What researchers want

About this publication

What researchers want
A literature study of researchers' requirements with respect to storage and access to research data

> *"…The key requirement from most researchers' perspectives is for services which are there when they need them, but do not interfere with the creative work at the heart of the research process." (Henty, 2008)*

SURFfoundation
PO Box 2290
NL-3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60

info@surf.nl
www.surf.nl

Author
- Martin Feijen

Editors
- Paul Gretton - Gretton & Willems Translations
- Keith Russell - SURFfoundation

SURF is the collaborative organisation for higher education institutions and research institutes aimed at breakthrough innovations in ICT (www.surf.nl/en)
This publication is online available through www.surffoundation.nl/en/publications

February 2011

Contents

Management Summary **4**

Managementsamenvatting **5**

1 Background and introduction **7**
1.1 Methodology 7
1.2 Reading notes 8

2 Storing and preserving research data **9**
2.1 The general European background 9
2.1.1 Riding the Wave (2010) 9
2.1.2 E-IRG (2009) 10
2.2 The researchers' perspective 11
2.2.1 Key findings of PARSE 11
2.2.2 Differences between disciplines 11
2.2.3 Data types and storage 12
2.2.4 Sharing across disciplines 13
2.2.5 Publishers 13
2.2.6 Preservation 13
2.3 Research data during and after the production phase 14

3 Data storage **17**
3.1 Reasons for preservation 18
3.2 Obstacles to preservation and sharing 19
3.3 Non-technical impediments 21
3.4 Needs 22

4 Summary of findings **26**
4.1 The context 26
4.2 The researchers' perspective 26
4.3 Production phase (during the research project) 26
4.3.1 Focus on data protection 26
4.3.2 Key success factors for support are: 27
4.4 Post-publication phase 27
4.4.1 Barriers 28
4.4.2 Focus on trust and control (data sharing) 28

5 Conclusions **29**

Annex 1 – List of sources **31**

Management Summary

In October 2010, the Dutch universities explored possible projects in the area of research data. One of the outcomes of this discussion was the decision to first investigate what researchers need with respect to storing and accessing research data. The present literature study is the result of that investigation. Fifteen sources were studied, consisting of reports from 2008-2010 covering the Netherlands, the UK, the USA, Australia and Europe.

All the stakeholders (funding agencies, data producers, data consumers, data centres) agree for various reasons that something needs to be done to improve research data storage. At the same time, there is no one solution available. **Although there are major differences in the way disciplines conduct their research, they also have a number of factors in common when it comes to data storage and access**. They all encounter both technical barriers, for example the use of obsolete software, and non-technical ones, such as fear of competition, lack of trust, lack of incentives, and lack of control. The literature shows that these non-technical barriers are more powerful than other impediments.

There is an important difference between data storage and access *during* a research project phase and data management *after* publication of the research results. **Storage and preservation are two distinct issues for researchers.** Researchers have expressed a clear need for support in day-to-day storage, but they see preservation as a different step, and one that lies somewhat outside their immediate scope of interest.

During the research project phase, researchers focus on protecting their valuable data. They have expressed their needs in this area and they do indeed need support, because they do not possess the skills, awareness, or knowledge to improve their day-to-day data management.

The literature leads us to conclude that **researchers can, indeed, benefit from support services in managing their digital data, but that these services must meet a number of requirements if they are to be successful**:

- Tools and services must be in tune with researchers' workflows, which are often discipline-specific (and sometimes even project-specific).
- Researchers resist top-down and/or mandatory schemes.
- Researchers favour a "cafeteria" model in which they can pick and choose from a set of services.
- Tools and services must be easy to use.
- Researchers must be in control of what happens to their data, who has access to it, and under which conditions. Consequently, they want to be sure that whoever is dealing with their data (data centre, library, etc.) will respect their interests.
- Researchers expect tools and services to support their day-to-day work within the research project, and long-term/public requirements must be subordinate to that interest.
- The benefits of the support must be clearly visible – not in three years' time, but now.
- Support must be local, hands-on, and available when needed.

Most researchers are unwilling to automatically accept responsibility for preserving their data after publication. The need for fine-grained access control remains very important. Researchers assume that local storage gives them more control over their data than remote storage in a data centre. At the same time, they admit that remote storage will probably alleviate some of the burden of data management. **In all cases, when the data is transferred to another party, researchers wish to remain in control of their data**.

Preservation of research data after the publication phase is possible only when storage during the research project phase has been well managed. In fact, **it makes sense to invest in better data management during the research phase because doing so will improve data preservation once the research phase has ended.**

Managementsamenvatting

In oktober 2010 hebben de Nederlandse universiteiten de mogelijkheden onderzocht voor diverse projecten op het terrein van onderzoeksdata. Een van de uitkomsten van deze discussie was de beslissing om eerst in kaart te brengen waar onderzoekers behoefte aan hebben bij de opslag van en de toegang tot onderzoeksdata. Dit literatuuronderzoek is het resultaat van deze inventarisatie. Er zijn vijftien bronnen onderzocht, waaronder rapportages in Nederland, het Verenigd Koninkrijk, de Verenigde Staten, Australië en Europa over de periode 2008-2010.

Alle betrokkenen (financiers, dataproducenten, datagebruikers en datacentra) zijn het om verschillende redenen eens dat er iets moet gebeuren om de opslag van onderzoeksgegevens te verbeteren. Er is echter niet een uniforme oplossing voorhanden. **Hoewel er grote verschillen bestaan in de wijze waarop binnen de diverse disciplines onderzoek wordt gedaan, zijn er ook gemeenschappelijke factoren wanneer het gaat om de opslag van en toegang tot onderzoeksdata.** Alle disciplines kampen met technische belemmeringen, zoals het gebruik van verouderde software, en niet-technische belemmeringen, zoals angst voor concurrentie, wantrouwen, gebrek aan beweegredenen en verlies van controle. Uit de literatuur blijkt dat met name deze niet-technische belemmeringen een groot struikelblok vormen.

Er is een belangrijk verschil tussen dataopslag en -toegang in de onderzoeksfase en het databeheer na publicatie van de onderzoeksresultaten. **Voor onderzoekers zijn opslag en archivering twee verschillende zaken!** Onderzoekers hebben een duidelijke behoefte aan ondersteuning bij de dagelijkse gegevensopslag. Zij zien archivering als een volgende stap, die enigszins buiten hun directe aandachtsveld ligt.

Tijdens de onderzoeksfase gaat de aandacht van de onderzoeker met name uit naar de bescherming van zijn waardevolle data. Onderzoekers hebben aangegeven waar zij op dit gebied behoefte aan hebben en zij hebben inderdaad ondersteuning nodig, omdat het hen aan de vaardigheden, de kennis en het bewustzijn ontbreekt om de dagelijkse opslag van hun data naar een hoger plan te tillen.

Uit het literatuuronderzoek kan worden opgemaakt dat **onderzoekers bij het beheer van hun digitale gegevens inderdaad profijt kunnen hebben van ondersteunende diensten, maar dat deze ondersteuning dan wel aan de volgende door hen gestelde eisen moet voldoen:**

- Tools en diensten moeten afgestemd zijn op de werkstromen van de onderzoekers, die vaak per discipline (en soms zelfs per project) verschillend zijn;
- Onderzoekers willen geen regeling die van bovenaf en/of verplicht wordt opgelegd;
- Onderzoekers willen het liefst een "cafeteria"-model waarbij zij zelf uit een pakket diensten kunnen kiezen;
- Tools en diensten moeten gebruiksvriendelijk zijn;
- Onderzoekers moeten controle hebben over wat er met hun gegevens gebeurt en met wie zij die gegevens onder welke voorwaarden uitwisselen. Dit betekent dat zij erop moeten kunnen vertrouwen dat degene die hun gegevens (datacentrum, bibliotheek, etc.) beheert, hun belangen respecteert;
- Onderzoekers verwachten dat de tools en de diensten een ondersteuning zijn bij hun dagelijkse werkzaamheden binnen het onderzoeksproject – langetermijnbelang/algemeen belang dient hieraan ondergeschikt te zijn;
- De voordelen van de ondersteuning moeten duidelijk zichtbaar zijn – niet over drie jaar, maar nu;
- De ondersteuning moet lokaal worden aangeboden, praktisch zijn en beschikbaar zijn op het moment dat het nodig is.

De meeste onderzoekers zijn niet zonder meer bereid om verantwoordelijkheid te dragen voor de opslag van hun data na publicatie. De noodzaak van fijnmazige toegangscontroles blijft erg belangrijk. De onderzoeker gaat ervan uit dat het lokaal opslaan van zijn data meer controle oplevert over de data dan opslag op afstand in een datacentrum. Aan de andere kant geeft hij toe

dat opslag op afstand de lasten van het databeheer deels wel zal wegnemen. **In elk geval wil de onderzoeker controle over zijn gegevens houden wanneer die gegevens aan derden worden toevertrouwd**.

Bewaring van onderzoeksgegevens na de publicatiefase is alleen mogelijk wanneer de opslag tijdens de onderzoeksfase goed is geregeld. **Het is zelfs zinvol om al tijdens de onderzoeksfase in een beter gegevensbeheer te investeren, omdat dit de kwaliteit van de gegevensbewaring na het onderzoek alleen maar ten goede komt.**

1 Background and introduction

In October 2010, the Dutch universities that participate in SURFfoundation's SURFshare programme discussed a possible project in the area of research data. One of the outcomes of this discussion was the decision to first investigate what researchers need with respect to storing and accessing research data. The present literature study is the result of that investigation.

There has been a steady stream of publications on the subject of research data storage and access in recent years. It was thought that a summary of these publications and reports describing what researchers want and need would provide a good foundation for a project or other activities. Questions that needed to be answered were:

- How are researchers currently handling research data storage and access?
- What problems have they encountered with regard to storage and access?
- What are their needs with regard to storage and access?
- What projects (in Europe and elsewhere) are dealing with this theme?

This study will take these questions as a starting point and attempt to find the answers. The focus will be on research conducted in the Netherlands, Europe, the USA and Australia exploring what researchers need to enable them to store their research data and make that data accessible. Answering this question will clarify any problem that may exist. Secondly, we will try to clarify the nature of the problem context: who is facing a problem and what is the problem? We have to test our assumptions. For example, we might assume that researchers should be more aware of the importance of preserving their data and therefore need to be trained, but in fact researchers may have a totally different perspective on what the problem is.

In short, this study will focus as much as possible on the researcher's perspective and define the problem from that starting point. This study will not investigate the needs of data managers, funding agencies, librarians or other parties. Technical or organisational issues and issues related to selection of data to be preserved have not been included in this study.

1.1 Methodology

This study draws on a variety of different sources. They are listed in the table below and in Annex 1: List of sources.

Table 1: Sources used in this report

Name	Year	Focus	Disciplines	Size of survey	Region
Vd Graaf	2010	Organisation	Not specified	Small	NL
Rombouts	2009	Data Centre	Technical Sciences	Small	NL
Verhaar	2010	Researcher	Arts & Media	Small	NL
Tjalsma	2010	Data Centre	Humanities & Technical Sciences	Small	NL
PARSE	2009	Researcher	All	1389	EU, USA, other
Beagrie	2009	Researcher	STM[1]	700	UK
RIN	2010	Researcher	Not specified	Unknown	UK
UKRDS	2010	Researcher	STM	700	UK
RIN/BL	2009	Researcher	7 disciplines	56	UK
OCLC	2010	Researcher	All	Unknown	USA
Henty	2008	Researcher	All	879	Australia
Kuula	2008	Open Access	Humanities	150	Finland
E-IRG	2009	Researcher	All	Literature	Europe
Riding the Wave	2010	Policy	Expert Group	Report	Europe

[1] Science, Technology and Medicine.

The sources have been selected based on desk research and input from members of the SURFshare community. We owe a special word of thanks to Inge Angevaare for her feedback and support.

The first step in our methodology was to analyse the size of each study and the discipline or disciplines concerned. We summarised the outcomes and conclusions and compared them with outcomes from other sources. We also analysed similarities and differences between studies in similar disciplines, as well as sources pertaining to other disciplines.

Many sources begin their conclusions by saying that there are major differences between disciplines. Most of the time it is very clear that the way things are done and the related problems and needs are specific – even very specific – to one discipline. Sometimes, however, issues can be identified that affect many, if not all, disciplines. In those cases it is possible to speak of "common" factors between disciplines, but this needs to be seen against the background of the differing interpretations of the various sources used.

The reader should bear in mind that each source used for this study has its own research objective and background. Although the questions posed to researchers appear to be similar in many cases, the way their answers are interpreted may be "coloured" by the research objective. Each source has its own distinct focus of analysis and interpretation, for example organisation of preservation, or quality of research data, or management of a data archive.

It is interesting to note that researchers are only the subject of research in the sources used. It seems that they have not themselves conducted research on their research data; no such reports were found in the sources used. There may also be cultural and local differences in perspective, but these are usually more implicit.

All sources offer a "snapshot" of the situation under review. In other words, this study is a summary of snapshots. In drafting this summary, I have tried to represent the facts and opinions as given in the sources, leaving out any personal interpretations and commentaries. Inevitably, however, the mere fact that I made a selection, leaving out things that may have been important on second thought, means that this summary is imperfect. I accept full responsibility for any misinterpretation of the materials presented in the sources used for this study.

1.2 Reading notes

The introductory chapter of this study summarises the general European background, based on two sources: the Riding the Wave report and the E-IRG report. The Wave report gives us a nice general overview of the state of affairs in Europe. The E-IRG report surveys current European projects and initiatives concerning data management. The second chapter focuses on the researchers' perspective. The PARSE report is a very comprehensive source and it offers an excellent glimpse of the opinions, experience and needs of researchers. This is followed by an investigation of the other sources, with a focus on researchers' experience, problems and needs with regard to data storage, preservation, accessibility, and sharing. The closing chapter presents final observations and conclusions.

Each source is identified by an abbreviated name and the year of publication, for example (E-GRID, 2009). All sources are listed in alphabetical order in Annex 1. They are referred to as "sources", whereas the literature study itself is referred to as "the study". Percentages are rounded off; for example, when the source reports 36.2%, this study will round it down to 36%. Whenever a secondary source is cited (such as a report or a quotation in a primary source), it will be listed in a footnote. When an original source is quoted in this study, the text will be in italics and between quotation marks.

2 Storing and preserving research data

2.1 The general European background

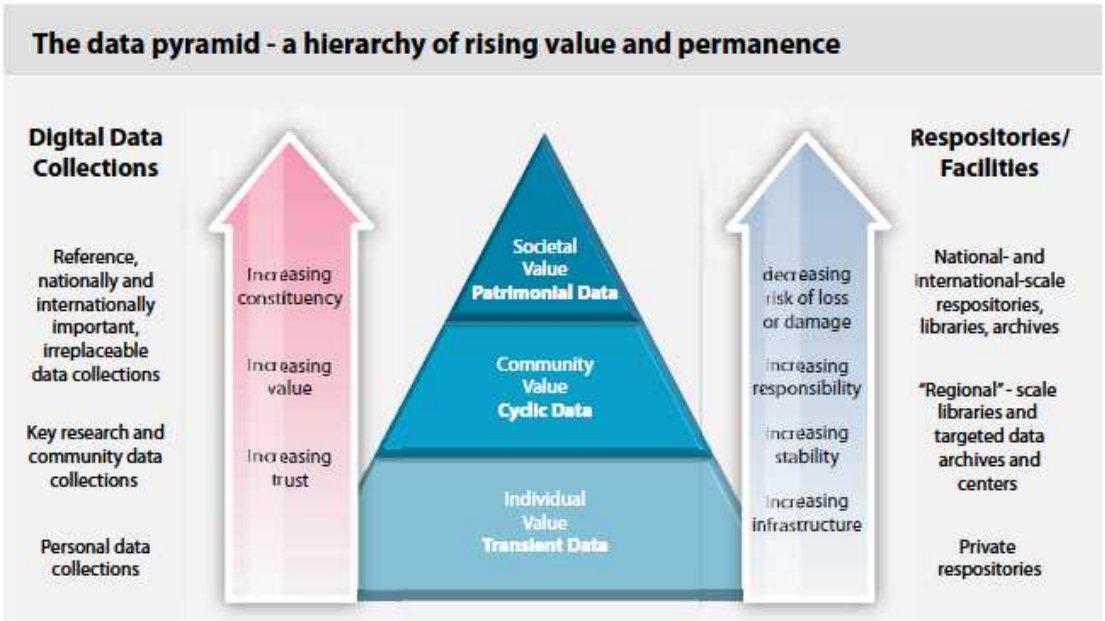
The introductory chapter of this study summarises the general European background, based on two sources: the Riding the Wave report and the E-IRG report. The Wave report gives us a nice general overview of the state of affairs in Europe. The E-IRG report surveys current European projects and initiatives concerning data management.

2.1.1 Riding the Wave (2010)

The Riding the Wave report, published in October 2010, provides a general picture of the most important current issues with regard to research data. The expert group that wrote this report offer a vision of how things might stand in 2030 and summarises the obstacles that stand in the way of this vision. The report was commissioned by the European Commission.

The authors emphasise new general trends in science:
"We all experience it: a rising tide of information, sweeping across our professions, our families, our globe. We create it, transmit it, store it, receive it, consume it – and then, often, reprocess it to start the cycle all over again. It gives us power unprecedented in human history to understand and control our world. But, equally, it challenges our institutions, upsets our work habits and imposes unpredictable stresses upon our lives and societies…
Most importantly, however, our focus is on scientific data because, when the information is so abundant, the very nature of research starts to change. A feedback loop between researchers and research results changes the pace and direction of discovery. The 'virtual lab' is already real, with the ability to undertake experiments on large instruments in other continents remotely in real time. Researchers with widely different backgrounds – from the humanities and social sciences to the physical, biological and engineering sciences – can collaborate on the same set of data from different perspectives. Indeed, we begin to see what some have called a 'fourth paradigm' of science – beyond observation, theory and simulation, and into a new realm of exploration driven by mining new insights from vast, diverse data sets…"

Data itself is seen as the infrastructure, consisting of several layers:



Source: *Riding the Wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission* (October 2010).

As part of the report, the authors offer a vision of how things might be in 2030, but in doing so they ask themselves: *"…How can we get researchers – or individuals – to contribute to the global data set? …To start with, this will require that they* ***trust*** *the system to preserve, protect and manage access to their data; an incentive can be the hope of gain from others' data, without fear of losing their own data. But for more valuable information, more direct incentives will be needed – from career advancement, to reputation to cash…"* They also provide a list of possible hurdles to overcome if their vision is to become reality. Note particularly the items about lack of trust and complexity.

What could jeopardise the vision?

Impediments	What we could do to overcome them
Lack of long term investment in critical components such as persistent identification	Identify new funding mechanisms Identify new sources of funding Identify risks and benefits associated with digitally encoded information
Lack of preparation	Ensure the required research is done in advance
Lack of willingness to co-operate across disciplines/ funders/ nations	Apply subsidiarity principle so we do not step on researchers' toes Take advantage of growing need of integration: within and across disciplines
Lack of published data	Provide ways for data producers to benefit from publishing their data
Lack of trust	Need ways of managing reputations Need ways of auditing and certifying repositories Need quality, impact, and trust metrics for datasets
Not enough data experts	Need to train data scientists and to make researchers aware of the importance of sharing their data
The infrastructure is not used	Work closely with real users and build according to their requirements Make data use interesting – for example integrating into games Use "data recommender" systems i.e. "you may also be interested in..."
Too complex to work	Do not aim for a single top down system Ensure effective governance and maintenance system (c.f. IETF)
Lack of coherent data description allowing re-use of data	Provide "forums" to define strategies at disciplinary and cross-disciplinary levels for metadata definition

Source: *Riding the Wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission* (October 2010).

2.1.2 E-IRG (2009)

The E-IRG report is very valuable in that it presents a picture of European data management initiatives. It describes many different European projects on data management involving researchers (18 in the social sciences, 12 in the health sciences, 33 in science and engineering). The aim of this report was not to acquire a better understanding of the needs of researchers, however, so the recommendations are not written from the researcher's perspective (for example, "*R8: researchers need to be urged to produce high quality metadata descriptions*").

The E-IRG report is important as a source of information about research trends that may be relevant for a better understanding of the needs of researchers, such as:

- *"…New requirements for cross-disciplinary research will require interoperability between different disciplines and different types of data;*
- *The focus is on curating data for reuse, not necessarily for long-term preservation..."*

Some discipline-specific items are:

- "*…in social sciences important datasets are often collected not by research teams but by government to inform policy makers. In some cases archives have negotiated access to these data for the wider research communities…"*
- *"…European e-Health Action Plan is an important driver for the creation of high quality electronic health record systems in the EU. The field will need to come to consensus on the length of time to store these records…*"
- "*…natural sciences and engineering are experiencing an unprecedented data avalanche due to fast evolution of sensor/detector technology and advances in IT that enable capture, analysis, storage and huge quantities of data. But distinction between raw and processed data is difficult because of pre-processing in or close to the detectors…*"

2.2 The researchers' perspective

For a summary of the current views of European researchers on their research data, we must turn to the PARSE.Insight Survey of 2009. The survey underlying the report was conducted among 1389 researchers from all disciplines as well as data managers (273) and publishers (178). The PARSE report is the most comprehensive source of information on what researchers think about research data so far.

According to this source, there are an estimated 1.33 million researchers in Europe. Statistically, it would take at least 385 respondents to obtain an adequate picture, and 1389 researchers (609 from EU Member States, 465 from the USA, and 315 from other countries) responded to the PARSE survey. The authors of the report therefore consider the survey as representative.

One of the aims of this survey was *"…to gather information on the practices, ideas, and needs of research communities regarding the preservation of digital research data…"*

2.2.1 Key findings of PARSE

"…Researchers consider the possibility of re-analysis of existing data as the most important driver for the preservation of research data; 91% of the respondents thought this to be either important or very important.

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

2.2.2 Differences between disciplines

Besides these key findings, the PARSE report contains an extensive overview of the various issues that researchers believe are connected with research data. Because of the importance of this overview for this study, the relevant texts of the PARSE report are reproduced below. They include an analysis of similarities and differences between disciplines and possible differences due to the age of the respondents.

➢ **Physical sciences compared with other disciplines**

"…When it comes to similarities and differences between disciplines, we compared the largest group (physical sciences) with the other categories of disciplines. In our survey the physical

sciences is an aggregation of the disciplines astronomy & astrophysics, chemistry, computer sciences, mathematics and physics. The following comparisons and differences were found:
- respondents from physical sciences seem to deal with more data within their current research project. Not only now, but also in 2 and 5 years;
- seem to be more eager to make their own data openly available. They scored about 10% higher than respondents from other disciplines;
- compared to other respondents, slightly less (about 5%) of the physical sciences respondents are interested in data outside their own discipline;
- about 10% more respondents of physical sciences stated that they use general search engines for finding new information on their research topic;
- regarding data that had already become inaccessible, physical sciences respondents stated that this is most often due to the fact that software to interpret the data is no longer available, while the rest of the respondents in research regard this more as the result of hardware problems;
- as to where to publish data, respondents from other disciplines stated that they are most willing to publish data in an archive of their own organization, while physical sciences respondents prefer to publish data in a specific archive of their own discipline;
- however, respondents from physical sciences and other respondents seem equally unaware of available external preservation facilities such as data archives or other services;
- regarding the influence of funders, a small majority of physical sciences respondents stated that funding organizations do provide mandatory procedures for managing and preserving digital research data. Others either thought funders didn't or they didn't know…"

- **Age**

"…only two bigger discrepancies (> 10% discrepancy) were found between less experienced (and potentially younger) researchers and those that have been working in research for more than 20 years. First, novice researchers seem to be much more eager to use research output from other disciplines than experienced researchers. Secondly, although online collaboratories are not very commonly used by either of the groups of researchers, novice researchers are more in favour of having these kinds of platforms preserved as well…"

- **Other differences**

"…Participants from all disciplines agreed (with the exception of behavioural scientists) that the results of publicly funded research should become public property and therefore be properly preserved. They also agreed that preservation will stimulate the advancement of science (new research can build on existing knowledge) and allows for re-analysis of existing data.

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

There is a significantly higher than average percentage of humanities researchers (75%) who feel that there is a need for an e-science infrastructure to counter the threats to digital preservation. Apart from an infrastructure, the respondents believe that training, guidelines and manuals would be useful to raise the level of knowledge…"

2.2.3 Data types and storage

The PARSE survey included questions about the types of data used in the research project phase and the way they were stored.
"...*Not surprisingly, MS Office documents are most often used by the respondents. What is a bit surprising perhaps is that still 6% of the respondents do not use MS Office documents. The other two of the top three most used data types are: network-based data (web sites, e-mail, chat history, etc.) and images (such as JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc). For both data types 79% of the respondents claimed to use them.*

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

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

2.2.4 Sharing across disciplines

When researchers were asked about sharing their data, PARSE reports:
"…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. Only 11% of the respondents make their data available for researchers within their research discipline. The majority of respondents do make their data available to researchers within their research collaborations and groups, but even 58% may not be considered a very high figure.

While the percentages of the respondents who share data are small, sharing does take place. However, the sharing of these data does not seem to take place through established digital archives, not even when they are specific to the discipline. The obvious conclusion would be that researchers want some sort of control over their data and they see many problems surrounding the sharing of data. The major problems researchers foresee in sharing their data through digital archives are legal issues (41%), misuse of data (41%), and incompatible data types (33%). Based on the responses, it looks like there still is a lot of distrust in the capability of digital archives to properly handle research data. The current practice is not to be explained by a disinterest of researchers in other people's data. 63% of the researchers, who do not currently make use of other researchers' data within their discipline, would like to do so in the (near) future — 40% for data from other disciplines.

When asked whether they ever truly needed digital research data by other researchers that was, for whatever reason, not available, 53% of the respondents answered yes…"

2.2.5 Publishers

PARSE also reports the responses of the 178 publishers who participated:
"…Many publishers outsource the digital preservation of their journals. Again, this is overall better arranged for larger publishers than for smaller publishers. A few of the organisations publishers outsource to are: Portico (30%), CLOCKSS (13%) and KB (7%).

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

2.2.6 Preservation

PARSE reports what managers of data centres do and how that compares with the data actually used by researchers:
"...This report is about preservation of data of any kind. It is interesting to see if there is a match between what researchers use and what data managers actually preserve for the long term. Normalizing the responses of this question for both research and data management shows some interesting differences (see Figure 63). Looking at what is used by researchers but not adequately

preserved by data managers (see blue peaks of second graph) indicates that: network-based data (such as websites), source code, computer applications and raw data are often not preserved well…"

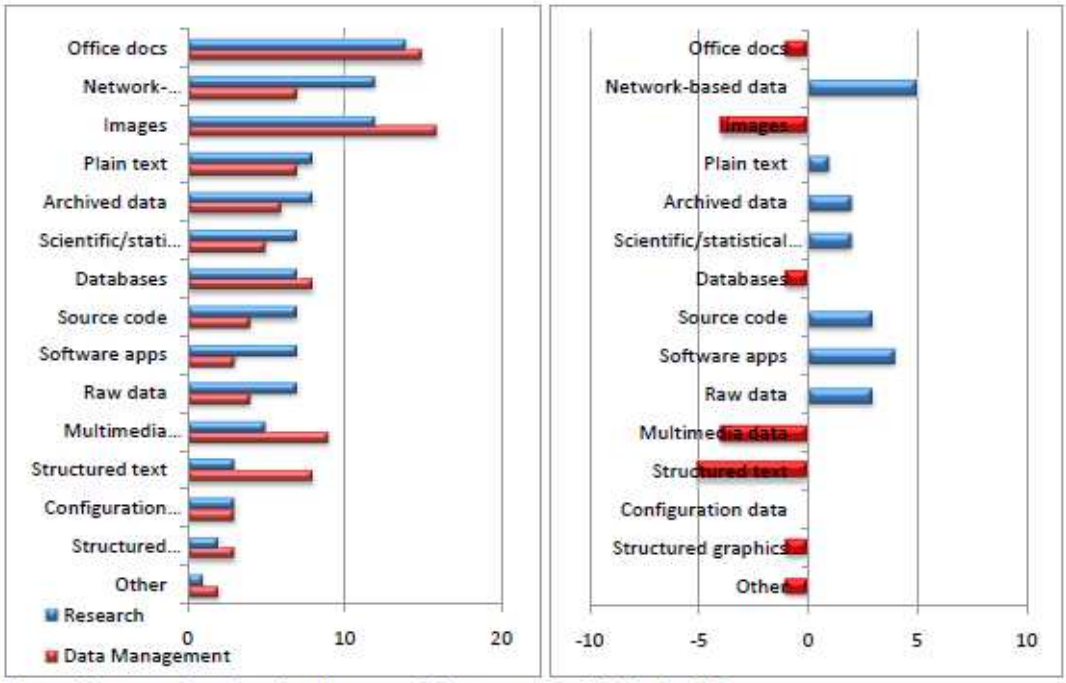


Figure 63: comparison of used and preserved data types, n1 = 1366, n2 = 206

Source: *Open access to and reuse of research data: the state of the art in Finland.* Kuula (2008).

2.3 Research data during and after the production phase

In considering research storage and access, it would be useful to have some idea of what we mean by the words "research data". Kuula and Borg (2008) addressed this question and present the following table on page 28 of their report:

Research publication	Research data
Information transformed into results	Information not transformed into results
Use requires basic software and instruments and their command	Use often requires special software and instruments and their command
Self-explanatory	Requires additional information and documentation if not archived
Should not include sensitive information	May include sensitive and confidential information
Use does not require permission	Use often requires permission
Ownership and copyright often clear	Ownership and copyright often unclear
Openly accessed by the scientific community for a fee or for free	Several degrees of openness (from completely open to closed)
Understood as scientific output (mentioned in the CV)	At the moment not understood as scientific output/merit even if the data were published (usually not mentioned in the CV)
Ready to be used by others as such	Use requires processing

Table 6.2
Research publication, research data and Open Access: a simplified difference chart

Source: *Open access to and reuse of research data: the state of the art in Finland.* Kuula (2008).

It is very interesting and relevant for this study to compare this with the PARSE report. Here, we find *"…In PARSE.Insight the term digital research data is used for all output in research. In practical terms, raw data, processed data and publications are all covered by the same term. A distinction between these sorts of research data is only made when necessary (for example when policies for publications are compared with other data)…"*

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

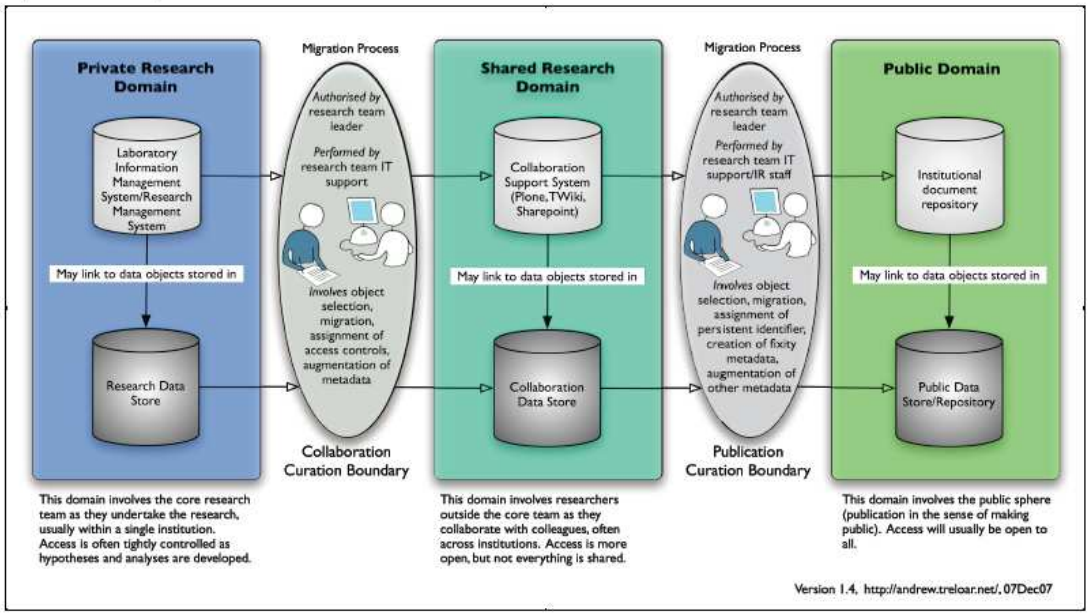
Rombouts (2009) also writes about this, using different wording and from a different perspective. He draws a distinction between dynamic data (as part of a research project that is active; the data is therefore dynamic) and static data (research has ended and data is stable). Protection of data during the research project phase is essential because data is frequently re-used during this phase. It is stored locally and this may lead to data loss or storage capacity problems. During the research project phase, responsibility for data storage and data management is in the hands of researchers.

Treloar refers to "data domains" in his well-known article:[2] *"…The first domain is the private research domain. This is where the immediate research team is working with its data and producing its results…The second domain is the shared research domain. Here the research team is prepared to open up a subset of its research results to other researchers to access and*

[2] Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". *Proceedings of VALA 2008*, Melbourne, February. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

analyse…The third domain is the public domain. At this point, the research is 'finished' in the sense that the resulting publications (and possibly linked data objects) are available for public viewing."

Figure 1: Domains, Data Stores and Curation Boundaries



Source: Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships*". Proceedings of VALA 2008*, Melbourne, February.

3 Data storage

What is it that researchers do with their research data? Are they satisfied with what they do and how they are doing it? Are they encountering problems? Do they want help? And if so, what kind of help, and in what form?

The PARSE report (2009) observed that most research data is held locally, on individual PCs and departmental servers. Convenience, backup (probably referring to storage on a departmental server) and ease of access are listed as important factors for using the preferred storage solution. Less than 20% of the respondents use a national/international data deposit facility.

Beagrie (2009) writes: *"...Local data management and preservation activity is very important with most data being held locally…"*

Kuula (2008) reports along the same lines for Finland: most researchers retain their data themselves (85%). A minority (28%) store the data in the department/institute without further processing or documentation. Generally speaking, there is no long-term preservation plan.

Henty (2008) concludes that researchers employ many different means to store and back up their data, often using unreliable and short-lived storage media such as USB/Flash drives, CD-ROMs and DVDs. In Henty's estimate, 38% use network storage, while 78% are responsible for their own data management, even after the end of the research phase, and see data management "*as another bureaucratic requirement being imposed on their time*". They complain about the difficulty of keeping track of large and diverse collections.

Significantly, this source concludes that there is still a wide range of non-digital data being collected, and that researchers need support in digitisation. Another important observation is that researchers are using a wide range of specialised software and that they are not aware of the implications of their software choices for later access to their data.

In the same vein as the findings of the PARSE report, Henty concludes that MS Office documents (Word or Excel) and databases (SPSS) are the most common data types used (two thirds of respondents). Another 50% use data that is automatically generated through software. About 40% have experimental data and e-mail and, although their numbers are diminishing, data collected from sensors, images, scans, fieldwork data, audio, video, websites, lab notes, and blogs.

In the USA the OCLC report (2010) shows that researchers use a "*hodge-podge of approaches*" such as storing hardcopy items (e.g. lab books or printouts) on bookcases in their office, retaining working files on the computer initially used to create the files, storing files on a variety of devices in the lab or at home, and storing files and data in sophisticated computer centres. There is no breakdown of these various approaches in percentages, but the report concludes: "…*Based on the comments during this study, it appears that universities are doing a uniformly poor job storing, maintaining, and providing access to the discoveries they are encouraging their faculty to pursue through the research process. Individual faculty members are unable to solve this problem themselves; meanwhile, many continue to store documents and data in a haphazard manner*…"

VdGraaf (2010) mentions the outcome of five data audits at the universities of Melbourne and Edinburgh in the departments of physics, theology, economics, social history and brain research: "…*In most cases there was insufficient storage space available on the servers. Researchers themselves stored many datasets in a more or less ad hoc manner on personal external storage devices with little chance of effective retrieval. The majority of the participants' data were very valuable and it would be difficult to regenerate the data in case of a loss. In some cases, particularly interviews and surveys, it would be impossible to regenerate the data. Searching for the data was difficult as most of the data was undocumented and there was not a well-defined folder structure. Metadata was sparse at best. A clear message that came through was the urgent*

need to develop greater awareness and understanding of data management within the university as well as guidance on best practice..."[3] He also mentions that researchers resisted "improved" data management, not only because it might take extra time, but also because it might generate extra costs in cases where the university ICT department charges for storage capacity on network servers.

3.1 Reasons for preservation

What reasons do researchers give for wanting to preserve their research data?

Recently, SURFfoundation published a report on guidelines for selecting research data (Tjalsma, 2010). Although the guidelines are not based on input from the research community, they provide a useful description of possible reasons for preservation: reuse, verification, or heritage (or a combination of the three).
*"…There may be an **obligation** to preserve the research data so that it can be **used/reused**. An obligation of this kind might be imposed by research funding bodies, academic publishers or others. If there is **no obligation**: are there valid reasons to preserve the research data so that it can be **used/reused for research purposes**? Are those reasons valid from the perspective of the research discipline where the data were created or academic disciplines other than the original research discipline? These reasons could be:*

- *Value of the data: potential value in terms of reuse, national/international standing and quality, originality, size, scale, costs of data production or innovative nature of the research.*
- *Uniqueness of the data: the data contain non-repeatable observations.*
- *Importance of the data for history, in particular the history of science.*

*Or there may be an **obligation** to preserve the research data for **verification purposes**. This could be an obligation based on an existing code of conduct for research, like the Netherlands Code of Conduct for Scientific Practice (2005), which prescribes storage of raw data for at least five years.*
*Or there may be reasons to preserve the research data for **general (basically non-academic) purposes**. For example, data important for cultural heritage reasons, for museums or for other presentations…"*

Looking at other sources, especially those based on input from research groups, we see that they too mention the above reasons as those given by researchers for the preservation of data, although in other words or for other motives.

In four universities in the UK (Beagrie, 2009), 33% of the researchers reported a grant requirement and 14% a legal requirement to retain their data. A surprisingly high number of respondents believe there is no requirement (about 26%) or do not know (36%) whether there is a grant or legal requirement to retain their research data.

Kuula (2008) describes the expected benefits of open access to research data. These include reduction of duplicate research efforts, cost reduction and more effective use of data, better cooperation and improved access to data. A large number of respondents are very (33%) or fairly (43%) positive about sharing their own data through open access. A majority (71%) said researchers should be encouraged to share their data, but there should be no obligation.

Verhaar (2010) claims that researchers consulted in his survey all agreed that transferring files to a data archive may have benefits: they would not have to manage the data themselves, it would enable others to re-use their data, it could lead to new, unexpected scientific output, and it would enhance verification of research conclusions. Researcher attitudes towards data sharing were found to be positive.

Waaijers (2010) reports that in the life sciences, 17% of the respondents claim to deposit their data because it is required by the scientific journal. Other disciplines show a similar outcome.

[3] Edinburgh Data Audit Implementation Project. See: http://ie-repository.jisc.ac.uk/283/

Rombouts (2010) and other sources mention the costs involved in generating research data. In many cases, collecting raw data is a time-consuming process and initial processing requires specialists and specific equipment. It can be very expensive to generate research data, and that fact alone has an impact on storage and use. For example, the owners of geodata are not very keen to allow free access to their data. They do, however, put data storage and preservation high on their list of priorities.

According to Verhaar (2010), some researchers noted that research data may be used for educational purposes, including, for example, to teach research methods. He mentions the possible benefits for cultural heritage institutions of opening their digitised resources as input for humanistic scholarship, but also to the general public in order to generate interest in their collections.

3.2 Obstacles to preservation and sharing

Hands-on experience in Dutch data centres (Rombouts, 2009) demonstrates that the researcher is the most important source of information about research data. But researchers have neither the time nor the necessary experience to document their research work or to structure the dataset according to standards, or even to add metadata. They seem to expect the data manager to resolve these issues.

In the humanities, says Verhaar (2010), "*…an important obstacle is that it is currently very difficult to discover what datasets other scientists are actually producing. As for releasing data, most researchers state this can only happen after they have been discussed them in a published text. This is because sources such as databases or models have often required much intellectual effort to produce, but earn no scholarly credits on their own. The value of humanistic studies is often related to the uniqueness or the originality of those ideas, which data sharing might undermine…*" Note that Verhaar mentions a lack of incentives: there is no academic reward for resources other than text publications.

Beagrie (2009) concludes that most researchers share their data (only 12% of his respondents do not) and that sharing take place mostly through informal peer exchange networks (e-mail, websites, USB) within research teams and with collaborators. Over 40% use other researchers' data via a data centre, although only 18% actually share their data via such a centre. Most data is seen as useful for a small number of users.

Henty (2008) reports that over 60% of researchers are willing to share their data, whether openly (9%), via negotiated access (44%), or after the end of a project (9%). Also, about 33% of researchers in his survey (which covered all disciplines) access their data in original or print form. "*Print is not dead,*" is Henty's conclusion.

Waaijers (2010) found that 71% of the respondents in his survey produced research data, 50% consumed it, and 60% had used data generated by others.

Kuula (2008) reports on barriers to preservation and what researchers perceive as the disadvantages of preservation and sharing:

- concerns about inadvertent misuse of data and consequent mistakes;
- lack of agreement on ownership of the data;
- competition for academic positions and funding;
- usability and technical problems (insufficient documentation, obsolete formats, damaged data);
- lack of informed consent and confidentiality.

Only a small minority (21%) thought that half of their data could be reused. Respondents confirmed that they were reusing data, but in most cases this was reuse of their own data, and then mostly for educational purposes.

Verhaar (2010), Rombouts (2009) and others refer to the privacy of research data as a limiting factor. This aspect is also important in the Parelsnoer Initiative,[4] in which personal medical data on patients and the course of their illness was collected and stored for analysis and exchange between medical specialists.

Verhaar (2010) provides an excellent summary of the main difficulties involved in data sharing (in the arts and media sector). On page 21 of his report, he writes:

- "*Scholars often create unstructured and non-standardised research annotations. Such resources are mostly unfit for reuse.*
- *A temporal dimension should be taken into account. Researchers are usually not willing to share their data before these data have been discussed in a formal publication.*
- *Copyright or privacy laws may sometimes complicate open access to research data. Scholars who wish to share their data should make sure that publication does not produce any legal or ethical issues.*
- *Incidentally, scholars may produce resources which are strongly application-specific, rendering reuse near to impossible.*
- *Researchers generally did not know exactly how they could archive their data. Most respondents were familiar with DANS, but had never actually deposited their own datasets in their archive.*
- *Usually, researchers do not assign metadata to the resources they produce.*
- *There is a lack of a clear mandate. Universities, publishers or funding agents do not actively stimulate scholars to share their data.*
- *Researchers are often unaware of the existence of relevant datasets. There is no central registry of primary resources…*"

Rombouts (2009) writes about hands-on experience in Dutch data centres and concludes that the main obstacle to publishing research data is the lack of professional (academic) credit. But even if the intention to share is there (or if it is an obligation imposed as part of funding), there is usually not enough time or funding. The search for funding for the next research project is a bigger priority.

Researchers may also be afraid of competition: someone else could use the data and base a publication on it before the data producer, thereby getting all the credit for work done by others. It is also possible that errors in the research process and/or research data might be detected.

In cases where a researcher depends on data provided by others for his research, he will be more willing to share his own data with others. But he will do this within his own network, with trusted parties, without interference from others (such as a data centre). Such trust-based sharing is mentioned in various other sources, e.g. Verhaar (2010), OCLC (2010).

In the UK life sciences (RIN/BL, 2009), concerns about data sharing are expressed as follows:
"*…Indeed, researchers highlight a number of barriers to sharing their research data, including concerns about potential misuse, ethical constraints, and intellectual property. Above all, they see data as a critical part of their 'intellectual capital', generated through a considerable investment of time, effort and skill. In a competitive environment, their willingness to share is therefore subject to reservations, in particular as to the control they have over the manner and timing of sharing…"*

"…Researchers are thus willing to share experimental data subject to two strong provisos.

- *First they are concerned that they need sufficient time to complete the analysis and, in some cases, to explore intellectual property rights (though when asked, they are often reluctant to define exactly how long this period of time should be).*
- *Second they want to publish their results before or simultaneously to publishing the data – and they want to be the ones who publish the data.*

[4] See: http://parelsnoer.org/

This is because researchers distinguish between two aspects of the data they collect: recorded 'facts' such as geographic positions, where sharing is unproblematic; and the attributes which they themselves provide and which add value to the data. Many researchers do not want to share the attributes at least until after a paper has been published…"

"…It is also notable that most of our researchers do not wish to re-use research data collected by other researchers because there are so many differences in experimental design and data collection practice…"

"*For all these reasons, researchers prefer collaborative arrangements and direct contact with potential users where differences and intricacies can be elucidated and understood, rather than making data freely available for re-use. This in turn raises questions about how much time should be spent on annotating data.*

Thus while willingness to share is part of the ethos of life science research, individuals like to choose what to share, with whom, and when. Lack of trust in wider 'cyberspace' is pervasive…"

3.3 Non-technical impediments

Resistance to data management following a research project is an example of a non-technical (cultural) issue related to daily data management, as seen by researchers. This paragraph will explore some of these issues.

PARSE (2009) concludes as follows:
"…As became clear from the researchers' survey, researchers tend not to store their digital data at external facilities. At the same, as we have seen, researchers are concerned about legal issues and misuse of their data when it is stored elsewhere. One conclusion we can draw from this is that there is a human, psychological dimension to the issue of preservation which has to do with trust. This cannot merely be tackled by technical solutions. It is another confirmation that a roadmap should include more than these technical solutions…"

The RIN report (2010) reveals that most researchers receive relatively little support in the area of data management, but there is a growing awareness that much needs to be done. They see library staff as less proactive in reaching out to researchers with customised information support, and staff at the research office as more proactive. There is a thriving e-mail culture: e-mail systems are used as a primary mechanism for a wide range of tasks, including managing personal information collections. Researchers feel confused by many different policies and practices relating to the management and deposit of supplementary and other material in addition to e-prints, including lab notes, images and data.

But there is even more to consider. The UK case studies (RIN/BL, 2009) of researchers in the life sciences show that there is a significant gap between how researchers behave and the policies and strategies of funders and service providers. This source concludes:

"*…Researchers use informal and trusted sources of advice from colleagues, rather than institutional service teams, to help identify information sources and resources.*

The use of social networking tools for scientific research purposes is far more limited than expected.

Data and information sharing activities are mainly driven by needs and benefits perceived as most important by life scientists rather than 'top-down' policies and strategies.

There are marked differences in the patterns of information use and exchange between research groups active in different areas of the life sciences, reinforcing the need to avoid standardised policy approaches…"

The authors of this report stress the importance of sustained and proactive contact between information professionals and research communities. Support needs to be based upon a close understanding of the researchers' work. Attention must be paid to the use of specific terms such as "data", "information", and "database", because it became clear that researchers and information specialists interpret such terms differently. Researchers believe that the best way to curate their data is to do so themselves. But they need to develop broader skills (training) that enable them to do this.

"...A key message from our work, therefore, is that policy intervention and support systems for researchers need to be built around the many different and successful tools and practices emerging within life science research communities themselves..."

3.4 Needs

Three Dutch universities of technology asked their researchers what they expected from a data centre (Rombouts, 2009). Besides quality of research data, organisation, standardisation, search functionalities and metadata, they mentioned:

- ease of use: storage must be simple;
- control: storage must be in a protected, private access area;
- control: the researcher should decide what is stored, when and where.

Verhaar (2010) writes "*…Developing facilities for the curation of data is complicated by the fact that different disciplines differ widely in terms of research methods and the type of data that are produced. A recent report produced by the SCARP project in the United Kingdom concluded that 'researchers' attitudes and practice with regard to the creation, sharing, reuse and long term care of data are closely linked to the discipline in which they work'. For this reason, 'a generic approach to data curation will not be sufficient to cope with the different data-related needs and expectations of researchers working in different disciplines other than at a superficial level'* (*Key Perspectives, 2010, p. 2) …These findings suggest that the key to enhancing the Dutch infrastructure for data curation lies in firstly exploring the needs of researchers working in fields across the entire academic spectrum, and, consequently, to develop facilities and guidelines which can adequately address these domain-specific needs…"*

Although Waaijers' research (2010) focused on quality of research data, it is interesting to note his list of actions intended to enhance the quality of research data. In the figure below, the actions are ranked according to the respondents' answers per discipline. It is understandable that researchers do not like actions that force them to do more work (e.g. the final two actions), but it is interesting to see that training is also not very popular, and that a code of conduct is fairly well received except in the physical sciences. In contrast, open access to datasets is well received in the physical sciences but not in the other sciences.

	Physical Sciences and Engineering	Social Sciences and Humanities	Life Sciences
Creation of data publications: peer-reviewed descriptions of datasets	+++	+++	++
Citation of datasets	+++	+++	++
Comments on the quality by data users, to be published with the dataset	+++	++	+++
Open Access availability of datasets, possibly with a no-use period	++++	++	
Creation of a code of conduct for researchers on data management and availability of datasets		++	+++
Peer review of a dataset as part of peer review of the publication	+	+	++
Data management training		+	++
Imposing data audits at regular intervals	−−−−	−−	+
Obligatory data management paragraph in proposals	−−−	−−	−

Source: *Kwaliteit van onderzoeksdata, een operationele benadering.* Leo Waaijers, Maurits van de Graaf. Preprint version (December 2010).

Focusing on training, we see (Henty, 2008) that there may be some specific needs: "*Training is sought in areas related to data management planning, either prior to a project or after, digitisation and data rescue (for older materials)...*"

Beagrie (2009) summarises the opinions expressed by 37 respondents at Oxford as follows:

- *"Overall, the vast majority of researchers interviewed thought that there are potential services that could help them to manage their data more effectively.*
- *Advice on practical issues related to managing data across their life cycle. This help would range from assistance in producing a data management/sharing plan; advice on best formats for data creation and options for storing and sharing data securely; to guidance on publishing and preserving these research data.*
- *A secure and user-friendly solution that supports storage of large volumes of data and their sharing in a controlled way that will permit the use of fine-grained access control mechanisms."*

In Australia (Henty, 2008), we find that most researchers are willing to share their data, but they would like sharing to be easier than it is now. They resist the bureaucratic requirements that they expect will be part and parcel of data management. The same picture emerges in the UK (RIN, 2010), where there seems to be a demand for simple tools that facilitate the sharing of documents and data with colleagues in other departments and institutions. Researchers also express concerns about preservation of their research data: most researchers receive relatively little support in this area. If support were to be offered as shared services, there is a strong proviso that these should be customer-focused and should meet local circumstances and needs. Researchers suggest that libraries could do more to promote their services. "…*The key requirement from most researchers' perspectives is for services which are there when they need them, but do not interfere with the creative work at the heart of the research process."* There is a flourishing e-mail culture and any support given must take this into account. Researchers will need to see obvious benefits; otherwise support services will not be used.

In the USA (OCLC, 2010), researchers described the particular research support services that they lack. Data management and storage topped the list of gaps in research tools and services. "…*Researchers Value Ease of Use and Increased Efficiency. Researchers live by satisficing. Because of tremendous pressures on their time, researchers adopt information tools and services that are easy to use and that simplify their work, even when those tools and services are not optimal,*

comprehensive, or on the 'approved' list preferred by their university…Despite all of the technological advances, researchers depend upon personal introductions and face-to-face interaction. Faculty and staff at all levels use direct contact as the initial step in any decision to work together. Technology cannot replace that human factor, and tools or services without that element are not well accepted…"

Most sources agree that data volume will increase and that there are (significant) differences between disciplines with regard to expected growth and storage needs. As an example, consider the findings of Beagrie (2009). He writes:
"…A strong need…to focus on long-term (>5 years) and medium-term (1-5 years) data curation and preservation was indicated. There was lower demand for short-term data storage (1-12 months). Overall a 360% growth in data volumes is anticipated over the next three years by researchers in the survey. There is some significant variation within and between disciplines. 48.9% of data is seen as having a useful life of under 10 years. Only about 27% is seen as having indefinite value and retention. There are significant differences across disciplines in retention…"

In the UK, the UKRDS initiative (2008) tried to establish whether there is a need for a national data management service and to determine its feasibility. In preparation, a distinction was made between four groups of stakeholders:
1. Research Funding Bodies
2. Research Data Creators
3. Research Data Consumers
4. Repository Providers

In view of the results of the UKRDS feasibility study, we discern some important findings that are relevant within the scope of this literature study. First of all, we see that the findings (mentioned elsewhere in this study) that Beagrie used in his article (Beagrie, 2009) are repeated in the UKRDS study. This was to be expected, because both are based on the same surveys, interviews and desk research. In addition, the UKRDS study concludes that:

"…Research Data Creators *- While research data creators were in many cases willing to share data sets they were keen to ensure that they had extracted as much research value as possible from their datasets before they were willing to share them further. They were also anxious about the potential effort and cost of making data sets shareable and how the skills, tools and safeguards on IPR and personal data would be made available or managed.*

Research Data Consumers *- Data is the life blood of research and every research group is critically dependent on access to data. In the most cases the availability of suitable data is a study in itself as there is no coherent method of searching for relevant data other than through references in publications and trawling the many institutions where repositories are maintained. In 'big science' much investment has been made and the relative narrowness of the major topics means that (volume aside) data access and processing is relatively simple. In arts and humanities and 'small science' the problem is more complex and intractable…"*
As part of its feasibility study, UKRDS lists the key success factors for a national service such as UKRDS. Several of these factors should be mentioned here as well, although some of them are closely related to a national service as a model:

Headline	Detail
Engagement with researchers and suppliers	The main measures of success for a service such as UKRDS will be the willingness of potential service providers to engage in the service delivery process through a UKRDS and the willingness of researchers to seek education in data management and opportunities to deposit their datasets with service providers via the UKRDS route.
Targeted delivery	It is essential for this to work and be measured in a reasonable fashion that availability is controlled and targeted at low risk low cost providers and researchers initially.
Relevant data	It is clear that there is a considerable degree of existing data management that can be adopted unchanged by UKRDS as it is well established and targeted at a highly specialist research community. Much of this is in "big science" such as astronomy, particle physics and the like. It will clearly mitigate against the success of any UKRDS if its advent has any negative impact on such work.
Clear added-value	It is likely that UKRDS will succeed by adding value to the work of arts and humanities, social sciences and small science. In this context there is some evidence that supporting overlap in dataset usage between social sciences and clinical medicine would be of benefit.
Pragmatic build-up	The major constraint in this work will be the need to engage with suitable institutions and national service providers; the four case study institutions and UKDA and ADS are thought to hold out the opportunity of a carefully controlled early implementation.
Embrace, don't threaten, value where it already exists	Any service such as UKRDS must recognise current areas of excellence such as ESDA/UKDA and not detract from them – the model should embrace and enhance them where possible.
Good communications to target community	It remains difficult for national subject centres to reach researchers in institutions. Institutions and their researchers are leading on data creation and use with the subject centres keen to engage them more in the preparation of research data for long-term preservation. Any service such as UKRDS must ensure its role is well communicated and constantly refreshed and ideally help to address the problem faced by all repositories.
Must be built to succeed, not to fail	Key elements are – an affordable, sustainable business model; focus on ease of use, with no heavy metadata overhead; presentation is key – UKRDS must ensure the message gets across that it is a journey, the first phase of which alone is likely to last three years, it is not a quick fix; there are unlikely to be any "quick wins" other than proper representation of UKRDS' intent.
Quality as well as quantity	Data stored in UKRDS must have a quality stamp on it, can be trusted, is secure and made accessible through simple but effective data mining tools.
Availability of resources skilled in data management	Sufficient support/advisory staff must be available who are familiar at discipline-level and costed into the funding models.

Source: *UK Research Data Service: Proposal and Business Plan for the Initial Pathfinder Development Phase* (May 2010).

It is interesting to see that the study does not similarly list the risks. There is one very concise statement about risk: "*The biggest risk is to do nothing*".

4 Summary of findings

This chapter summarises the most important findings of the literature study.

4.1 The context

Looking at the overall picture, we see that the current literature on research data is full of references to a "data avalanche". Science will be more data-driven than ever. Data will in fact become the infrastructure. There will be more cross-disciplinary research.

There are growing concerns about data storage, data access, retrieval of stored data, and preservation of data for re-use in the future. Not everybody knows how these problems should be solved, however, or even where to start solving them. Researchers share some of these concerns in the short term, but often do not worry beyond the end of their research project.

There are major differences in the way the disciplines conduct their research, but they also have various factors in common when it comes to data storage and access. There are technical barriers, for example the use of obsolete software, and non-technical barriers, such as the fear of competition, lack of trust, lack of incentives, and lack of control. The literature shows that these non-technical barriers are more powerful than other impediments.

4.2 The researchers' perspective

Many sources provide evidence of an important distinction between data storage and access *during* the research project phase and data management *after* publication of research results. From the researchers' perspective, this distinction is very important and has a major impact on the way they perceive the issues involved, the proposed solutions, and their attitudes.

4.3 Production phase (during the research project)

If we look at the issues from the researchers' perspective, we note that researchers distinguish between raw data, the recorded facts, and processed data. In most cases, raw data can be seen as "pure" data, although if data is recorded by sensors, some processing has already taken place in the sensor, controlled by the parameters installed by the researcher. Processed data includes attributes added by the researcher.

It should be noted that for the purposes of this study, the "research project phase" may also denote the entire *production phase* of longitudinal data or data that is gathered during an individual's lifetime.

4.3.1 Focus on data protection

In all cases, during the research project or production phase the research data must be stored and easily accessible for the researcher and those to whom he has granted access to his data. Local storage is, for many researchers, the preferred storage option, even if it is not very reliable. Convenience and being in control apparently are more important to researchers. Management of the research data is in the hands of the researcher, and perhaps that is how it should be, since he knows his data like no one else does. Only a data specialist familiar with the research process and methods used by the research group may be capable of providing any real support. The bottom line is that a researcher does not wish to be interrupted in what he wants to do most: his research.

Researchers identify a number of ways in which data management services could help them make their research work more efficient and less time-consuming. Bearing in mind that each discipline has different needs, researchers mention:

- support in digitising non-digital materials that are needed in digital form for the research work;
- support in keeping track of large collections of research data, e.g. collections of digitised originals;
- support in managing data more efficiently;
- sufficient storage space available immediately when needed, preferably without extra costs or red tape;
- support in retrieving stored data;
- support in sharing research data;
- fine-grained control access mechanisms for sharing.

And most important of all: protection of valuable research data from the first stage of the research project until publication of the results. Protection here means:

- protection against data loss by hardware of software failure or human error;
- protection against unauthorised access to or use of the data (reading, printing, changing, deleting);
- ensuring that the data is easily available for re-analysis and processing.

4.3.2 Key success factors for support are:

- Support must be local and discipline-specific!
- Support must be based on sustained and proactive contact with the researcher, and not interfere with his research work. Face-to face interaction is important.
- The benefits of the support offered must be clearly visible and tangible.
- The support should not be offered as a top-down "toolbox", but rather as part of existing good practices that are already in place.
- Support should also take into account that there is a flourishing e-mail culture in most disciplines. Many researchers use e-mail as a basic tool for sharing their work and for communicating with colleagues, including data exchange with trusted parties.
- When it comes to sharing research data, it is important to make a clear distinction between the raw, unprocessed data and the processed data that contains attributes or other elements that cannot easily be shared with others for a variety of reasons.
- Researchers want to be in control of the parameters that handle access to their data.
- It is important to build trust, because researchers perceive handing over their data to another party as giving away their intellectual capital.

4.4 Post-publication phase

Once the research project has ended and the results have been published (the method of publication is irrelevant here), a new phase begins. The researcher is probably already spending a lot of time securing funding for the next research project. In most cases, data from the previous project will stay where it was at the end of the research project – where, more often than not, the storage situation is unreliable and the data is likely to deteriorate over time. Although all who were involved know that data will probably be lost forever, there is no time to take protective measures. Researchers have the feeling that they are not in a position to solve this problem, nor do they tend to accept responsibility for it. Their willingness to take responsibility is not highly developed.

Depending on the specific situation, a copy of the research data may have to be transferred to the publisher, and another copy may need to be sent to a data centre. Documentation and metadata may be required, but again – there is no time to do all this.

Regardless of the location of the data, either local or in a data centre, there is a basic willingness to share data with others, including with people outside the research group. Once publication is a fact, this willingness increases. The need for fine-grained access control remains very important, however. The researcher assumes that local storage of data gives him more control over his data

than remote storage in a data centre. At the same time, he admits that remote storage will probably alleviate some of the burden of data management. In all cases, when the data is transferred to another party, the researcher wants to stay in control of his data.

4.4.1 Barriers

In most cases, researchers become less interested in their research data once the research has been published. There is a growing awareness that there should be more focus, but it seems that this awareness is stronger in circles *outside* the research groups. This is understandable when we look at things from the researchers' perspective, in particular the barriers they mention in the various surveys. Note that these barriers relate mostly to data sharing and data preservation that are beyond the researcher's control.

The biggest barriers mentioned are non-technical in nature and could be described as "human" factors, partly generic, partly cultural, partly local or discipline-based. Examples are:

- fear of competition and, in connection with that fear, a protective attitude towards career, finances, professional status;
- protection of data as intellectual capital;
- fear of misuse of the data;
- complicated legal issues (ownership);
- lack of incentive to overcome these fears: there is no reward for spending extra time on data preservation and data sharing;
- privacy issues, e.g. in the case of medical data;
- no time, no funding to spend extra time on data preservation;
- differences in research methodology leading to incompatibility of data.

There are also technical barriers to reusing data, either real or perceived. Some examples are:

- incompatible data types;
- obsolete data formats, software, hardware or combination of all of these;
- missing or inadequate documentation/metadata.

4.4.2 Focus on trust and control (data sharing)

In the view of researchers, the possibility of re-analysis of research data is the most important driver for data preservation. The barriers to actually doing so are clear. When there is a legal or grant obligation to deposit the data, the non-technical barriers become less relevant, but the technical barriers remain. Researchers need help solving the technical problems listed above. In addition, they need help meeting the requirements of the data centre. The specific needs in this case are:

- help with documentation;
- help with adding the relevant metadata;
- help with structuring the data files according to a standard.

When a researcher sends data to a publisher, we may assume that the data is safely stored. Most publishers have outsourced their data archive to an external partner, however, and the data files are sent to this partner as is. The technical issues are similar and the requirements of the data centre may not be met.

There is some evidence that training and guidelines are welcome, but the disciplines differ in this respect.

5 Conclusions

The initial research question for this study was: what do researchers need when it comes to research data storage and access? In addition, other questions arose: How are researchers currently handling research data storage and access? What problems have they encountered with regard to storage and access, and what are their needs? What projects (in Europe and elsewhere) are dealing with this issue?

Based on the sources that we have studied, our conclusions are the following.

1. There is growing awareness of the need for data storage and data preservation and a growing sense of urgency concerning this issue. All stakeholders (funding agencies, data producers, data consumers, data centres) agree, for various reasons, that something needs to be done to improve the current situation. At the same time, no one solution has been found.

2. There are major differences in the way disciplines conduct their research, but they also have factors in common when it comes to data storage and access. There are technical barriers, for example the use of obsolete software, and non-technical barriers, such as the fear of competition, lack of trust, lack of incentives, and lack of control. The literature shows that these non-technical barriers are more powerful than other impediments.

3. There is an important difference between data storage and access *during* a research project phase and data management *after* the publication of research results. Researchers view this distinction as a very important one, and it has a huge impact on how they perceive the issues involved, the solutions proposed, and their attitudes. Storage and preservation are two distinct issues for researchers. They have expressed a clear need for support in day-to-day storage; on the other hand, they see preservation as a different step, one that falls somewhat outside their immediate scope of interest.

4. During the research project phase, the researcher's focus is on protecting his valuable data. Researchers have expressed their needs in this area, and they do in fact need support, because they do not possess the skills, awareness, or knowledge to improve their day-to-day data management.

5. The literature leads us to conclude that researchers can, indeed, benefit from support services in managing their digital data, but that these services must meet a number of requirements if they are to be successful:
   - Tools and services must be in tune with researchers' workflows, which are often discipline-specific (and sometimes even project-specific).
   - Researchers resist top-down and/or mandatory schemes.
   - Researchers favour a "cafeteria" model whereby they can pick and choose from a set of services.
   - Tools and services must be easy to use.
   - Researchers must be in control of what happens to their data, who has access to it, and under which conditions. Consequently, they want to be sure that whoever is dealing with their data (data centre, library, etc.) will respect their interests.
   - Researchers expect tools and services to support their day-to-day work within the research project, and long-term/public requirements must be subordinate to that interest.
   - The benefits of the support must clearly visible – not in three years' time, but now.
   - Support must be local, hands-on, and available when needed.

6. Preservation of research data after the publication phase is possible only when storage during the research project phase has been well managed. In fact, it makes sense to invest in better data management during the research phase because doing so will improve data preservation once the research phase has ended.

7. Most researchers are unwilling to automatically accept responsibility for preserving their data after publication. The need for fine-grained access control remains very important. Researchers assume that local storage gives them more control over their data than remote storage in a data centre. At the same time, they admit that remote storage will probably alleviate some of the burden of data management. In all cases, when the data is transferred to another party, researchers wish to remain in control of their data.

8. There are many other barriers to data preservation and data sharing. These are mostly non-technical in nature and are therefore more difficult to address. Once researchers have seen the results of better (from their perspective) data management during the research phase, however, and once they realise that the support they have accepted is actually beneficial for their work, we expect some of the non-technical barriers to become less important.

Annex 1 – List of sources

Beagrie, 2009[5]
Research data preservation and access: the views of researchers. Beagrie N., Beagrie R. and Rowlands I., Ariadne (July 2009).
http://www.ariadne.ac.uk/issue60/beagrie-et-al/

E-IRG, 2009
E-IRG Report on data management. Data Management Task Force (December 2009).
http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

van der Graaf, 2010
Organisatorische aspecten duurzame opslag en beschikbaarstelling onderzoeksdata. Maurits van der Graaf, Stichting SURF (November 2010).
http://www.surffoundation.nl/nl/publicaties/Pages/Organisatorischeaspectenduurzameopslagenbeschikbaarstellingonderzoeksdata.aspx

Henty, 2008
Investigating Data Management Practices in Australian Universities. Margaret Henty et al., APSR (July 2008).
www.apsr.edu.au/orca/investigating_data_management.pdf

Kuula, 2008
Open access to and reuse of research data: the state of the art in Finland (2008).
http://www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs07_OECD_en.pdf

OCLC, 2010
A Slice of Research Life: Information Support for Research in the United States. Susan Kroll and Rick Forsman, OCLC (June 2010).
http://www.oclc.org/research/news/2010-06-16.htm

PARSE, 2009
PARSE.Insight: INSIGHT into Issues of Permanent Access to the Records of Science in Europe (2009).
http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

RIN, 2010[6]
Research Support Services in UK Universities. A Research Information Network Report (October 2010).
http://www.rin.ac.uk/news/what-research-support-services-do-researchers-need

RIN/BL, 2010[7]
Patterns of information use and exchange: case studies of researchers in the life sciences. A report by the Research Information Network and the British Library (November 2009).
http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studie

Rombouts, 2009
Waardevolle Data en Diensten. Rombouts J. et al., Eindrapport (2009).
http://3tu.typo3.3xo.eu/fileadmin/documenten/Eindrapportage_WDenD_v10_170709.pdf

[5] This article based on the same survey as the UKRDS study.
[6] This is based on research carried out at Leicester, University College London, York, and Warwick.
[7] The case studies were collected among seven research teams from a wide range of disciplines, from botany to clinical neuroscience.

Tjalsma, 2010
Selection of Research Data: Guidelines for appraising and selecting research data. Heiko Tjalsma and Jeroen Rombouts, SURFfoundation (July 2010).
http://www.surffoundation.nl/nl/publicaties/Pages/StudieSelectionofResearchData.aspx

UKRDS, 2010[8]
UK Research Data Service: Proposal and Business Plan for the Initial Pathfinder Development Phase (May 2010).
http://www.ukrds.ac.uk/resources/download/id/47

Verhaar, 2010
Data Curation in Arts and Media Research. Peter Verhaar et al., SURF (2010).
http://www.surffoundation.nl/nl/publicaties/Pages/StudieDataCurationinArtsandMediaResearch.aspx

Waaijers, 2010
Kwaliteit van onderzoeksdata, een operationele benadering. Leo Waaijers, Maurits van de Graaf. Preprint version (December 2010).
http://www.surffoundation.nl/nl/publicaties/Pages/Verkennendonderzoek.aspx

WAVE, 2010
Riding the Wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission (October 2010).
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[8] The universities of Bristol, Leeds, Leicester and Oxford participated in this report.