Insight into digital preservation of research output in Europe

PARSE
insight

Permanent Access to the Records of Science in Europe

SEVENTH FRAMEWORK PROGRAMME

Insight Report

Project Number	223758
Project Title	PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Insight report
Deliverable Number	D3.6
Contributing Work package	WP3: Community Insight
Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	30 April 2010 (M26)
Actual Delivery Date	18 June 2010 (M28)
Author(s)	Thaesis (external company), Jeffrey van der Hoeven

Abstract

This report (deliverable 3.6 of PARSE.Insight) describes the overall results of the surveys and interviews conducted by PARSE.Insight to gain insight into digital preservation of research output in Europe. Major surveys were held within three stakeholder domains: research, publishing and data management. In addition, several interviews have been conducted together with desk research to gain insight into research funding and policies around preservation of research output; 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, interview, results, research, publishing, data management, funding, discipline, data, sharing, publication

Contributors

Person	Role	Partner	Contribution
Thaesis (external company)	Author	KB	Main author of document
Jeffrey van der Hoeven	Co-author	KB	Contributing author
Moritz Gomm	Reviewer	FUH	Review of document
Peter Igo-Kemenes	Reviewer	CERN	Review of document
Ingrid Dillo	Reviewer	KB	Review of document
Eefke Smit	Reviewer	STM	Review of document

Document Approval

Person	Role	Partner
Peter Igo-Kemenes	Reviewer	CERN
Moritz Gomm	Reviewer	FUH

Revision History

Issue	Author	Date	Description
0.1	Thaesis	09-04-2010	Initial version
0.2	Thaesis	29-04-2010	Incorporated comments from KB
0.3	Jeffrey van der Hoeven	07-05-2010	Updated sections based on extra KB input
0.4	Jeffrey van der Hoeven	01-06-2010	Incorporated comments CERN, FUH and most of KB
0.5	Jeffrey van der Hoeven	17-06-2010	Incorporated comments KB and STM
1.0	Jeffrey van der Hoeven	18-06-2010	Finalised document, including comments DNB, wrote executive summary, updated overview of insights

I. Executive Summary

This report is the final result on the PARSE.Insight study into the current state of affairs and needs regarding digital preservation of research output in Europe. The study has been conducted from March 2008 until June 2010. Within this study we subdivided the Europe's research landscape in four stakeholders: researchers, data managers, publishers, and funders. Several methods were deployed to gather information of these stakeholders on their practices, ideas, and needs to guarantee long-term access to research output. These methods comprise desk research, in-depth interviews, case studies into three specific research communities and large-scale surveys.

All stakeholders in this research agree that preservation of research output is important. Reasons such as *it may stimulate the advancement of science* and *it allows for re-analysis* were acknowledged. But that preservation is by no means simple, was commonly understood as well. Threats to digital preservation such as *lack of sustainable hardware, software* and *evidence may be lost because the origin and authenticity of the data may be uncertain* were acknowledged by all stakeholders although they vary sometimes. For example, in the disciplines of High Energy Physics and Earth Observations, experiments simply cannot be redone easily (if at all). For a complete list of reasons and threats, see further onwards in this report.

The current state of affairs on preservation of research output in Europe is diverse and fragmented. Preservation of publications is covered pretty well by publishers and data managers, but looking at the broader spectrum of research data (e.g. data sets, software) the outcome is less bright. Some data repositories do exist, but not every discipline is covered and the organisation differs per country and discipline. Only a few policies on preservation procedures and guidelines exist and the ones under development are done mostly in isolation while every country has to deal with this. Funders can play a significant role in this but currently are more focused on access to data on the short term than looking at data preservation for future generations. However, they do recognise that this is important.

To cope with current threats of preservation, all stakeholders agree that a science data infrastructure is required. But the roles covering who should do what in this matter is far from clear. Roles should be defined more explicitly and business models should be developed. For this, strong coordination is required. Also, awareness should be raised on what exactly digital preservation is and what should be done.

If done right, there is a huge potential for a better research environment. Researchers spoke out their desire to use each other's data. However, they are reluctant to share their own data as they see legal issues and misinterpretation of their data as important hurdles. They are unfamiliar with archiving and do not like to lose much time working on that. Clearly, challenges lie ahead.

See section 5.3 for an overview of insights, recommendations and references to the roadmap.

II. Table of Contents

1. 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 traceable, 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.

1.1.   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 being deployed in this area, such as newly founded digital archives and enabling persistent identification of scientific papers. However, a coherent vision shared by all stakeholders is missing. The aim of the PARSE.Insight project is to define a roadmap and recommendations for developing the e-infrastructure for research in order to maintain the long-term accessibility and usability of scientific digital information in Europe.

To understand the current situation in Europe, PARSE.Insight carried out a broad study. Several large-scale surveys, case studies, desk research and interviews have been carried out targeted to specific stakeholders in research. This report highlights the cross-stakeholder analysis and offers a number of conclusions. Based on that, a gap analysis can be performed to measure gaps between 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.

The PARSE.Insight project was initiated by the Alliance for Permanent Access to the Records of Science[1]. Results of the project will also be taken into account by the Alliance and its members.

1.2.   About this report

The aim of this report is to analyse and summarise all insights gained during research of PARSE.insight. Special attention is given to the role of research funders in digital preservation as this stakeholder in research was not yet covered in previous reports of the project. Other

---

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

stakeholders covered briefly are researchers, data managers and publishers. Based on the findings, conclusions are drawn and recommendations are given which can be used for refining the roadmap and paving the way for a science data infrastructure in Europe.

Parts of previous research results have been made public already and can be found on the PARSE.insight website[2].

1.3.   The structure of this report

Firstly, an outline of the four stakeholder groups that are central in this research (researchers, data managers, publishers and funders) is given in chapter 2. After that, chapter 3 provides the research results for four specific topics that were addressed:

- Perceptions of Digital Preservations (reasons and threats)
- Current practices of Digital Preservation
- The future outlook of Digital Preservation and the need for an international infrastructure
- The roles and responsibilities for organising and financing Digital Preservation

In Chapter 4, the research results of the various stakeholders are compared and analysed. In Chapter 5, conclusions are drawn and implications for the *PARSE.Insight Roadmap for an e-science data infrastructure* are discussed in terms of four focus areas. These are:

1. Awareness
2. Technological needs
3. Rules and best practices
4. Funding

At the end of the report, the methodology of this research is further explained and a list of references is given.

1.4.   Terminology

Digital research data in the context of PARSE.Insight is the term used for all output in research. In practical terms, raw data, processed data, publications and post publication material, 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).

---

[2] PARSE.Insight website. Available at: www.parse-insight.eu.

Digital preservation is the set of processes and activities that ensure continued access to information in digital form. It 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). Furthermore, it entails several activities such as taking into account environmental changes (preservation watch), preservation planning (what needs to be done when), and preservation actions (e.g. migration, emulation).

Open access is about the absence of access entitlements – it means that information (e.g. publications, raw data) is accessible to anyone without technical, legal or financial restrictions. It should be made clear that open access is not the same as digital preservation, which deals with accessibility over time, but regardless of access entitlements. Open Access can regard data, grey literature and publications. The relation between the two is that Open Access material is equally dependent on good digital preservation as is any other digital information with or without access entitlements.

2. Stakeholders in research

2.1.   Four stakeholders defined

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.

The PARSE.Insight project aims at European stakeholders within the research community[3]. This encompasses stakeholders from all member states of the European Union and all research disciplines. Four major stakeholders are recognized in research (see Figure 1):

- Researchers
- Data managers
- Publishers
- Funders
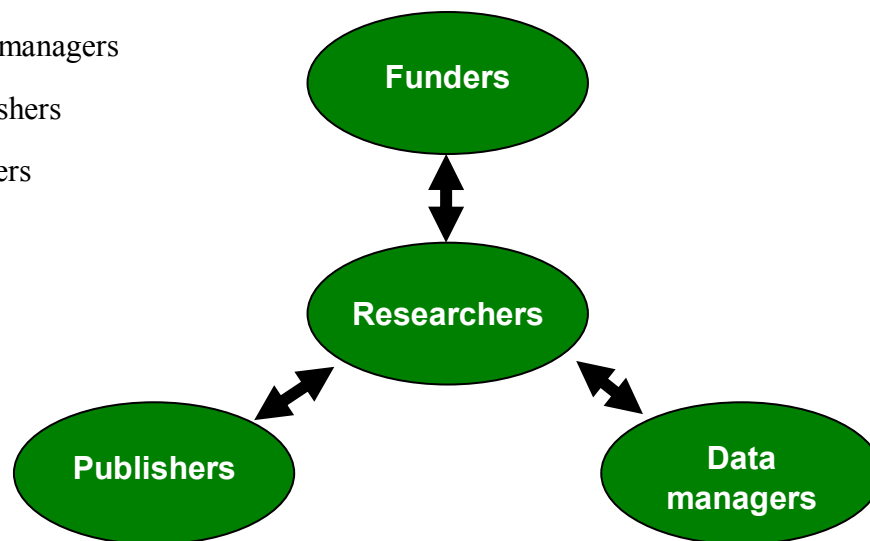
Figure 1: Generalised view on stakeholders in research

Within this group of stakeholders, the researcher is the central actor. Researchers provide publishers and data managers with the necessary research output, while being financed by funders. Meanwhile, they are consumers of disseminated research output as well.

---

[3] For a geographical view on all stakeholders within research in Europe and abroad, PARSE.Insight created an interactive map, available at: http://www.parse-insight.eu/imap_intro.php.

2.2.   Researchers

All researchers create, use and reuse research data. Research data is the key output of these activities.

As creators of data, researchers are responsible for[4]:

- managing data for the duration of the project

- making the data available in a form that can be used by others

- using standards where possible

- complying with data policies

- disseminating their research work by writing articles and other publications

As users of data, researchers are responsible for:

- adhering to any license and restrictions of use

- acknowledging data creators and curators

- proper citation of and reference to previous work and data as found in publications

- managing every derived data

- providing feedback to the research community and data archives

2.3.   Data Managers

The category *data managers* covers profit and non-profit data archives, traditional memory institutions (libraries, archives & museums) as well as research and development in preservation technology itself.

Data archives collect data, make them accessible to researchers, and preserve them for future use. Hence, data archives are centres of expertise in data acquisition, preservation, dissemination and promotion. Data archives are setup as (cross-) institutional, national or disciplinary archive and are part of a network of national and international organisations with whom they maintain contact in order to keep knowledge current regarding new data collections, information

---

[4] The description of roles and responsibilities refers to Liz Lyon, Dealing with Data: Roles, Rights, Responsibilities and Relationships, 2007.

management techniques and new computing technology. There are several national and international data archives in Europe[5].

Memory institutions follow a similar workflow as data archives but have a broader focus. They collect all kind of information of humanity's culture, make that accessible to researchers and are responsible for long-term preservation. Archives, museums and libraries are paradigmatic examples of traditional memory institutions. Most of them preserve digital collections as well[6]. These collections cover both born digital documents and digitized material.

Data archives and memory institutions have to manage data for the long term. In this context, they are responsible for:

- Identification, validation and selection

- performing preservation watch, planning and actions to ensure data remains understandable

- supporting ingest and metadata capture

- protecting the rights of data creators

- providing access

- supporting re-use via special tools

- participation in the development of standards

- training

2.4. Publishers

The category *publishers* covers publishers of academic books and journals and is regarded to be the most important stakeholder for disseminating the results of research.

With the advent of digital information systems and the internet, the scope of publishing has expanded to include digital resources, such as the digital versions of books and periodicals. Publishing is undergoing major changes, emerging from the transition of print to digital format. Currently, publishers are experimenting with interlinking publications with original research data, and integrating new kinds of digital research output into publications. Most publishers have preservation arrangements for all their publications in place, largely outsourced to third party services from National Libraries (for example KB) or specialised (non profit) organisations (like

---

[5] The PARSE.Insight interactive map in digital preservation shows some of the stakeholders within research in Europe, see http://www.parse-insight.eu/imap_intro.php

[6] PARSE.Insight survey report, section 7.3 *preservation – state of affairs*, pages. 42 - 43

Portico). In most cases, supplementary material that is submitted by the author together with the manuscript receives the same preservation treatment as the scientific article. But only 12% of all supplementary data attached to journals are officially covered by a data preservation policy[7]. Publishers do understand the importance of preservation but do not regard preservation of data as an activity in their own area of expertise.

Responsibilities of publishers in the context of preservation are to:

- organise the dissemination of research output

- manage all publications under their copyright

- ensure sustainable business model for this form of research dissemination

- acknowledge data creators and curators

- adhere to any licenses and restrictions of use

- protect the rights of data creators

- support ingest and metadata capture as well as re-use via special tools

- arrange for digital preservation of the publications

2.5.   Funders

The category *funders* comprises organisations that are involved in the process of management, funding and establishing policies in long-term preservation of digital research data. Funders provide capital (funds) for a person, project, or a private or public institution. Funds can be allocated for short-term as well as for medium and long-term purposes.

The main actors in funding of European research are the European Commission and national governments. Most national research funding agencies in Europe are associated in the European Science Foundation (ESF)[8], which promotes scientific collaboration between countries. Within the European Union, the European Research Council (ERC) is the major research funder. On national level, the government is often represented by agencies amongst which the available research funds are divided. These agencies either have a general character or they are discipline specific. In some cases though, funding is also directly assigned by national governments. The typical research funding structure, both on national as on EU level, can be visualised as in figure

---

[7] PARSE.Insight D3.4 Survey report, page 64

[8] ESF members and observers, available at: http://www.esf.org/research-areas/marine-sciences/about-us/members-and-observers.html

2^9. In addition, a small part of scientific research is funded by non-profit foundations and private institutions. In contrast to the United States, this part of research funding is negligible in Europe since it accounts for less than one percent of total research funding[10].
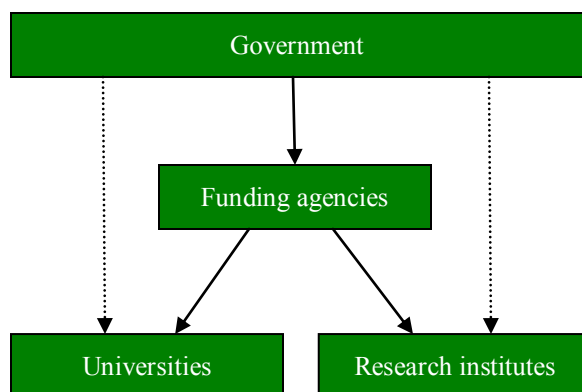

Figure 2: Typical research funding structure

Next to funding, policies can be understood as political, management, financial, and administrative mechanisms, arranged to reach explicit goals. They form deliberate plans of action to guide decisions as well as to achieve rational outcomes. Many funding bodies, data management organisations and research institutions, no matter the scientific discipline, have policies in place concerning accessibility of research data (e.g. the Safeguarding Good Scientific Practice[11] (1998) by the German Research Foundation (DFG), the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[12] (2003) signed by several international scientific organisations and the Declaration on Access to Research Data from Public Funding[13] by the Organisation for Economic Co-operation and Development (OECD)). However, policies concerning long-term digital preservation are far less often available and seem to be in an early stage of development.

In this, funding organisations are important actors in long-term preservation of digital research data. They can set the scene for new guidelines and regulations regarding digital preservation. In this field, they have the following responsibilities:

- to consider wider policy perspectives

---

[9] Overview of research funding structures in Europe, available at:
http://cordis.europa.eu/erawatch/index.cfm?fuseaction=ri.home

[10] http://cordis.europa.eu/erawatch/index.cfm?fuseaction=ri.content&topicID=4&countryCode=EU

[11] http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf.

[12] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[13] http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

- to develop policies, either in co-operation with other stakeholders or by themselves

- to monitor and enforce policies

- to act as advocate for data curation and fund expert advisory services

3. Insight into preservation

Through different parts of research PARSE.Insight gained insight into the four defined stakeholders. This chapter brings these insights together. Graphs of the statistics can be found in the S*urvey report* (D3.4) and on the PARSE.Insight website.

3.1. Researchers

3.1.1. Perceptions

While the reasons for preserving digital research data are often regarded as self-evident by the specialists, it is useful to know what the various stakeholders think about these reasons. In the large scale survey, the stakeholders 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.

Researchers consider the possibility of *re-analysis of existing data* as the most important driver for preservation of research data (91%), 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 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 research data* as either an important or a very important reason to preserve research data.

Within the research communities, the importance that is attributed to the various reasons for long-term preservation differs to some extent. This has also been put forward in the *PARSE.Insight Case Studies report*[14]. Within the disciplines of High Energy Physics (HEP) and Earth Observations (EO), the uniqueness of measured data is regarded more important than in other disciplines. For HEP assumes that those measurements cannot be repeated at will due to enormous financial, technological and human efforts invested in each experimental project. The discipline Earth Observations is even more critical as their observations cannot be repeated under the same conditions because things on Earth are constantly changing. Within the discipline of Book Studies on the other hand, preservation was regarded to be vital for stimulating the advancement of science. Without preservation, new research cannot build on existing knowledge.

Besides the various reasons stakeholders may see as important for digital preservation, they also identify different threats. It is important to address these threats, so that they can be taken into account when future policies are formulated.

Seven potential threats were identified:

- 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.

There seems to be a high degree of awareness on the major threats to long-term preservation of digital research data. Between 56% and 80% of the responses among researchers 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.

---

[14] This report can be found on the PARSE.Insight website: www.parse-insight.eu.

Zooming in on the research 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. Researchers in the humanities and physics however seem mostly concerned with the threat that *future users may be unable to understand the data*. In the *Case Studies report* this is noted by HEP-researchers as *"Adequate documentation capturing the experimental conditions and the 'insider knowledge' of researchers is perceived as a key issue for a flawless re-analysis of the preserved data."*[15]. Researchers from the agriculture & nutrition 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 researchers. Many - especially socio-cultural and social sciences researchers - consider the possibility that *organisations or projects may cease to exist* as a major threat to the preservation of digital research data.

3.1.2.   The state of affairs

To be able to determine what is needed for the preservation of research data, it is important to learn more about the day-to-day work practice of researchers. What data do they preserve? Where do they preserve the data and are they aware of the volumes of data they are preserving? Do they share data?

Not surprisingly, office documents are the kind of data most researchers make use of (94%). 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). For both data types 79% of the respondents claimed to use them. More interestingly, almost half of respondents use source code, software applications, raw data and databases. It is likely that these forms of digital files offer significant challenges in terms of usability and keeping them understandable for future users.

It proved to be a difficult question to answer for researchers how much data they actually store for their current project and how much this will grow in the next five years. About 11% of the respondents really had no idea about the amount of data they currently store and 17% did not know how to estimate the amount of stored data in five years. The biggest group of researchers currently store between 1GB and 1TB (40%), while 6% store even more data. This indicates that the total amount of stored data by researchers is enormous. More importantly, this amount will only increase in the future. For the HEP community, several tens of Petabytes per year will be generated in coming years, mostly a result of experiments with the Large Hadron Collider (LHC). In the area of Earth Observations the number of satellite fleets by the European Space Agency (ESA) will be massively increased during the upcoming missions. This will oblige the

---

[15] PARSE.Insight Case studies report, section 5.3 What aspirations were formulated? - Is there a gap between wishes and reality? Pag. 118.

space agencies to deal with an exponential growth in the volume of EO data archives (from about 3PB archived today up to 20PB expected in the next 10 years)[16].

When asked where researchers keep their research data, the most common 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%). Only 20 % of the respondents submit data to a digital archive. This is a telling figure which is even more meaningful in the context of the questions relating to sharing data.

As it turns out, 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. 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. 11% of the respondents only 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. This conclusion can be justified by the major problems researchers foresee in sharing their data through digital archives: *legal issues* (41%), *misuse of data* (41%), and *incompatible data types* (33%). Based on the responses, it seems that researchers are not familiar with data archives and if they are, there still is a lot of distrust in the capability of digital archives to properly handle research data.

The current practice is not to be explained by a disinterest of researchers in other people's data. The reason for this is that 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 and 40% would like to use 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. In Earth Observations, one of the researchers noted that *"SAR data from the Shuttle Imaging Radar mission C (SIR-C) that flew on the Shuttle twice in 1994 is effectively lost because the computer hardware to read and process the original data is obsolete and no new system has been created to read the data."[17]*. This clearly shows the fragility of research data over time.

The infrastructure available to researchers differs to a large extent per research discipline, as was pointed out by the case studies that were conducted. In the High Energy Physics research

---

[16] PARSE.Insight Case studies report, section Earth Observations Case study, page 44

[17] PARSE.Insight Case studies report, section Earth Observations Case study, page 50

discipline, measurement data are preserved in a number of institutional repositories of research centres and academies. However, these repositories are not equipped for long-term guardianship and hold the data for a few years only beyond the lifetime of the experimental projects (the time judged necessary for the completion of research analyses). In other words, no infrastructure is in place that allows for long-term preservation of research data. Within Earth Observations, the European Space Agency (ESA) is acting as coordinator of preservation at European level, warranting long-term accessibility of its Earth Observations archives. In Book Studies, although it has taken advantage of the digitization of library materials, books and metadata catalogues of recent years, no specialised preservation system is yet in place. In this research discipline, digital research data are stored decentralized, e.g. in repositories of universities, but mostly still only on researchers' private or official computer.

3.1.3.   The outlook and the need for an international infrastructure

The amount of data is growing, while currently far from all researchers store their data in digital repositories. This can partly be explained by a lack of trust in those digital archives and shows that part of the problem is psychological, but it may also be that researchers are unfamiliar with existing digital archives or that there are simply not enough archives.

PARSE.Insight is based on the premise that an e-science infrastructure might deal with many of the threats to preservation of digital research data. It is an important observation to note that 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. This is supported by the results of the three case studies. For HEP 83% of the theorists and 73% of the experimentalists support the idea for an infrastructure, in EO 94% of the responding researchers agree and in Social Sciences and Humanities 75% supports it.

When asked what such an infrastructure should look like, the answers are anything but uniform. Many researchers admit 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.

3.1.4.   Roles and responsibilities

A majority of the researchers believes that their national government should pay the bill for the preservation of research data (61%). As alternative, the researchers' organisation (41%) and the European Union (36%) are mentioned. If publications are treated separately, the responding researchers still think the national government is the most likely organisation to fund preservation activities (57%). However, in case of publications another group believe that the brunt of the costs for the preservation of publications should be borne by publishers (42%) or the research community itself (35%).

It is interesting to see that researchers attribute a greater role to the government (national or EU) when it comes to the preservation of research data, compared to publications. Researchers might assume here that the preservation of research data is something that is not or should not be done by private institutions and that the government should take a leading role to safeguard data over the long term.

3.2.   Data managers

3.2.1.   Perceptions

Six of the seven reasons formulated were regarded as either important or very important by 76% to 98% of the respondents dealing with data management. 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. Data managers find the *potential economic value* of research data a bit less important. Still, 62% of the respondents regard economic value as either an important or very important reason.

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.

3.2.2.   The state of affairs

Of the various kinds of digital material data managers store, doctoral and diploma theses are among the most preserved (69% of data managers claim to store those), 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%) are preserved less often.

The data formats that are currently used to store data are most often digital still images (81%), office documents (74%), and audiovisual materials (46%). As it turns out, there is quite a difference between what researchers use and data managers store. It seems that researchers do not provide data managers with all types of data they make use of, e.g. source code and raw data sets. In other words, some kind of selection already seems to be taking place at this stage.

When asked whether their organisations have policies and procedures in place which determine what kind of data is accepted for storage and how and when it needs to be submitted, 63% of the

respondents answered affirmative. Nevertheless this still means that 37% of data managers do not have such policies in place or are not aware of that.

Looking at what these policies entail, it is interesting to see that liability arrangements are often not specified when data is lost or affected. Only 34% of data managers acknowledged that their policies for storing data include arrangements for liability, which might be an important reason for researchers for not storing certain kinds of data. Furthermore, 73% of data managers pointed out that within their policies it is not possible to see who has previously enhanced, annotated or had access to the data. 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.

3.2.3.   The outlook and the need for an international infrastructure

The greater part of data managers do not think they are well prepared for the future regarding digital preservation of research output. 59% of the data managers argue that the tools and infrastructure available to them do not suffice for the digital preservation objectives they have to achieve. Less than half of them (47%) think that their current infrastructures will scale with future requirements, while 35% do not believe they will.

A comparable number of data managers (60%) agree with researchers that an international infrastructure for data preservation and access should indeed be built to help guard against existing threats. Data managers seem to agree with researchers on the idea of building some sort of central international repository/database for research data.

3.2.4.   Roles and responsibilities

According to data managers, the National Library (71%) is the organisation of choice to take on responsibility for preservation of research data. Next to the national libraries, the researcher's institute (60%) and research libraries (56%) are also regarded as important players in the organisation of digital preservation. The question as to who should be paying for it all, is answered as follows: national government (77%), research funders (51%), and the EU (42%).

3.3.   Publishers

When analysing publishers' responses, it is important to investigate not only the number of publishers but also the amount of journals they represent. Some 25,400 peer-reviewed journals

are published world-wide[18], 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 publishers with only one or a few titles on their list. Therefore, we differentiate the results on small (less than 50 journals) and large publishers (more than 50 journals). When percentages are mentioned, the first figure denotes the <u>small publishers</u>, the second the <u>large publishers</u>. If possible, statistics are also translated into the number of journals covered.

3.3.1. Perceptions

The most important reason for preservation regarded by publishers is the *stimulation of the advancement of science*. 96% of both small as large publishers regarded this either important or very important. The top three of most important reasons for the publishers is completed by *future validation purposes* (92% of small publishers and 88% of large publishers) and the possibility of re-analysis of existing data (92% and 96%). Publishers in both sizes seem to 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 little disagreement between large and small publishers on the most important threats to digital preservation. When looking at the answers 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%. Without such a custodian, publishers would not be assured of having access to the research data which are part of the added value of their scientific papers, which explains why this is regarded to be such an important threat to preservation. Both equally (72%) fear that the *lack of sustainable hardware, software or support of computer environment may make the information inaccessible*. On a third place, 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 always reply similarly. The most noticeable difference of opinion is apparent in their response to the 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.

3.3.2. The state of affairs

---

[18] The STM report – An overview of scientific and scholarly journal publishing, September 2009, available at: http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf.

If we look at the number of journals covered in the responses to the survey by publishers, we count approx 8,500 journals (or a third of all scientific journals published). Of these 8,500 journals, more than 90% are covered in preservation policies and measures. A similar percentage is subject to disaster recovery measures. These measures regard first and foremost the official publications.

For underlying or supplementary material as sometimes added by the author to the journal manuscript, the situation is less well arranged.

Within the wide range of formats that authors submit, the traditional office documents are accepted by a majority of publishers, as well as digital still images. Less than half of the publishers accept scientific/statistical data formats and multimedia data such as audio and video. This holds for both large and small publishers. Databases, source code and raw data are less frequently accepted by publishers. And when these research data are stored, in most cases they are not subject to any explicit preservation arrangement. Almost 70% of all publishers asserted that they do not have any preservation arrangements for the underlying digital research data in place, they treat this material similar as the journal articles. In terms of number of journals this reflects about 8,000 journals for which the underlying data is not covered by a specific data preservation policy. When it comes to the preservation of digital publications on the other hand, almost 93% or around 8,500 of the published journals are covered by their publisher's preservation policies.

Publishers deploy different preservation strategies. Large publishers often outsource long-term preservation of their journals to a third party (52%). Only 23% of the small publishers responded to have arranged a preservation strategy. When outsourced, Portico, CLOCKSS/LOCKSS and the e-Depot of the national library of the Netherlands are external parties that were mentioned most (ranging between 25% and 44%). If publishers do not outsource their preservation activities, normalisation (i.e. converting file formats to common de facto standards) is mostly used. Migration is used less as a strategy (between 15% and 28%) and emulation only in 8% of the cases.

3.3.3.   The outlook and the need for an international infrastructure

As the publishing world is in motion, PARSE.Insight asked publishers about their opinion on what the world will look like in the near future. Several scenarios were presented of which the hybrid model was opted for most (32% of small and 43% of large publishers). This scenario states that *the current journal model will remain dominant, but that it will consist of a combination of subscription-based and open access journals*. Next to this, 19% of the small publishers and 26% of the large ones believe that the publishing process as such (based on journals, peer-review, etc) will not change much. Small publishers though, had a higher score (20% vs 11%) on the scenario stating that *most research results will be Open Access and available for free via institutional repositories*. Of all publishers, 63% believes that journals will

become more interactive and will include multimedia and containing more underlying research data in all kind of formats.

Given this assumption, it is likely that publishers will extend their preservation policies to all kind of media published by them.

What is clear from the responses is that out of the surveyed publishers, the greater part (between 64% and 75%) are convinced that some kind of international infrastructure for data preservation and access should be developed. This is a high figure, compared to researchers and data managers, but can be explained by the necessity of publishers of having access to research data to cross-link scientific journals with data. Obviously, an international infrastructure would be very helpful for publishers since they would be able to adjust their publishing practices to it.

Looking into views of what such an infrastructure should look like, it is worth noting that publishers specifically describe 'libraries' as part of the infrastructure solution and value 'access' and 'standards' explicitly.

3.3.4.   Roles and responsibilities

When it comes to the preservation of publications, publishers see themselves as the primary responsible party (69% - 73%). National libraries (59%-66%) and specialised external organizations such as Portico (31-52%) are also mentioned as important parties. The majority of the publishers believe that preservation of publications should be paid for with public money. Between 56% and 63% of the publishers assume that national governments should pay for digital preservation. Between 41% and 46% of the publishers believe that the national library should carry the financial burden of digital preservation. Yet publishers also assert they carry a financial responsibility themselves (between 42% and 47% of the publishers believe so).

Preservation of other research output (e.g. underlying data sets), is the main responsibility of researchers themselves (48% - 52%), say publishers. If the author is not held responsible, then at least the author's institute should make sure the research data are kept safe (43%). Third important after the researchers and their institutes are the publishers themselves. And similar to the responses in the other categories of stakeholders (researchers and funders), it is the national government that should ultimately pay the bill (46% - 55%). Second most important for financing digital preservation of research data are the researcher's institutes (29% - 40%). The EU is mentioned as the third most important party (18% - 27%).

3.4.   Funders

3.4.1.   Perceptions

As opposed to researchers, data managers and publishers, a large number of responses on this topic could not be gathered from funders. However, from several funder policy documents that are available in the various countries in Europe on the topic of the management of research outputs, some statements on the issues funders regard as relevant can be made. It must be specified here that most of the documents primarily address the issue of (open) access, rather than long-term preservation per se. In many cases preservation seems to be an important long-term goal that is assumed to be reached when access is taken care of properly. Data archiving in openly accessible repositories is then regarded to be the vehicle by which access can be achieved and by which research data will be available over the long term. This is a misconception as short-term access is no guarantee for access on the long term.

Dissemination and exchange of information are mentioned as important advantages of free access to scientific information[19]. Maximized dissemination as a result of open access, might lead to increased impact of the research results and citations[20]. It will increase research effectiveness across discipline boundaries[21]. Furthermore, the efficiency of scientific discovery will be improved when access to research data is secured over the long term, maximizing the return of investment in research by public research funding bodies[22]. Moreover, publicly-funded research data are a public good, produced in the public interest[23]. Finally, secure long-term preservation will provide opportunities for independent verification and fresh analysis[24].

From the seven reasons mentioned earlier, most of them are thus recognized by funders as well, except for the uniqueness of research data which was not mentioned explicitly. In one of the interviews conducted with research funders it was even asserted that in some research disciplines the uniqueness of data is not important at all, since most experiments can be reproduced at any time[25]. Obviously, the highly complex HEP experiments as well as Earth Observation

---

[19] http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_Recommendations_OpenAccess_200805.pdf.

[20] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/ESRC_Open_Access_Repository_Policy_tcm6-31316.pdf.

[21] http://www.rin.ac.uk/system/files/attachments/Research-funders-outputs-report.pdf.

[22] ftp://ftp.cordis.europa.eu/pub/fp7/docs/open-access-pilot_en.pdf.

[23] http://www.bbsrc.ac.uk/web/FILES/Policies/data_sharing_policy.pdf.

[24] http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf.

[25] Interview with Deutsche Forschungsgemeinschaft (DFG).

measurements on a long time-baseline have to be regarded as important exceptions and other cases will exist as well.

The economic value of research data, re-analysis of existing data, future validation, advancement of science, interdisciplinary collaborations and the public property of research data are definitively acknowledged by funders (i.e. supported by all interviewed funders).

On the topic of threats to preservation, some important statements are made by research funders as well. The Research Councils United Kingdom (RCUK) for example, stresses the possibility of misconduct by individuals within the research community (i.e. researchers) in the execution of research as well as the preservation of research data, e.g. through deliberate intention, recklessness or gross negligence[26]. This may lead to users not being able to understand the research data or to identify the location of the data[27].

Next to individuals who fall short, another potential threat to preservation concerns interoperability. According to the OECD, technological and semantic interoperability is a key issue for long-term international access to and use of research data[28]. As long as there is no broad international agreement on interoperability or appropriately designed technological infrastructure, digital preservation will not be very effective.

The risk that *users may be unable to understand or use the data in the future due to semantics or formats* was regarded as the most important threat to digital preservation in one of the interviews that were conducted. The observation that even researchers themselves sometimes do not understand the data to the full extent any more after some years was argued to be the primary reason for this view point and emphasizes the need for proper digital preservation[29].

Another threat to digital preservation is related to the hardware that is used for storing data, which changes over time, may result in loss of information. The example of floppy disks that were no longer readable when new computers were developed was mentioned several times and shows how changing hardware can be problematic for preservation of research data. It stresses the fear of not being able to access research data, even while preservation was assumed to be taken care of.

---

[26] http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/goodresearchconductcode.pdf.

[27] The RCUK regards researchers to be responsible for the preservation of research data (together with their research institute) and has specified the requirements researchers need to attend to concerning the management of research output in their Code of Conduct.

[28] http://www.oecd.org/dataoecd/9/61/38500813.pdf.

[29] Interview with Czech Republic Academy of Sciences.

Finally, it was also acknowledged that some fear exists that *the current custodians of data may at some stage fail and let us down*. The loss of NASA's original records of the first moon landing is such an example, in which better preservation policies might have made a big difference[30].

> *"The problem with floppy disks and the loss of NASA's records of the first moon landing are two of the most striking examples of what can happen when digital preservation is not taken care of properly."*

3.4.2.   The state of affairs

In contrast to researchers, data managers and publishers, for research funders it is not directly obvious how digital research output should link to their activities. However, it is increasingly being recognized that funders also play an important role in the preservation of research output. It is sometimes even asserted that there is a central role for research funding bodies in Europe[31].

Research funders have been adopting various policies for publicly funded research. Generally, these policies tend to express the need for open access and archiving rather than long term preservation per se. At least in 11 European countries, open access policies have been adopted in some form, while in others open access policies are currently being developed for publicly funded research[32]. The differences in these policies can be on various levels. Some funders for example enforce archiving of their funded research results in openly accessible archives, while others only encourage this behaviour. In Germany for example, encouragement is used rather than enforcement, simply because for some research disciplines no repositories exist.

> *"When no repositories exist, how would it be possible to enforce researchers to archive their research data?"*

Researchers in Germany are encouraged to archive their research data via a bottom-up approach, while awareness of the importance is promoted. Best practices of archiving are rewarded[33]. In

---

[30] Interview with Dutch Ministry of Education.

[31] http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

[32] Austria, Belgium, France, Germany, Ireland, Italy, Norway, Spain, Sweden, Switzerland, and the United Kingdom. The Netherlands Organization of Scientific Research has spoken out in favor of open access and is developing policies accordingly.

[33] This was acknowledged in an interview with DFG.

Spain on the other hand, where it is argued that it should not be expected that preservation activities are organized from within the research community, a top-down approach is suggested[34].

Most funders specify which data should be deposited, but not always. When it comes to the depositing of publications, for some funders it is a requirement that the publication is peer-reviewed, while others accept all publications. The file format in which publications are submitted is sometimes specified, but often not. Finally, some funders specify when data or publications should be archived (e.g. within six months after publication), while others leave that up to the researchers themselves. In other words, there is no such thing as a standard policy for open access and archiving[35]. Cohesion between the policies of funding organisations in different countries seems to be missing. It is surprising that these national funding organisations are not aligned internationally, being important research funders in Europe.

While most funders require that their researchers make sure the research data are openly accessible and in many cases also that publications are archived, explicit requirements relating to digital preservation are less frequently made[36]. Long-term digital preservation is in the picture, but often regarded as the next step[37]. The line of thought behind this could be that funders are of the opinion that open access first needs to be taken care of, before long-term preservation becomes worth thinking about[38]. Or, in the words of a research funder:

> *"What use is preservation when nobody is able to access the research data?"*

It might also be the case that funders assume that by archiving in openly accessible repositories, research output will at the same time remain preserved for the long term, while this is definitely not the case. Without proper strategies on how to preserve research output, open access by itself is not of great value, which is something that is not being recognized explicitly by many research funders. The concepts of open access and long-term preservation seem to be confused within the funder community, urging the need for more awareness on the topic.

---

[34] Interview with the Spanish Research Council.

[35] For a detailed overview of open access and archiving policies in various European countries, see http://www.sherpa.ac.uk/juliet/index.php?sortby=name/.

[36] For a useful example of the differences between research funder policies regarding open access and preservation in the UK, see: http://www.rin.ac.uk/system/files/attachments/Research-funders-outputs-report.pdf.

[37] This was acknowledged by in the interview conducted with the Dutch Ministry of Education.

[38] Interview with Spanish Research Council.

Only in the UK were some formal and explicit statements regarding long-term preservation found. The Economic and Social Research Council (ESRC) for example, stipulates to ensure that all research outputs are accessible and readable for a minimum of 15 years after deposition[39]. The UK partnership of seven research councils (RCUK) states that this minimum should be ten years for any research project, while the data of projects of clinical or major social, environmental or heritage importance should be retained at least twenty years[40].

The absence of many explicit and formal policies relating to digital preservation does also not mean that preservation is not being put into practice. It might well be the case that researchers do make sure their research data and publications are preserved over the long term and that their funders have a role in that, while it is not formally specified in any policy document.

There are strong indications that this is the case. In the large scale survey that was sent to researchers, they were asked whether their funding organisation required them to preserve digital research data. Interestingly enough, quite a number of researchers stated that they were required to take care of preservation by their funders. Of all respondents (more than a thousand), almost 25% indicated that this was the case with their funder (see figure 3).
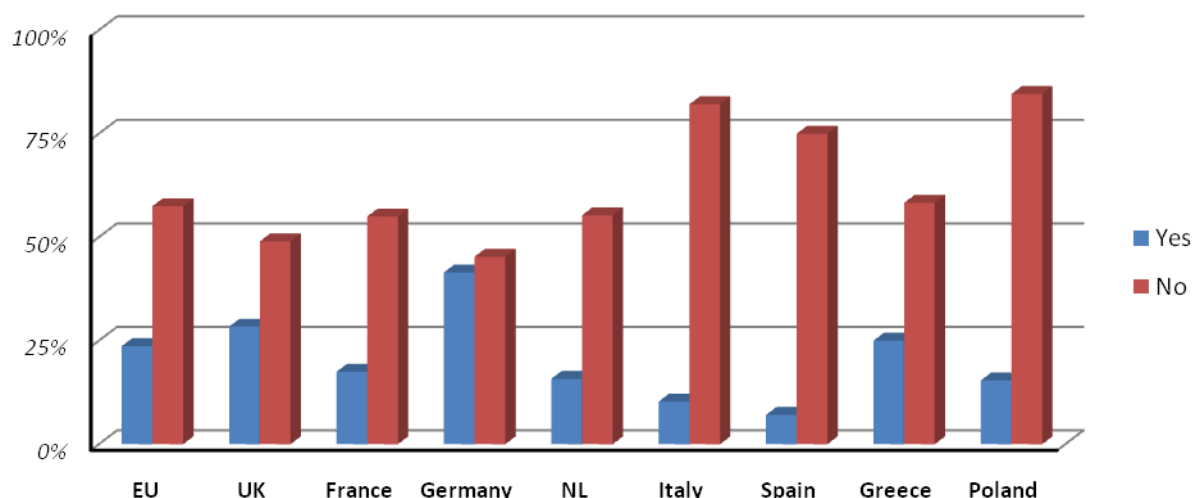

Figure 3: Do funders require researchers to take care of digital preservation? (N = 1,022)

In Germany, the figure was highest (41%) while in Western Europe in general the requirements of funders seems to be somewhat more stringent than in Southern and Eastern Europe. Looking at the differences in availability of policy documents and the adoption of open access and

---

[39] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/ESRC_Open_Access_Repository_Policy_tcm6-31316.pdf.

[40] http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/goodresearchconductcode.pdf.

archiving policies, it is not such a surprise to find here that these regions are not on the same levels as their Western European counterparts. In an interview conducted with an Eastern European research funder, it was confirmed that policies on digital preservation are only starting to be developed. When digital preservation is taking place, it is primarily organised bottom-up from within the research communities without real pressure or incentives from the national research funders[41].

When funders require their researchers to preserve their results and data digitally, it is remarkable to see that in less than 40% of the cases the researcher is given explicit instructions on how to do so. When asked whether their funding organisation provides them mandatory procedures for managing and preserving research data, in most of the cases the answer is no. Apparently, funders do quite often ask researchers to make sure they preserve their research data properly for the long term; how they do it is mostly left up to the researchers. Again, staggering differences appear when comparing countries in Europe. Whereas in Western Europe researchers claimed to be provided with procedures from their funders, this was far less often the case in other parts of Europe (see figure 4).
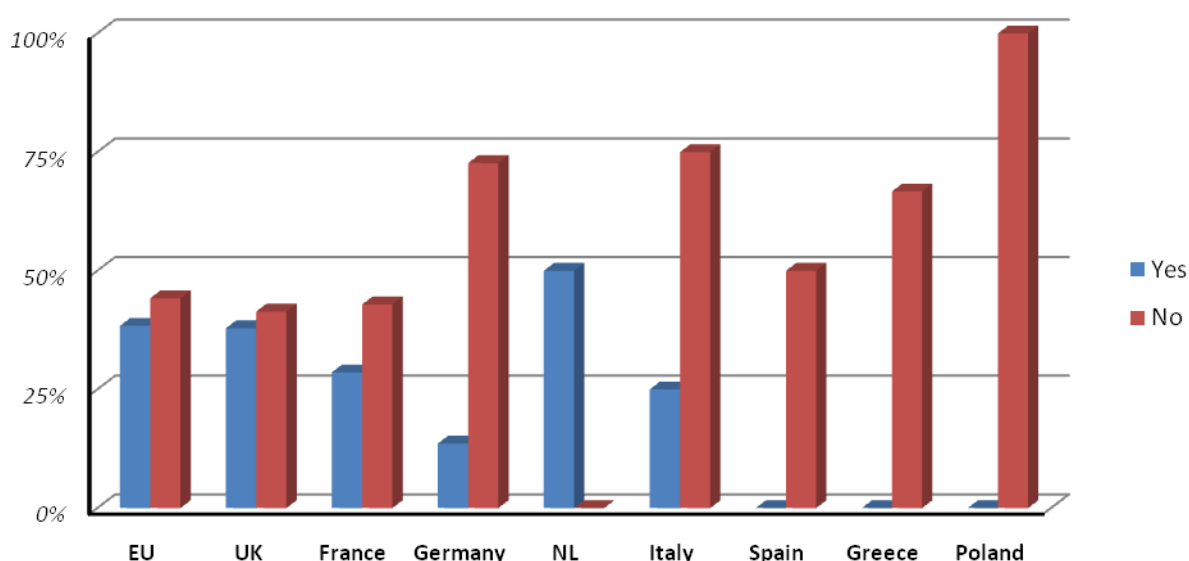

Figure 4: Do funders provide researchers with procedures for digital preservation? (N = 242)

---

[41] Interview with Czech Republic Academy of Sciences.

3.4.3.   The outlook and the need for an international infrastructure

As was mentioned before, most of the attention of funders currently is focused on open access instead of on long-term preservation. The importance of long-term preservation is acknowledged, but the greater part of funders does not explicitly require researchers to ensure their research is preserved. And if they make statements about archiving, they encourage rather than enforce this behaviour. However, policies regarding digital preservation are being developed in many countries and several ideas are ready to be tested to stimulate digital preservation.

For example, it was suggested that it could be an interesting idea to hold back a proportion of the research funding until digital preservation is realised and in this way make sure research data are preserved over the long term[42].

> *"The carrot and stick method might prove useful when it comes to encouraging researchers to take care of digital preservation. Release 80% of the research funding upfront, while waiting with the last 20% until research data and publication are secured for the future."*

Another idea that was put forward is to stimulate researchers to organise digital preservation of their research data by awarding best practices. Incentives would then be used to motivate researchers to archive their research results as well as underlying data sets in repositories, resulting in more cross-references and citations[43].

An important observation resulting from interviews with research funders is that the issue of selection needs to be taken into account as well when contemplating potential preservation infrastructures. Obviously, not all research data can or should be preserved, but how to decide on what is relevant information to store and what is not? The idea of having *expert panels* decide per research discipline which data should be stored was proposed several times[44]. Researchers should be part of these expert panels, since they are at the very core of the process, but funders and policy makers should be included as well.

---

[42] This idea was suggested in an interview with the Dutch Ministry of Education.

[43] The attractiveness for researchers of such a system was confirmed in several interviews with research funders.

[44] Interviews with the Dutch Ministry of Education, DFG and the Spanish Research Council.

Regarding the need for an international infrastructure to ensure an efficient international preservation of research data, it cannot be asserted that within the funding community a strong desire is expressed. At this stage, international agreements on research output only cover the subject of open access[45]. In one of the interviews conducted with research funders, it was asserted that it is too early to be discussing supranational organisation of digital preservation[46]. Central coordination on European level of what is being done in the field of digital preservation on the other hand would definitively be welcome.

3.4.4.   Roles and responsibilities

Within the funding community, different views exist on the responsibilities and the role of research funders relating to digital preservation. As was discussed before, various policies have been developed by research funders in different countries, encouraging researchers to make sure their research results are accessible and archived. This shows that these research funders regard themselves to be responsible at least to some extent for the preservation of research output.

On the other hand, it is in many cases unclear what measures funders take if researchers do not abide by their rules. Even if measures are in place, for instance withholding a part of the funding until a researcher has deposited his data, it is unclear how these policies are enforced in practice. Also, most research funders do not incorporate any preservation policy issues at the application stage of research proposals, which could motivate the researcher to think about preservation from the beginning. Hence, research funders do not assume responsibility for ensuring digital preservation of research data for the full extent.

Apart from the responsibility issue, another question is that of finance. None of the policies on the topic of digital preservation mention the availability of earmarked budgets/funding for preservation activities. The BBSRC policy does have a paragraph on funding for the sharing of data, which states: *"BBSRC recognizes that data sharing has time and cost implications. Funding to support the management and sharing of research data (for example staffing, physical resources such as storage and networking capability) can be requested as part of the full economic cost of a research project."*[47]

Yet all these funding activities are focused on the researcher and the research projects. None of the policies mentions anything about funding activities after the project's lifetime, i.e. funding for data archives or libraries. It is therefore not clear whether research funders think of themselves as being responsible for financing the digital preservation of their funded research. This was also confirmed in one of the interviews conducted as part of this research project.

---

[45] The Berlin Declaration is an example of international cooperation on the subject of open access by research funders, see http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[46] Interview with DFG.

[47] http://www.bbsrc.ac.uk/web/FILES/Policies/data_sharing_policy.pdf.

Although in the end it is the national governments who finance research, it is assumed to be up to the actual researchers and their institutes as well to make sure their work is kept in a safe place and can still be accessed in the future. Or, in the words of a research funder[48]:

> *"Ultimately, researchers strive for immortality. Should it not be left up to them to achieve it?"*

At the same time, research funders also acknowledge that the primary task of researchers is to do research. While it may be important that preservation is taken care of, it should not result in an additional heavy burden which may be obstructive for the research output itself[49]. Hence, central coordination is required and efficient ways to organise digital preservation need to be looked for.

---

[48] This research funder wanted to remain anonymous.

[49] Interview with Czech Republic Academy of Sciences.

4. Analysis

In this chapter, the most important observations from the research results will be more closely examined.

4.1.   Stakeholder cross analyses

First, the views of the various stakeholder groups with respect to the reasons for digital preservation are compared. Figure 5 summarizes the views of the three stakeholders from which a large number of responses to the above question was collected. Since only qualitative information regarding the views of funders on this topic was available rather than a large set of quantitative responses, funders have not been included in the graph. The numbers on the horizontal axis correspond with the seven reasons listed earlier (see page 16 of this report).
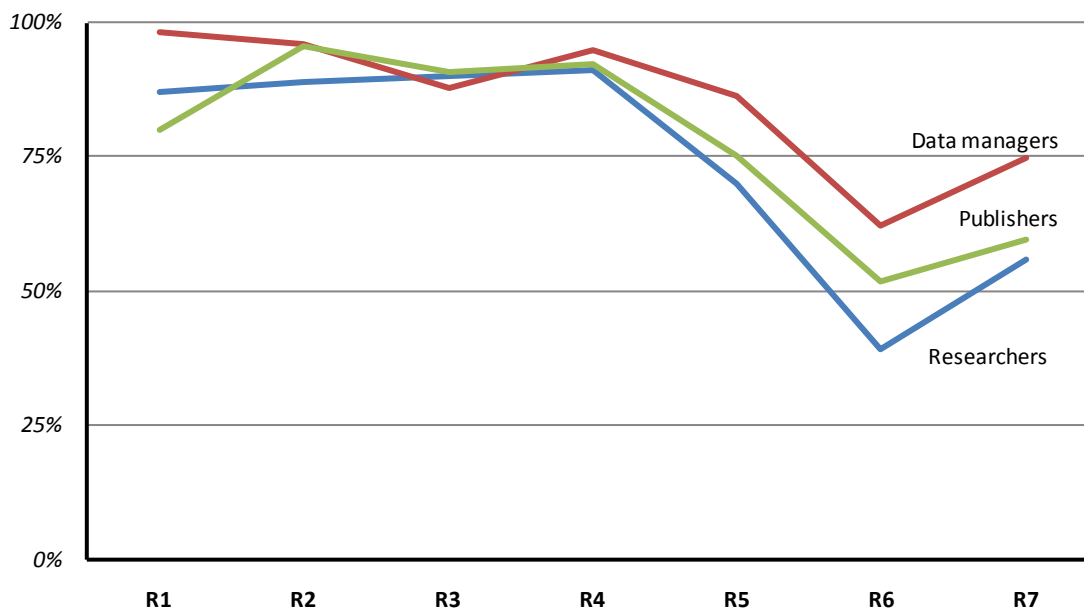
Figure 5: The reasons for digital preservation compared for three stakeholders

In general, the various stakeholders agree on what they regard as important reasons for digital preservation. Digital preservation is assumed to be important, because it allows for re-analysis of existing data, may stimulate interdisciplinary collaboration and it may serve validation purposes in the future. Moreover, it will stimulate the advancement of science. Finally, because research is mostly publicly funded, the results should become public property and hence should be properly preserved. The uniqueness of research data and the economic value are regarded less important by these three stakeholders. Within the funding community however, the economic value of research data is definitively acknowledged as was confirmed by various funders.

Page 35 of 51

Regarding threats to digital preservation (see figure 6), quite some consensus exist as well. Among all stakeholders, there is a fear that we will no longer be able to understand or to use data in the future because of semantics, formats, hardware or software obsolescence or that we will not know where to find the data any more. It is also feared that current custodians of data might cease to exist. The potential risk that access and use restriction policies might not be respected in the future seems to be worrying the stakeholders less. Again, the seven numbers on the horizontal axis correspond with the seven threats listed earlier (see page 17 of this report).
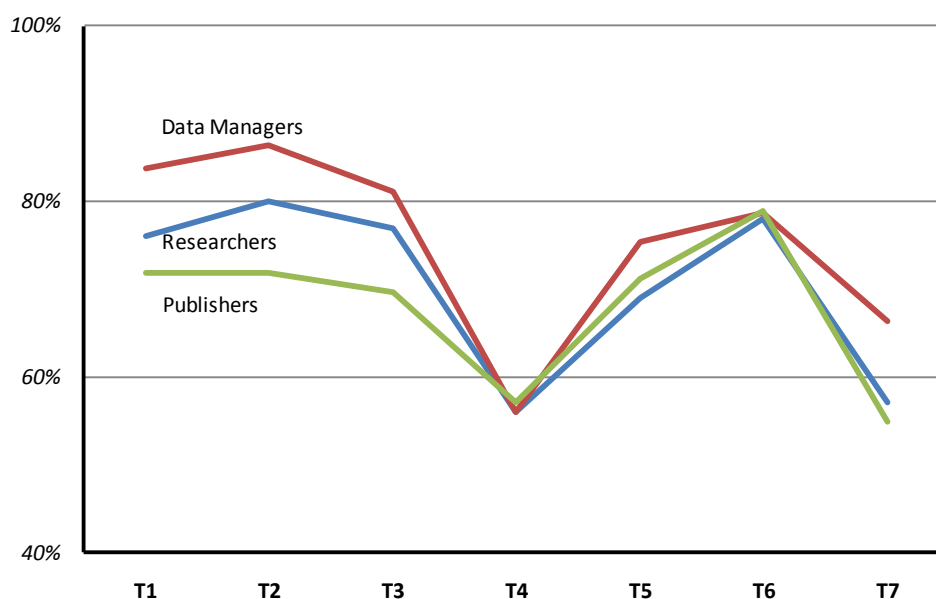

Figure 6: The threats to digital preservation compared for three stakeholders

On the important question of whether an international infrastructure is needed that can help guard us against these threats, the opinions of the stakeholders are unambiguously and unanimously positive.

The stakeholders differ more in opinion when it comes to deciding who should pay for the preservation activities. Figure 8 displays how the stakeholders assess the responsibilities of the four most mentioned parties for financing digital preservation.
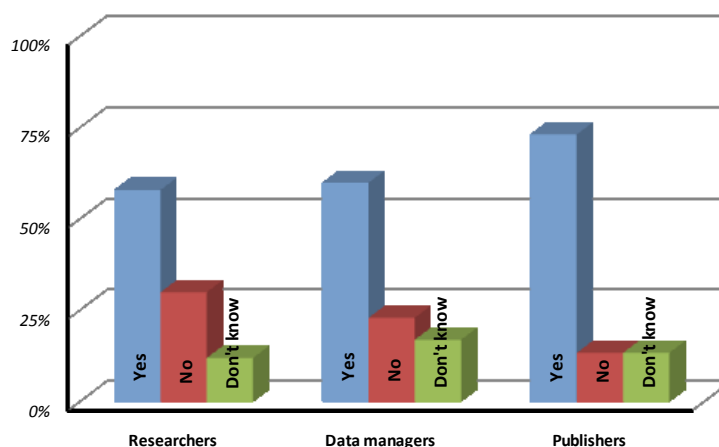

Figure 7: Is an international infrastructure for digital preservation needed?
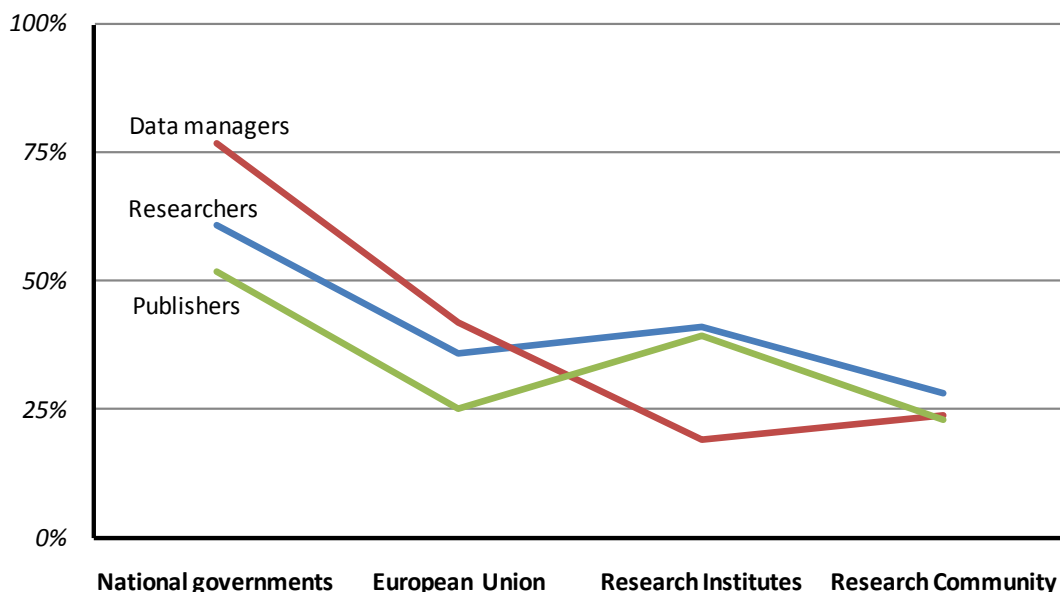
Figure 8: Who should pay for the preservation of digital research data? (multiple answers possible)

All stakeholders unmistakably place responsibility for financing digital preservation on national governments rather than on the European Union. Although most respondents claim that the biggest responsibility is with the national governments, data managers put more emphasis on this financing party than do publishers and researchers. On the other hand, data managers argue that the role of research institutes should be smaller than researchers and data managers assert.

4.2.   Important observations

Other important valuable insights can be found. First, a significant discrepancy exists between the data which researchers actually use and the data they send to data managers for storage or preservation. Network-based data, source code, computer applications and raw data are forms of data that are used by researchers, but not too often received by data managers for preservation. It may well be the case that the absence of liability arrangements (in case data gets lost or damaged) with many data managers, is the reason that researchers fear misuse and legal issues. Researchers do not think data archives are the safest place to store their data and rather keep their information stored on their own computers. As a result, researchers tend not to share their research data, even though most of them acknowledge that they sometimes need digital research data that are not publicly available. Within their own discipline, 63% of researchers that currently do not make use of each other's research data, expressed their interest in this possibility of data sharing, while 40% of researchers argued to be interested in data from other disciplines. It is therefore clear that within the research community there seems to be a strong desire for a shared safe research data infrastructure.

A general observation that relates to this issue is that the focus of most policies of research organisations and policy makers is on open access rather than on digital preservation. The value of open access and data sharing increasingly seems to be recognised within different stakeholder groups. The way in which open access is envisioned, seems to be through encouragement rather than through enforcement. Whether researchers really feel compelled to open up their research data and make their results openly accessible remains unclear. Digital preservation is often regarded to be the next step, after open access, even though the importance of long-term preservation is definitely acknowledged. However, focusing on access only will not be enough to ensure that digital data will still be accessible for future generations. Crucial to this is that the difference between 'short-term' access and long-term preservation should be better understood. Especially among research funders, the two concepts are often confused. Hence, awareness creation of what long-term preservation actually means is a must.

Another reason why digital preservation is not yet attended to so much may be that there is no clarity about which data should be preserved and which should not. It should be clear that digital preservation starts with selection, because without limits data preservation would undoubtedly lead to a data overload and control would be a serious issue to worry about. But how can one decide which data to store and which not as this is a highly discipline-dependent decision and also depends on the needs of future users.

Moreover, the responsibilities regarding long-term preservation are in many cases not clear. Surely, it is the researchers who create the research output and publishers that disseminate publications, but who should make sure they remain accessible and preserved over the long term? The effectiveness of data archiving and preservation policies vary to a large extent per research discipline and depend on the willingness of researchers and the availability of repositories. As was discussed earlier, in some research disciplines there seem not to be any repositories and the profession of data curator is simply non-existent[50].

> *"The profession of data curator is missing."*

Another important observation relates to the role of funders within the research value chain. As was shown earlier, research often only commences when funders have agreed on financing a particular project or when they have actually released the funds. In this role, they are at the very starting point of any research project and have the ability to approve or reject project proposals. In this role, they thus have the ability to make certain demands regarding the way this research is going to be conducted. In other words, research funders can influence researchers on where and how to archive their research results. However, this is not often found in practice, although encouragement or even enforcement of preservation was proposed by several research funders in

---

[50] Interview with DFG.

interviews[51]. Demands regarding digital preservation are rarely part of the project proposal criteria. Even not for this European project. In this way, digital preservation is not being attended to the full extent.

A final observation concerns the different policies and procedures which relate to digital preservation in Europe. As was discussed before, no uniform policies exist. No consensus exists on questions such as what, when and how data have to be preserved. At the same time, both research institutes as policy makers and funders are developing their own sets of policies, while an integrated approach seems to be missing. Also, quite some differences seem to exist between regions in Europe in the development of these policies. While on central European level and in Western European countries policies or at least the best intensions have been made explicit on the requirements for digital preservation, less attention has been paid to this topic within Southern and Eastern European countries. This was not only reflected by the absence of explicit policy documents on digital preservation in these countries and the interviews that were conducted, but also became clear from the responses of the survey results among researchers in the various countries. Researchers from Western Europe are required to take care of digital preservation by their funders more often and have to follow certain procedures, than their counterparts from other parts of Europe.

---

[51] In the interview with the Spanish Research Council, the term "strongly encouraging" was used as an alternative to "enforcing".

5. Conclusions and implications for the Roadmap

5.1.   Conclusions

From the analysis, some straightforward general conclusions can be drawn. First, all stakeholders within research agree that there are several strong reasons for digital preservation. The reasons that are regarded most important are:

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

Then, the stakeholders argue that some serious threats exist that need to be tackled, in order to ensure digital preservation. The most important threats are:

1. users may be unable to understand or use the data e.g. the semantics, format or algorithms involved;
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;
4. loss of ability to identify the location of data;
5. the current custodian of the data, whether an organisation or project, may cease to exist at some point in the future;

Furthermore, the various stakeholder groups assert that an **international infrastructure is needed** that can help in guarding against these threats. Moreover, digital preservation should primarily be financed by national governments and the European Union.

5.2.   Implications for the Roadmap

Apart from these general conclusions, some more specific outcomes can be addressed as well. As introduced in Chapter 3, we will discuss these outcomes in terms of four focus areas.

5.2.1.   Awareness

Long-term preservation seems to be confused with open access or even with access in general. It might sometimes be assumed that by having research output archived in accessible

repositories, long-term preservation will automatically follow. However, this is not the case. Digital preservation arrangements are a requirement to ensure access in any form over the long term. This is independent from the business model chosen for access. Therefore, more emphasis needs to be put on preservation instead of (open) access only and awareness needs to be created of its importance.

Awareness should be raised on the development of policies within different countries. Currently, effort is too fragmented leading to different views and implementations on what, how and when digital research output needs to be preserved.

5.2.2.   Rules and best practices

As was exposed in this research, organisations dealing with data are not always regarded to be the safest place to store data, while liability arrangements are often lacking. As a result, researchers are not archiving their research data to the full extent and sharing is not taking place so much, even while it is desired by the researchers themselves. An obvious solution to this problem is to make sure these data archives become safer and that **liability arrangements** are made explicit, so that researchers can be more confident that their research data are in good hands and that misuse of their data will not be tolerated. Only when this confidence in data archives will have been increased, can the idea of encouraging or even enforcing researchers to deposit their research data be given consideration. This requests the development of **audit and certification procedures**.

For publications specifically, publishers and libraries made a good step forward in preservation. By now, **most publications are covered by a preservation policy** and publishers have taken preservation measures by arranging agreements with several international 'safe places'.

As was put forward in the research as well, **the role of data curator needs to be formalized** and **responsibilities need to be made explicit**. A central role here needs to be attributed to the national governments, who can coordinate national preservation activities, but to the European Union as well, who can coordinate international cooperation efforts.

The amount of data produced and used by researchers is enormous and is only going to increase in the future. Not only does this stress the importance of digital preservation, it also points out that **selection is a highly serious issue** that needs to be included in the debate. Without careful selection of what exactly we want to preserve digitally, it will never be clear how we need to give shape to future policies regarding digital preservation. A premature suggestion might be to **form expert panels within the various research disciplines** to start thinking about this

selection issue, since every research discipline is different and has different needs for digital preservation[52].

Central coordination is required for several reasons. Most importantly, policies regarding digital preservation are being developed in different countries at the same time, resulting in countless different policies and confusion among researchers, especially because research communities are not tied to national borders. Consequently, in some regions the development of these policies has advanced much further than in other regions. Efforts are being duplicated, while the results are not unambiguous. It may be clear that coordination between research funders and policy makers in this respect would be desirable.

Although most researchers are aware of the importance of long-term preservation, it might not be a bad idea to stimulate preservation activities by **rewarding best practices**. This way, groups of researchers can learn from each other's preservation activities while their data sets become better accessible, resulting in more citations and cross-references.

5.2.3. Technical infrastructure

Although all stakeholders agreed that a science data infrastructure is required, the current situation of such an **infrastructure is fragmented and diverse**. In some countries, data repositories have emerged (national, institutional or discipline-specific) but coverage is weak. There is **little conformation on data formats** and **persistent linking is often not possible**. This makes preservation very hard. Moreover, descriptive information about the data itself is not structured leading to the risk of misunderstanding the data over a longer term.

With several safe places for publications in the world, this kind of research output is mostly well-preserved today. However, when an official publication heavily relies on underlying research data **there should be an opportunity to link to that data**. And conversely, any data that leads to a publication should be made accessible to the readers of that article. It would be helpful if all stakeholders involved, from researchers, to journal editors, to publishers to data managers come to common practices and conventions for that.

5.2.4. Funding

A **more prominent role should be attributed to the funders of research**. Since they are at the very beginning of the research chain, they should take the opportunity to pose more serious demands to researchers concerning digital preservation. **Withholding part of the research funds** until both publication and underlying research data are carefully archived and openly accessible is a very obvious method by which research funders can contribute to the long-term

---

[52] This idea was suggested by both the Dutch Ministry of Education and the Czech Republic Academy of Sciences.

preservation of research results. Moreover, research funders should **make clear policies** explicit regarding what, when and how research data need to be archived, so that no misunderstanding can arise on these issues amongst researchers. Although the current focus of research funders is primarily on open access rather than on preservation, it is argued here that research funders should be urged to **expand their horizons to digital preservation** and take on their responsibility as gatekeepers of research projects. Also, they should be more explicit towards researchers on the procedures of preservation.

5.3. Overview of insights & recommendations

Based on section 5.2 *implications for the roadmap*, the following table summarises the insights gained during our research and lists the desired actions to be taken, appointed to actors and relating to the PARSE.Insight roadmap. The abbreviations used in the last column stand for Organisational (ORG), Policy (POL), Technical (TEC) and Financial (FIN).

Insights	Desired courses of action	Actors	Relation to roadmap
Existing policies are mostly focused on open access rather than on long-term preservation. Open access and long-term preservation are also often confused.	**Awareness** Create understanding for the concept of preservation: as an objective to ensure accessibility of information for the long term future. Avoid confusion with Open Access which is about access entitlements. Preservation, or long-term access is important regardless of the access entitlements. Those with a main interest in Open Access should understand that Open Access is equally dependent on digital preservation, just like all other (non-open access) digital information. **Rules & best practices** Show that long-term preservation is worthwhile for different stakeholders. Cases can be scientific advantages for researchers, efficiency gains for funders, etc.	National governments, EU	ORG: awareness raising via consortium-based organisational structure and/or EU POL: focus on existing standards (OAIS, certification)
Various policies for preservation of and access to research output are developed at the same time. Duplicated efforts result in little efficiency.	**Rules & best practices** An integrated and international approach is desired, in which policies are geared to one another to ensure efficiency and rapid development of policies. Identify and analyse needs of stakeholders, focus on shared interests and cooperation possibilities.	National governments, EU, data managers	ORG: exchange best practices via consortium-based organisational structure and/or EU POL: deployment and adoption.
Some countries are far behind compared to others.	**Awareness** Make organisations aware that digital preservation is not only a technical challenge but also requires adjustments to their policies and procedures. **Rules & best practices** Best practices in preservation should be identified and shared, so that organisations and countries can learn	National governments, EU	ORG: exchange best practices via consortium-based organisational structure and/or EU POL: stimulate policy development in all countries.

	from each other and gaps between countries will become smaller. Build a cross-domain virtual platform for researchers to learn about best practices in sharing and archiving of data.		
Initiatives on preservation are taking place worldwide. Alignment is not yet reached.	**Awareness** Reach out and ensure cross continental collaboration with projects and organisations such as in US, Asia and Australia. **Funding** Assign special budget for international cooperation.	EU, data managers	ORG: deployment and adoption of infrastructure. FIN: business models
Weak coverage of digital repositories. In some research disciplines no digital repositories are present.	**Awareness** Encourage storage of data by creating awareness of available repositories. **Rules & best practices** Agreements between research institutes and data managers for archiving research data. Develop common standards. **Technical infrastructure** Make sure there are enough trusted digital repositories that can deal with discipline specific data. **Funding** Inject money into trusted organisations on national level or domain specific for the creation of trusted repositories	National governments, EU, Data managers	ORG: encourage data managers, encourage and enforce researchers. TEC: development of infrastructural components. FIN: national and European funding sources
No commonly accepted procedure for certification and accreditation of safe and reliable repositories exist.	**Rules & best practices** Start with a basic set of requirements for becoming a trusted digital repository. Support the development of an international standard and certification process.	National governments, Data managers	POL: development of audit and certification standards.
Limited use of those repositories that are available. In most disciplines the majority of researchers keep their research data on personal computers at work or	**Technical infrastructure** Support researchers with easy-to-use tools for data selection and archiving preparations (limiting misinterpretation of data and fragility of data formats). **Rules & best practices**	Data managers, Researchers	ORG: carrots and sticks towards researchers. POL: certification. FIN: earmarks, requirements on

departmental storage media, without reliable preservation efforts.	Liability arrangements must be provided so that researchers can feel comfortable when archiving data. Agree on a checklist for digital archives to become "trusted". Develop training courses to teach researchers how to cope with digital data (awareness raising). Develop e-learning modules for training researchers to work with data sets and how to archive and share them. **Funding** Make preservation of research data a requirement for receiving a research grant.		funding.
Researchers are reluctant to share their data. Most prominent arguments heard are lack of awareness of available repositories, afraid of misuse/misunderstanding of data and legal issues.	**Awareness** Work on incentives to encourage researchers to share their data, such as giving the possibility to receive credits for that. **Rules & best practices** Define and apply standards for openness of data (aka *Creative Commons* for data). Define and apply standards for exchange of data sets across research institutes and repositories. Demonstrate linking and citability of data sets within and across disciplines. Work on improvements of the *Community Framework* (legality) for preserving and sharing data. **Technical infrastructure** Ensure that data sets can be persistently linked, similar as currently is happening for online publications. Implement DRM techniques to ensure that only those allowed can access the data. Furthermore, take into account that data can contain private information which needs to be made anonymous. **Funding** Develop business models for safe and easy sharing of data.	Data managers, Industry (e.g. publishers), Legislation, Funders	ORG: business models. POL: standards. TEC: DRM, persistent identifiers, trusted digital repositories, knowledge gap manager.

Data is kept in a huge variety of file formats. Only a few disciplines have managed to work towards a common standard.	**Rules & best practices** Develop guidelines for researchers and their institutes to come to common practice data formats which are suitable for archiving. Work on development of new *de facto* or *de jure* standards for data sets. **Technical infrastructure** Create strategies for retaining access to the data by monitoring the changes on file formats and software and hardware dependencies (*preservation watch* function), make plans to cope with the changes (preservation planning function) and to take appropriate actions if needed (*preservation action* function such as migration or emulation).	Data managers, National governments, Funders, Industry	POL: standards. TEC: infrastructure, emulation/migration strategies
Amount of digital research data will grow explosively in the coming years. However, not all data needs to be preserved, but there is no consensus on what to keep for the long term.	**Rules & best practices** Commission expert panels (for each discipline) to support the selection process of what needs to be preserved and what not. Learn about collection build-up and management taking place at archives and libraries. Define the role of data curator. **Technical infrastructure** Ensure scalable solutions for storage, processing capacity and network bandwidth. **Funding** Funding should scale similar to the technical needs of the infrastructure.	National governments, EU, Researchers, Data Managers	ORG: expert panels, data Curation. POL: standards. TEC: scalable infrastructure FIN: structural funding
No clarity exists about who should do what in preservation. In other words, who is responsible or takes responsibility.	**Rules & best practices** Business models should be developed to understand what each stakeholder can offer and would like to carry on. Stakeholders can achieve much more collaborative efforts (e.g. linking publications and data). Best practices can be taken from the publishers and libraries. Boundary conditions should be defined.	EU, National governments, Industry (e.g. publishers)	ORG: business models TEC: brokerage systems
Financial responsibility of digital preservation is not defined. For example, many data managers expect funding to become an issue in the next few	**Awareness** National governments should realise this is an issue for data managers on national and European level. **Funding**	National governments, EU	ORG: business models TEC: infrastructural components

years.	A transition from project funding to structural funding for digital preservation should be made.		FIN: structural funding
Funders do not fully use their potential in shaping preservation activities. They have a prominent position in the research value chain and thereby the power to promote and shape preservation activities.	**Rules & best practices** Funders should provide procedures to researchers on how preservation is to be given shape. **Funding** Funders should incorporate preservation activities as part of project proposals. National governments and the EU should take on their responsibility and budget for preservation activities.	Funders, National governments, EU	ORG: business models FIN: earmarking, preservation requirements.

6. Methodology

In this chapter, the methodology by which this research was set up is explained. The various ways in which data were gathered are outlined. Also, the validity of the research will be commented upon, as well as some limitations of the research.

6.1.   Research methods

The method by which the research was conducted has four components:

- Desk research

- Surveys

- Case studies

- Interviews

Desk research provided the overviews of the research structures on European and national levels, as well with some basic feeling with existing ideas and concerns relating to digital preservation in various countries. Many relevant organizations even have policy documents available on their websites, expressing their views on preservation and displaying their procedures. However, policy documents and fact sheets only tell part of the story. Moreover, not all countries have these documents readily available. This is why surveys, interviews and case studies have been conducted as well.

Separate surveys were developed for each stakeholder, sharing a common core of questions, because it was deemed important 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.

Within the research community, three case studies have been performed, in order to provide some in-depth knowledge on the different views to digital preservation within the various research disciplines. In this way, the research disciplines of High Energy Physics, Earth Observation and Social Sciences & Humanities have been explored in great detail[53].

Finally, interviews have been conducted with key informants within the funding community of several European countries, covering most parts of Europe. Since the large-scale surveys did not

---

[53] PARSE.Insight D3.3 Case Studies report

result in many responses from funders, the interviews provided us with an opportunity to investigate the views and concerns of the funding society.

6.2.　Validity of results

By deploying various research methods quite an amount of responses from the various stakeholders from all over Europe has been elicited, which can give an indication of the validity of the research results. Especially the large scale survey among researchers, data managers and publishers has provided large enough a sample, to be able to draw conclusions from the research. From the research survey, 1,389 responses were elicited, while the survey among data managers gave 262 responses and the publishing survey resulted in 178 responses. All parts from Europe were represented in these surveys. The interviews and case studies that have been conducted, combined with desk research have given the additional depth to the analysis.

As has been mentioned before, there were not many responses among funders to the survey that was sent to them. This was compensated by analyzing policy documents and conducting interviews with key players within the funding community from various European countries. However, it may be clear that the insights that were derived from these interviews and policy documents cannot be attributed the same level of validity as to the other three stakeholder groups.

7. References

7.1.   Related documents PARSE.Insight

- PARSE.Insight D2.1 Draft Roadmap
- PARSE.Insight D3.2 Inventory of Communities
- PARSE.Insight D3.3 Case Studies Report
- PARSE.Insight D3.4 Survey Report
- PARSE.Insight D3.5 Interim Insight Report

7.2.   Interviews with research funders

- Dutch Ministry of Education (OC&W), 08.02.2010
- Deutsche Forschungsgemeinschaft (DFG), 31.03.2010
- Czech Republic Research Academy (ASCR), 07.04.2010
- Spanish Research Council (CSIC), 26.04.2010

7.3.   List of documents on funder policies

- http://cordis.europa.eu/erawatch/index.cfm?fuseaction=ri.content&topicID=4&countryCode=EU

- http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf

- http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

- http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

- http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_Recommendations_OpenAccess_200805.pdf

- http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/ESRC_Open_Access_Repository_Policy_tcm6-31316.pdf

- http://www.rin.ac.uk/system/files/attachments/Research-funders-outputs-report.pdf

- ftp://ftp.cordis.europa.eu/pub/fp7/docs/open-access-pilot_en.pdf

- http://www.bbsrc.ac.uk/web/FILES/Policies/data_sharing_policy.pdf

- http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf

- http://www.oecd.org/dataoecd/9/61/38500813.pdf

- http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/goodresearchconductcode.pdf

- http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

- http://www.sherpa.ac.uk/juliet/index.php?sortby=name/

- http://cordis.europa.eu/erawatch/index.cfm?fuseaction=ri.home