First insights into digital preservation of research output in Europe
Table of Contents

1 INTRODUCTION 3
   ABOUT PARSE.Insight 3
   WHAT DO WE MEAN WITH PRESERVATION OF DATA? 3
   WHY THIS DOCUMENT? 3
   WHAT DID WE DO? 3

2 RESEARCHERS 4
   GENERAL SURVEY RESULTS 4
   CASE STUDY 1: HIGH ENERGY PHYSICS 7
   CASE STUDY 2: EARTH OBSERVATION 10
   CASE STUDY 3: SOCIAL SCIENCES & HUMANITIES 12
      Psycholinguistics 12
      Book Studies 13

3 DATA MANAGERS 16

4 PUBLISHERS 18

5 IMPLICATIONS FOR THE ROADMAP 19
   STIMULATING THE DEBATE 19
   PARTICIPATE IN THIS DISCUSSION! 19

APPENDIX: METHOD 20
   WORKFLOW & DISTRIBUTION 20
   STATISTICAL VALUE 20

Acknowledgements

Gaining insight in Europe’s stakeholders in research is a difficult task. With help of many people and organisations PARSE.Insight was able to gather a lot of information by offering the opportunity to use their distribution channels. In special, we would like to thank the following organisations:

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

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, access and re-use of digital information in science, from primary data through analysis to the final publications resulting from the research.

The aim of the PARSE.Insight project is to develop a roadmap and recommendations for guiding the European Commission’s strategy in developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe.

PARSE.Insight is closely linked to the Alliance for Permanent Access to the Records of Science in Europe¹.

What is digital preservation of research data?

In this report research data encompasses the whole cycle of research-related, from raw (machine) data to publications.

Digital preservation of research data here means the careful storage of all research output in such a way that it remains accessible, usable and understandable over the long term.

Why this document?

This document is an interim version of the Insight Report, a deliverable of the PARSE.Insight project which is due early 2010. It presents some of the highlights of the surveys and case studies done by PARSE.Insight project members over the past months. The results provide evidence for the ideas presented in the PARSE.Insight Roadmap and lay bare existing barriers in the preservation preserving and sharing of research data.

This report is intended to stimulate discussion on the interim conclusions and help the PARSE.Insight project refine them. At the end there are a series of questions, arising from the surveys to date, that the project team hopes will act as a focus for debate.”

What did we do?

Our approach is twofold: firstly, we aim to gain insight into the needs, practices, and requirements of major stakeholders in research—researchers, publishers, data managers and funders. Secondly, we performed three case studies into three specific scientific communities: High Energy Physics (HEP), Earth Observation (EO), and Social Sciences & Humanities (SSH).

This document presents the highlights of the survey results from three stakeholder surveys: research, data management, and publishing. The section on research also contains the first insights into the communities of the three case studies. Finally, by way of conclusion, some implications for the Roadmap are given.

In addition to the afore-mentioned stakeholders, we also recognised funders as a major stakeholder. A survey for funders was also developed, but because of the much smaller size of this group, different and more targeted channels (e.g. workshops, interviews) are also being employed to reach them and ascertain their views. Funders results are not present in this interim report, but will be taken into account in the final version of the Insight Report.

A word on our method can be found in the appendix at the end of this document, where we discuss the confidence which we have in our conclusions, recognising that, despite the large number of responses, we must still be quite careful.

¹ http://www.alliancepermanentaccess.eu
2 Researchers

General survey results

From October 2008 through March 2009, a large survey on digital preservation in research was conducted. In total, 1,397 people responded and filled in the survey completely. The following statements highlight the outcome of the survey.

By and large there is a consensus among researchers that data from publicly funded research should be shared and preserved. About 87% of the respondents believe that if research is publicly funded, the results should become public property and therefore properly preserved.

The fact that preservation may stimulate the advancement of science and that it allows for re-analysis of existing data as well as future validation are either important or very important reasons for preserving the data according to the respondents. Researchers seem to care rather less about the potential economic value of the research data, although only 19% thinks that this is not important, at least in the long run (see figure 1).

There is also high degree of awareness on the major threats to long-term preservation of digital research data. Between 56% and 80% of the responses indicate that all threats we formulated are recognised as either “Important” or “Very Important.” (see figure 2). Access and use restrictions is regarded as the least important threat to preservation, while the lack of sustainable hardware, software or support is recognised as the most important threat to preservation.

59% of respondents also believe that some kind of international infrastructure for data preservation and access should be built to help guard against the above-mentioned threats. Yet, when asked what such an infrastructure should look like, the answers were anything but uniform. Many admitted not to know, but the idea of building some sort of international central archive or storage location with

PARSE.Insight: first insights into preservation of research output in Europe | www.parse-insight.eu | Project: FP7-2007-223758
open access was espoused several times. In figure 3, you’ll find a tag cloud of the free text answers to this question.

However, countering the threats to digital preservation encompasses more than dealing with the technical aspects. A large part of the problem is also psychological. And there is ample proof for that in the survey as well. When asked where researchers store their research data (see figure 4), the most important ‘locations’, in order of the number of responses, are: personal computer at work (81%), portable storage carrier (66%), organisational server (59%), and computer at home (51%).

Only 20% submit their data to a digital archive. This is a telling figure which gets more meaning in the context of the questions we asked about sharing data. As it turns out the 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. Only 11% of the respondents make their data available for researchers within their research discipline. The majority of respondents do make their data available to researchers within their research collaborations and groups, but even 58% may not be a very high figure here as well.

As we noticed earlier the sharing of these data does not seem to take place through established digital archives, even when they are specific to the discipline. In other words, researchers seem to want some sort of control over their data and they see many problems surrounding the sharing of data. The major problems researchers foresee are legal issues (41%), misuse of data (41%), and incompatible data types (33%). So, there still is a lot of distrust in the capability of digital archives to properly handle research data (see figure 5).
Besides the technical and human elements digital preservation involves costs. Respondents had to provide their views on who is responsible for the preservation of research data and publications. A majority of the respondents believe that their national government should pay the bill for the preservation of research data (61%) and publications (57%). The top three for research data is completed by the researchers’ organisations (41%) and the European Union (36%).

![Fig 5: Perceived risks in sharing data](chart.png)
**Case study 1: High Energy Physics**

Particle Physics investigates the fundamental constituents of matter and radiation and the interactions between them. It is often referred to as high energy physics (HEP) because many elementary particles do not occur under normal circumstances on earth but can be created during energetic collisions of other particles (as is done in particle accelerators) through Einstein's principle of equivalence of energy and mass.

HEP presents a particularly complex case in the debate on digital preservation due to the huge amount and outstanding complexity of the data generated by present-day accelerator facilities; it may thus be regarded as a “worst case scenario” for digital preservation.

The online survey for the High Energy Physics (HEP) case study, conducted by CERN, was launched in October 2008 and ran for three months. It was advertised through the mailing lists of large experimental collaborations and through a link on the SPIRES web page, a database of particle physics literature. A large fraction of the HEP community, which is estimated to include about 20,000 active physicists, was thus reached. The survey yielded 1197 responses: 883 by experimental physicists and 314 by theoretical physicists.

The distribution by country of the respondents reflects approximately that of the active physicists in the field: 41% come from countries of the European Union, 23% from the United States, and 23% from the rest of the world, while 13% spend most of their working time at CERN which was presented as an additional “country” in this study.

The distribution of respondents in their career path is rather flat, with 22.9% Ph.D. students, 23.3% post-doctoral fellows, 28.0% researchers with permanent positions, and 25.7% professors. Experimental physicists were also asked to specify the large projects in which they are or were involved. The respondents are therefore representative of the entire HEP community: geographically, by career status and age group, and by covering a large spectrum of HEP experiments.

Respondents were asked how important they regard the issue of digital preservation. It is remarkable that about 69% of the respondents perceive preservation as “very important” or even “crucial”. The distributions are the same if the respondents are divided in “age groups” (less or more than 5 years of experience in HEP). See figure 6.

![Fig. 6: importance of data preservation](image)

**In your opinion, how important is the issue of data preservation?**

- **Irrelevant**: 0.4% (0.9%)
- **Moderately important**: 3.3% (11.7%)
- **Important**: 15.2% (29.7%)
- **Very important**: 41.7% (40.5%)
- **Crucial**: 24.3% (49.3%)

**Theorists** vs **Experimentalists**
Respondents were asked to indicate the level of importance for four formulated reasons for preservation. A level of “very important” or “crucial” was indicated for:

- re-analysis of preserved data to test future theories: 74%;
- combination of preserved data with future data: 63%;
- use of preserved data for future independent checks of results: 60%;
- use of preserved data for teaching or outreach: 27%.

Each of the first three use cases is perceived to be more important by theorists than by experimentalists. Especially the difference for the use of preserved data that allows independent checks of results is striking. The first use case is deemed “very important” or “crucial” by 81% of the theorists but only 54% of experimentalists. The relatively low interest in re-using preserved data for teaching or outreach is somewhat surprising.

The scientific case for data preservation and re-use is embodied by the fact that 54% of the theorists and 44% of the experimentalists think that access to data from past experiments could have improved their scientific results. At the same time, 46% of respondents think that important HEP data have been lost in the past.

What should be preserved and where? Figure 7 summarizes the answers to the question regarding the level of abstraction at which data should be preserved. Moving downwards from the top, the complexity of the data is increasing while the level of abstraction decreases. The distribution is remarkably flat. In the light of the high complexity of the problem of re-using older data, it is surprising that as many as 66% of the respondents would preserve event-by-event information and 45% even raw data.

The overwhelming majority of respondents (68%) would like data to be preserved at a “neutral” platform such the equivalent of ADS, arXiv, CDS or SPIRES, adapted to the preservation of and access to data. This preference is even stronger in the case of theorists (86%). Storage at a site connected to the experiment or laboratory where the data have been produced is the second choice, with a stronger preference from the side of the experimentalists.

**Fig 7: what should be preserved?**

<table>
<thead>
<tr>
<th>Information from published tables and figures, e.g. numerical information in electronic form</th>
<th>Theorists</th>
<th>Experimentalists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backup information which did not fit in your publication (e.g., additional numerical information, figures and tables, comparison of data and simulation)</td>
<td>62.3%</td>
<td>74.9%</td>
</tr>
<tr>
<td>Multi-dimensional distributions (differential cross sections in many variables, likelihood distributions) which cannot be fully detailed in the publication</td>
<td>60.4%</td>
<td>60.4%</td>
</tr>
<tr>
<td>Event-by-event higher level objects (e.g. Four-vectors), together with appropriate information allowing some re-analysis of the data</td>
<td>69.7%</td>
<td>69.7%</td>
</tr>
<tr>
<td>Raw data (together with appropriate access and interpretation “tools”) allowing complete re-analysis of the data</td>
<td>45.5%</td>
<td>45.5%</td>
</tr>
</tbody>
</table>

This might reflect a certain level of concern of theorists regarding the stability of data when hosted by the producers themselves. Hardly any of the respondents prefer platforms managed by journal publishers.
Almost all respondents (94%) think that the additional effort needed for preparing data for preservation in a re-usable form is substantial (more than 1% of the overall effort invested in producing and analyzing the data) whereas 43% think that the supplementary effort is more than 10%. These percentages are independent of the experimental facilities where respondents work. This finding is not surprising and confirms the important financial and person-power implications of a large-scale preservation program in HEP. This situation is made more complex by the timing that researchers indicate as crucial for the success of a data preservation program: 41% of the respondents think that the effort towards data preservation has to be deployed concurrently to data taking, while 28% think that preservation should be prepared even before the actual data taking. It is remarkable that participants to the LHC experiments strongly indicate that data preservation should be addressed before data taking, which means now!

Against this background of positive attitude towards data preservation and acknowledgement of the large challenge it presents, it is sobering to find that only 16% of respondents think that their experiment/collaboration/organization is able to produce the required financial effort and person-power to tackle this issue, 7% think that this is not the case; the large majority of respondents, 77%, just does not know, showing a clear gap between awareness and action.

The survey also aimed to quantify issues clouding the potential of third-party re-use of preserved data. Two potential areas of concern emerge: the sharing of credit and responsibilities between the producers and re-users of the data, and the validity of the results derived from the analysis of preserved data. A relatively low number of respondents (24%) are “very concerned” or “gravely concerned” that preserved data could be used without giving proper credit to the authors.

On the other hand, 45% of the respondents are “very concerned” or “gravely concerned” that data re-use may in general lead to an inflation of incorrect results. Experimentalists are by far more concerned (51%) than theorists (29%). At the same time, respondents are also concerned by producing themselves uncorrected results by misinterpreting preserved data. As many as 53% are “very concerned” or “gravely concerned” about this possibility and, of these, experimentalists are again more concerned than theorists.
Case Study 2: Earth Observation

Used to monitor the status of natural and built environments, the Earth Observation (EO) gathers information about Earth’s physical, chemical and biological systems. Earth observations provide objective coverage across both space and time and are fundamental input to Earth Science’s spheres (atmosphere, hydrosphere, geosphere, biosphere, environmental science). Examples of their application include:

- preventing geohazards (e.g. floods, earthquakes, volcanoes)
- measuring land-use (e.g. deforestation, urban expansion)
- tracking biodiversity (e.g. oceanography, carbon cycle)
- monitoring climate change
- forecasting weather

Over the last decade, data acquired from space have become powerful scientific tools to enable improved understanding and management of the Earth and its resources. Thanks to improved means for accessing and use EO data, the Earth Science community is increasingly requiring access to historical datasets, time-series of data spanning 20 years and more, thus calling for the availability of any EO dataset, without time constrains.

In Europe, ESA is active since decades on the preservation of its EO archives, and the involvement as representative of the EO case study in PARSE.Insight is in-line with the established strategy and actions taken to ensure accessibility to EO data in the long-term. Having an insight on the topic of long-term availability of environmental data in the Earth Science community contributes to fill the gap between EO data generation, archiving and maintenance processes and their exploitation, understanding the data users’ standpoint and requirements on historical environmental data and non space data.

ESA issued a public survey with the primary target to know more about current and envisaged exploitation of historical environmental data streams by the scientific community.

The attendance show a very high interest on the topic, and of all respondents 65% left their names to be contacted again later. Respondents work worldwide, with a geographic spread far beyond ESA Member States. Participants are mainly scientists and researchers, working in research institutes, academia, national authorities, and international (non-)governmental organisations. Representatives from industry and the commercial sector, students and the general public, who is being showing an increased interest on EO applications to day life and environment protection, also participated actively.

13% of respondents stated to be new to the topic of data preservation, while the majority agreed that the scientific research, climate change study, and disaster evaluation are the fields that can profit at most from the availability of historical environmental data.

The vast majority declared to constantly need access historical environmental data and 84% stated to require at least access once, detailing their own experiences. About 25% of the respondents reported an interesting set of occasions of data losses.

Regarding the major threats to environmental data preservation, the lack of sustainable hardware, software or support of computer environment and the eventual loss of the data custodian are considered very important ones. Users’ inability to understand or use the data and the uncertainty of data’s origin and authenticity are considered anyway important threats by the majority (see figure 8).
The survey also investigated the opportunity of infrastructures for environmental data preservation. More than 90% of participants agreed that an infrastructure for data preservation would impact their daily work and further motivated their choices. Respondents demonstrated high awareness on the fact that making available and maintaining historical environmental data series means to reliably store them, to enable easy access both to data and knowledge, including applied processing, metadata, auxiliary data, and traceability and to make their analysis and re-processing possible.

Summing up on the many answers received, it is clear that the scientific community want to access historical environmental data and historical time series of earth observations. Secondly, the community wants to make its experiences on historical data exploitation known to ESA, and participate by reporting examples and suggestions. Lastly, the EO data users are aware and updated on the current infrastructures’ constraints and require timely solutions.
Case Study 3: Social Sciences & Humanities

To get a good coverage of the research disciplines spectrum, our third case study focuses on the social sciences & humanities. Within this case study two disciplines are covered: psycholinguistics, done by the George-August-Universität Göttingen and the Max Planck Gesellschaft (MPG), and book studies, done by the Deutsche Nationalbibliothek (DNB).

Psycholinguistics

Psycholinguistics is a field of study which combines the disciplines of psychology and linguistics. It is concerned with the relationship between the human mind and language as it examines the processes that occur in the brain while producing and perceiving both written and spoken discourse. Psycholinguistics as a separate branch of study emerged in the late 1950s and 1960s as a result of the Chomskyan revolution.

Psycholinguists use their own data they have generated during experiments or observations or use data that is stored in language archives. A lot of data archives within and outside Europe house materials that are intended to document and describe human languages, such as wordlists, lexicons, annotated signals, interlinear texts, paradigms, field notes, and linguistic descriptions.

The George-August-Universität Göttingen and the Max Planck Gesellschaft (MPG) ran a large survey among researchers in psycholinguistics and related stakeholders such as preservation officers and archivists. Three distribution channels were used: direct mailings by invitation (3,100 people), community-specific mailing lists, and international psycholinguistic associations. The survey ran from 21 January 2009 through 9 March 2009 and elicited 205 completed responses. In addition, various interviews have already been conducted as well.

20 % of the research respondents stated that digital data already has become unusable within their organizational unit which shows that data actually is at risk. In most cases data stored on an optical or magnetic disk has become unusable, mostly due to technical changes. Some respondents did not loose data but had problems to back up data that has been stored in more than one location/on several media types. Hence, not only technical changes endanger storage, preservation and re-use of data, but also organisational problems.

Most often data has been lost because the software for interpreting it is no longer available (75 %). Of the respondents 60 % also noted data loss due to lack of contextual information, like for example manuals or notes. At least 11 % of the respondents know that data has been lost, but cannot state the reasons. This ignorance is a risk as well: The less researchers know about their data, the more difficult it is to keep data alive.

Regarding the threats to preservation, lack of sustainable hardware, software or support of computer environment (42 %) and the current custodian of the data may cease to exist at some point in the future (40 %) were considered very important threats. Marked as important threats were users may be unable to understand or use the data (44 %), that evidence of data may be lost because the origin and authenticity of the data may be uncertain (42%) or that the ability to identify the location of data gets lost (47 %). The threat that those trusted to look after the digital holdings may fail to do so is considered less important (36 %).

About three-quarters of the respondents (75 %) think that an international infrastructure for digital data preservation and access should be built to provide protection against these threats. Apart from an infrastructure, more knowledge/expertise is needed to guarantee that valuable digital research data is preserved for access and use in the future (92 %).

In general, some respondents already have ideas of what such an e-infrastructure should look like: It should be persistent, easy to use and less bureaucratic. Fur-
thermore, the e-infrastructure should be based on standards (e.g. standard formats) and help guard against legal problems. The interests of organisations or projects that provide the data should also be preserved. Furthermore, it was stated that more interoperation between repositories is needed. According to one respondent, “the answer lies in a network rather than in one single infrastructure.”

To get a better picture of what an infrastructure could support, some basic elements (tools and features) were suggested. Out of that list, the research respondents defined research tools for language resources and storage as the most important elements. Other elements for the long-term preservation of scientific data, such as tools for metadata generation and persistent identifier systems, have been judged less important which is remarkable as these elements are very important in creating a sustainable infrastructure.

On average, all respondents consider training, more knowledge, more financial and operational resources and more digital repositories/archives equally important (training: 69 %; more knowledge: 73 %, more financial and operational resources: 72 %, digital repositories/archives: 69 %). Only a few researchers (2 %) do not think that this will help guard against the threats mentioned.

**Book Studies**

Book History, Printing Science and Technology, Bibliography, Publishing and Economy, History of Reading – the variety of names refer to the diversity of approaches in the discipline dedicated to the study of the book. Book Studies developed in the 19th century as an ancillary science from the study of literature. Today the discipline describes itself as historical Media-Science, which takes part on the exploration of a cultural history of knowledge. Traditional objects of research as the invention of printing by Johannes Gutenberg are included as well as actual tendencies on the book market.

The Book Studies survey was open from March though May 2009. Altogether, 999 researchers from Book Studies, mostly from the European Union, were invited to fill in the survey. About 124 people (12%) completed the survey. Researchers on the book from 27 countries participated. The United Kingdom, Germany, the Netherlands and the United States of America are the best-represented countries in this case study survey. Together, researchers from these countries constitute 66% of all respondents (UK 15%, Germany 23%, Netherlands 7%, and USA 21%). These countries all have a strong tradition in Book Studies and have numerous research centres and Book Studies related associations.

54% of the respondents mentioned that their institute is currently involved in activities aimed at preserving digital resources or plan such activities. 84 respondents described the activities planned. 23 of them stated “digitisation of library materials”, “scanning books” or other digitisation activities, which do not correspond with our understanding of digital preservation. Only 38 people described “real” preservation activities like “selection of a digital preservation tool” or “establishing a trusted digital repository”. This shows that the concept of digital preservation is still rather unclear to a significant number of researchers.

There is a large consensus among the researchers that the preservation of digital resources is important. Regarding the reasons, almost three-quarters of the respondents share the conviction that results from publicly funded research should become public and therefore be properly preserved. The same number of respondents believes that preserved data will stimulate the advancement of science, since new research builds on existing knowledge. Respondents seem to care less about the potential economic value of digital resources. Only 45% considered this an important or very important reason for digital preservation, and 12% considered it unimportant.
When asked what kind of materials should be preserved, 93% agreed that the preservation of scientific e-journals is very important or important, and 91% thought so for digitized books. Other resources which were held important or very important by the participants were online publications like websites and online theses (76%), trade journals of the book branch (75%), statistical data (74%) and grey literature like preprints or conference proceedings (70%). Both the long-term preservation of unpublished data and of information and communication platforms was considered important or very important by roughly 50%. Preservation of e-mail was considered as less important by the majority of respondents, which is an astonishing result for a discipline that traditionally depends on sources of communication, e.g. between authors and publishers.

There is also high degree of awareness on the major threats to long term preservation of digital research data among Book studies researchers. Figure 9 shows that of the threats listed, most are regarded as ‘very important’ or ‘important’.

75% of the respondents in Book Studies research believe that an international e-infrastructure for digital preservation and access should be built to help guard against some of the above mentioned threats. When asked what such an e-infrastructure should look like, the whole spectrum from a centralized EU-agency to a completely decentralized e-infrastructure was pictured.

In addition to these specific threats, respondents also consider more general threats as the lack of structural funding (82%), the lack of technical support (67%) and the lack of continuity of organisations (58%) as important.

In order to find out what is already in place in terms of e-infrastructure elements in the Book Studies, we asked the participants how they locate and access resources related to their research. Most of the respondents use both “general search engines” like Google and Yahoo (89%) and “institutional database and search facilities” like incunabula databases (88%), which reveals that the researchers of the book make reasonable use of the great number of specialised offerings for book researchers, but are also open for the “quick and dirty” search engines.

However, most of the respondents (53%) stated that they are not satisfied with the existing services for researching digitized sources. Two different reasons can be found in the according answers: One group of people complains that not enough material has been digitized already. The other group expresses the desire that the multitude of specialised research offerings should be unified within a central database or that it should at least more frequently be possible to search across the individual databases.

When inquired what is there in terms of e-infrastructure to support the creation, publishing and preservation of sustainable research papers. It turned out that 43% of the respondents did not know if any of such elements as preservation policies, repositories, platform for data exchange were available to them. Only 22% could say that they were satisfied with the existing elements and only 18% were sure they needed more or different services. The majority of 59% did not know if the existing tools and elements fit their needs.

**Fig. 9: threats to preservation in the field of Book Studies**

<table>
<thead>
<tr>
<th>Threat</th>
<th>(Very) Important</th>
<th>(Less) Important</th>
<th>Don't Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>The ones we trust to look after the digital holdings may let us down</td>
<td>65%</td>
<td>25%</td>
<td>7%</td>
</tr>
<tr>
<td>The current custodian of the data, whether an organisation or project, may cease to exist at some point in time</td>
<td>86%</td>
<td>11%</td>
<td>3%</td>
</tr>
<tr>
<td>Loss of ability to identify the location of data</td>
<td>71%</td>
<td>27%</td>
<td>2%</td>
</tr>
<tr>
<td>Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future</td>
<td>64%</td>
<td>31%</td>
<td>5%</td>
</tr>
<tr>
<td>Evidence may be lost because the origin and authenticity of the data may be uncertain</td>
<td>40%</td>
<td>59%</td>
<td>1%</td>
</tr>
<tr>
<td>Lack of sustainable hardware, software or support of computer environment may make the information unreadable</td>
<td>40%</td>
<td>59%</td>
<td>1%</td>
</tr>
<tr>
<td>Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved</td>
<td>46%</td>
<td>49%</td>
<td>5%</td>
</tr>
</tbody>
</table>

PARSE.Insight: first insights into preservation of research output in Europe | www.parse-insight.eu | Project: FP7-2007-223758
Only few researchers seem to have experiences with professional preservation facilities for their research materials. When asked where they store their digital documents and data, the most important storage location is the computer at work, where 77% store their documents and data, followed by portable storage carriers (75%) and the computer at home (70%). Only 10% submit their data and documents to a digital archive and 11% submit them to an external web service. However, these figures show that there is a certain level of redundancy – at least some people store their data in multiple places. The little use of digital archives could be a result of the low level of information about long-term preservation in general but could also simply indicate that there are not enough digital archives available to researchers of the Book.
3 Data Managers

Of data managers, 273 responded to the survey and 146 respondents completed it.

Like researchers, data managers agree that public funding is an important reason to preserve digital research data. 98% of the respondents believe it to be either a very important or an important reason for preservation. The data managers also think that it is important that preservation may stimulate the advancement of science and that it allows for re-analysis of existing data. Also, like researchers, data managers don’t regard economic value as an important driver for preservation.

It is interesting to see whether data managers have different opinions on the threats to digital preservation than researchers. Of the seven threats formulated, 75% to 86% of the respondents thought that five out of the seven threats are either important or very important. The threat that the current custodian of the data may cease to exist is regarded as the most important threat by respondents, closely followed by the lack of sustainable hardware, software, or support and the threat that users may be unable to understand or use the data.

Interestingly, 66% of the respondents in the data managers survey believe that the threat that the ones we trust to look after the digital holdings may let us down is either an important or very important threat (see figure 10). Does this point to a lack of faith in data managers’ capability to provide guarantees for a safe and sustainable curation of data? Or, do data managers see their role differently from the one we ascribe to them. Data managers are least worried with the threat that access and use restrictions are not respected in the future.

Considering the high awareness of the threats to digital preservation, do data managers believe that an e-science infrastructure may help counter these threats? 60% of the respondents believe it does and 23% doesn’t know. Yet, when asked what such infrastructure should look like, data managers – perhaps not surpris-
ingly – seem less likely to answer don’t know than the researchers. What was striking was the emphasis on an international solution, EU or even global.

We also asked respondents whether more is needed than can be facilitated by an e-science infrastructure. More resources, more expertise, and more training are most often marked as the things needed most next to an e-science infrastructure (see figure 11).
4 Publishers

The number of publishers who responded to the survey was 186 of which 137 completed the survey.

The majority of the large publishers (89%) have a policy for the preservation of digital publications in place, as opposed to 56% of the small publishers. Looking closer at what these policies entail or what kind of preservation strategies these publishers have implemented, we notice that 57% of the large publishers and 23% of the small publishers outsource the preservation of their digital publications to a third party service (KB’s e-Depot, Portico, etc.). Many publishers apply even more than one strategy; the two alternative strategies most often marked were “normalisation” (25% for small publishers and 36% for large publisher) and “migration” (13% for small publishers and 19% for large publishers). Yet, 7% of the large publishers admit not to have a preservation policy in place. For small publishers the percentage is significantly higher, namely 28%.

We also asked what kind of materials the publishers think should be preserved. Research articles (95%; 93%) and books (80%; 89%) are—perhaps not surprisingly—the most often marked choices. A substantive amount of publishers, though, also believe that illustrative material (59%; 61%) and even data sets / auxiliary material (55%; 57%) should be preserved.

There is more or less a consensus among publishers that underlying digital data is important to the publications and will become even more important in the (near) future. A majority of large publishers (71%) already allow authors to submit underlying digital research data, together with their manuscripts, to the journal, which comes available for free upon publication of the article. 58% of the small publishers also allow for this. However, asked if they have preservation arrangements in place for these research data, a majority of the publishers (68%; 71%) state that such arrangements do not (yet) exist.

Compared to researchers, publishers (74%; 67%) are more convinced that an infrastructure will help counter the threats to digital preservation. In contrast to the fact that many publishers do take care of digital preservation of the publications, most believe that other parties are better equipped to handle the preservation of research data. The publishers can at best play a facilitative role. This is in line with the high percentage of publishers who do NOT have a preservation policy in place for research data submitted with journal articles. Publishers believe that authors (49%; 37%) and the author’s institute (41%; 37%) can best take the role of preserving the research results. They also believe it should be paid for by government funding (54%; 44%).

On future business models in general the publishers believe the future will be dominated by a hybrid model, in which subscription-based and open access journals will both exist. Yet large publishers (78%) more firmly believe in this scenario than small publishers (58%). In spite of the changes publishers expect, they do seem to be confident that the journal will keep its important role in research as the “minutes of science,” as one publisher put it.

2 For the publishers results we wanted to make a distinction between open access publishers and STM publishers. Yet any crude distinction, like the one between the DOAJ and the International STM Association, would violate reality. Open access journals are not a sole property of the open access publishers organised in DOAJ. Most traditional STM publishers have catalogues which include open access journals. A more useful distinction would be between large and (very) small publishers. The STM market is one of large publishers, whereas open access publishers, usually only publish one or a few titles. So, in the results we made the distinction between publishers who publish less than 50 journals and those who publish more. The distinction is based on one of the survey’s questions.

3 From here on, when placed in brackets, the first percentage means small publishers, the second means large publishers.

4 There is a striking difference between the STM publishers (47%) and the DOAJ publishers (85%) here.
5 Implications for the Roadmap

Sharing and preserving research data seems desirable, but many hurdles prevent it from happening on a large scale. Reverse salients — critical unsolved problems — may be technical, but are also frequently social or organizational in nature [ref Understanding Infrastructure: Dynamics, Tensions and Design. The reasons for and threats to preservation mentioned in previous sections seem real to all stakeholders although the importance varies. In general, there is consensus about the need for an e-infrastructure for research to help guard against those threats. This is supporting evidence for what is already outlined by the draft Roadmap of PARSE.Insight.

However, what the infrastructure should look like is still vague and uncertain for many of the stakeholders. Suggestions have been done that vary from technical components up to a better legal framework. More clear seems the responsibility of who should do what and who should pay for setting up an infrastructure. By far, the EC and national governments seem favourite candidates to make it happen.

Stimulating the debate

Although surveys and case studies created valuable insight into how researchers work and what stakeholders think, there are still many questions are open for debate. We would like to hear your opinion on the following topics:

Regarding distrust in external organisations dealing with preservation:
3. Is this distrust justified?
4. How can it be countered?

Regarding the gap in desire and practice of sharing data:
5. How can openness of data be stimulated?
6. What would the role of funders be?

Regarding the case studies:
7. Do you recognise the situation?
8. Is this true in the areas you know about?

Regarding publishing:
9. Do you agree with their opinion that other parties are better suited for taking care of digital preservation?
10. Do you believe that the future business model of publishers will become a hybrid of open access / subscription-based?
11. That underlying research data should be linked with the publication?

Roadmap implications:
12. Do you agree with the broad thrust of the Roadmap?
13. What do you disagree with?
14. What is missing?
15. What details would you like to be emphasised?

Participate in this discussion!

You can take part in these discussions in two ways:
• Via our PARSE.Insight forum. Go to www.parse-insight.eu/forum
• Via our workshops at several occasions in 2009 and 2010

http://www.si.umich.edu/~pne/PDF/ui.pdf

PARSE.Insight: first insights into preservation of research output in Europe | www.parse-insight.eu | Project: FP7-2007-223758
Appendix: method

Workflow & distribution

In short, our workflow consists of desk research, surveys, interviews and three case studies. This interim report presents the early results of the surveys we sent to researchers, data managers and publishers. In addition, the highlights of the results of the three case studies are presented. The next stage is to gain insight into how research is funded by conducting interviews with people at financial organisations that finance research activities. Furthermore, in-depth discussions at workshops and in forums should help us gain a more refined perspective on the preservation practices, needs & requirements within European research communities.

A survey denotes a method of gathering information from a sample of individuals. The sample represents the total population being studied. There are different methods for survey data collection but in this document survey means an online questionnaire that is distributed by e-mail or any other means of online distribution, e.g. mailing lists.

The PARSE.Insight surveys are aimed at the (European) research communities. This encompasses all member states of the European Union and all disciplines. We consider both elements—country and discipline—for the survey’s representative sample. If we noticed a lack of responses from certain disciplines or countries when we were collecting data, we tried to locate new distribution channels that specifically targeted group(s) of people whose responses were lacking. The surveys were designed, distributed and processed by using the web application Survey Monkey.

The research survey generated the largest response. With the Elsevier mailing list of approximately 35,000 journals editors, the research survey elicited 1,379 responses. The survey that targeted data management was sent to research libraries, data centres, and archives via several mailings, among them the LIBER mailing list. 273 people responded to that survey and 146 respondents completed the survey. The publishers’ survey was distributed to a selective number of individuals from the members list of the International Association of STM Publishers and to the publishers of the Directory of Open Access Journals (DOAJ). When combining all publishing distribution channels, the number of publishers who responded is 186 of which 137 completed the survey.

Statistical value

Even with this large number of responses, we do not know the size of the whole population. Therefore, it is not possible to be certain about how representative the results are. We can be sure that the respondents are those willing to fill in surveys but those too busy or otherwise unwilling to complete the surveys will be under-represented. In other words, we may only hear the loudest voices. Another concern is that we have had to provide some structure to the responses by means of multiple choice questions. There is clearly a danger that by doing so we could have pre-determined the answers to some extent.

Although it is not possible to eliminate these concerns entirely, we have several methods to cope with those concerns:

- We can compare our results with those from the case studies surveys, in particular the High Energy Physics survey as we have a better idea of the total size of the population.
- We can compare different groupings of disciplines to see to what extent there are agreements.
- We have provided an odd number of choices (5) to tempt people to choose the middle option – the fact that they did not indicate a well considered choice.
- We have provided free text options to allow respondents to express their own ideas.

Thus, while we cannot claim that our results are absolutely representative, nevertheless we believe that our conclusions are quite reliable.