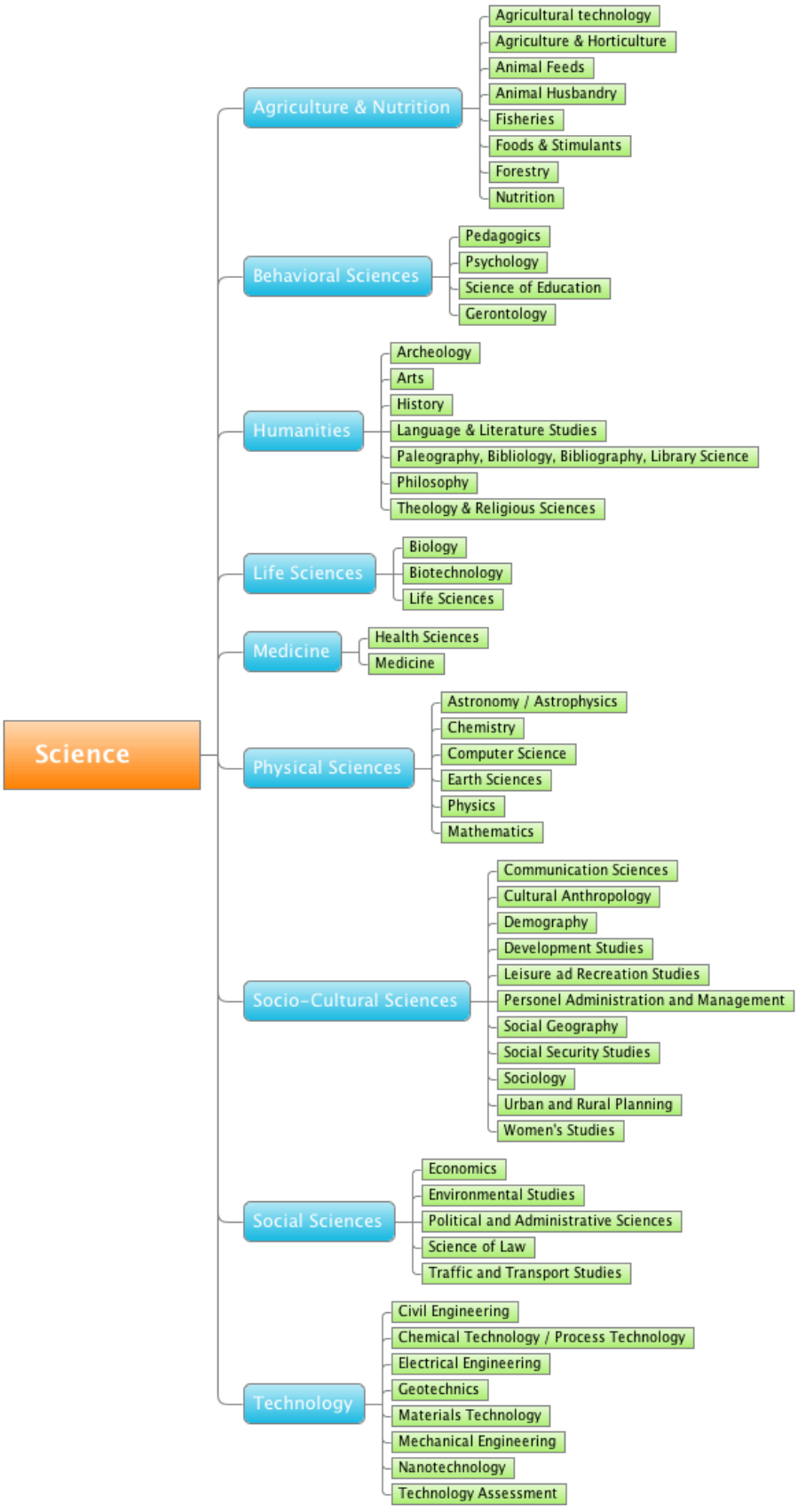
An implication for the roadmap may be that in order to safeguard proper preservation measures by the publisher stakeholders, it is important to ensure healthy and sustainable business models for journal publishing. Many smaller publishers seem to be less aware of preservation activities and probably do not yet include the costs for it in their business models.

Regarding funding, all stakeholders seem to agree that more resources are needed and that national governments and the European Union should pay the bill for preservation activities.

10 List of figures

11 List of tables

Appendix 1: Classification of disciplines



Appendix 2: EU member states (2009)

Source: EU website: http://europa.eu/abc/european_countries/index_en.htm