Opportunities for Data Exchange

REPORT ON INTEGRATION OF DATA AND PUBLICATIONS

October 17th, 2011

Susan Reilly [a, *], *Wouter Schallier* [a], *Sabine Schrimpf* [b], *Eefke Smit* [c]
Max Wilkinson [d]

[a] *LIBER – Association of European Research Libraries, Koninklijke Bibliotheek, National Library Of The Netherlands. Po Box 90407. 2509 Lk The Hague. The Netherlands*
[b] *Deutsche Nationalbibliothek Informationstechnik, Adickesallee 1. D-60322 Frankfurt am Main. Germany*
[c] *The International Association of STM Publishers, Prama House, 267 Banbury Road. Oxford OX2 7HT. United Kingdom*
[d] *The British Library, 96 Euston Road. LONDON NW1 2DB. United Kingdom*
* *Corresponding author: Susan.Reilly@KB.nl*

Abstract

Scholarly communication is the foundation of modern research where empirical evidence is interpreted and communicated as published hypothesis driven research. Many current and recent reports highlight the impact of advancing technology on modern research and consequences this has on scholarly communication. As part of the ODE project this report sought to coalesce current though and opinions from numerous and diverse sources to reveal opportunities for supporting a more connected and integrated scholarly record. Four perspectives were considered, those of the Researcher who generates or reuses primary data, Publishers who provide the mechanisms to communicate research activities and Libraries & Data enters who maintain and preserve the evidence that underpins scholarly communication and the published record. This report finds the landscape fragmented and complex where competing interests can sometimes confuse and confound requirements, needs and expectations. Equally the report identifies clear opportunity for all stakeholders to directly enable a more joined up and vital scholarly record of modern research.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report sets out to identify examples of integration between datasets and publications. Findings from existing studies carried out by PARSE.Insight, RIN, SURF and various recent publications are synthesized and examined in relation to three distinct disciplinary groups in order to identify opportunities in the integration of data. These groups are Researchers, Publishers and Libraries/Data centres. Opportunities identified for each group have been scoped against seven criteria:

1. Availability
2. Findability
3. Interpretability
4. Reusability
5. Citability
6. Curation
7. Preservation

Opportunities to improve the linking of data and publications have been identified for each stakeholder group and mapped against each of the criteria in tables at the end of this summary.

Based on an examination of the available research and literature, incentives and barriers relating to data exchange are identified for each disciplinary group.

The content of a draft of this report formed the basis of a workshop in June 2011 with professionals from research libraries. The workshop served to validate this opportunities and issues identified in this report.

From a researcher perspective, the value of data is that of a first class research object which represents the basis of their research. Researchers discover and use data and analyses from others to formulate new and testable hypothesis before extending the evidence base with empirical data. The implications of first class research objects are that they require preservation, recognition, validation, curation and dissemination which then improve their availability, findability, interpretability and re-usability. Researchers perceive and enforce their creator right over the data, choose when and with whom they share it and wish to maintain this control. This need for control is based on perceived legal barriers and misuse, or absence of a trust network common in other forms of scholarly communication; it may be a mixture of both. Researchers want somewhere safe to put their data while maintaining control in order to avoid legal redress and professional misuse, but expect some central organisational structure to pay for these infrastructures. They recognise that many lack sufficient skills to manage their data appropriately, but, importantly, are enthusiastic to change this situation. Researchers see the benefit in joining publications with data in a more formal and agreed convention, but there must be a recognition and credit mechanism for this. They accept this joining as good professional practice and agree that data supporting traditional publication should be available with the publication. Technology can reduce the latency to joining data to publications but there is a lack of common best practice conventions for scholarly publications. Distilled into statements, our desk research has revealed five abstract researcher requirements for integrating data and publication.

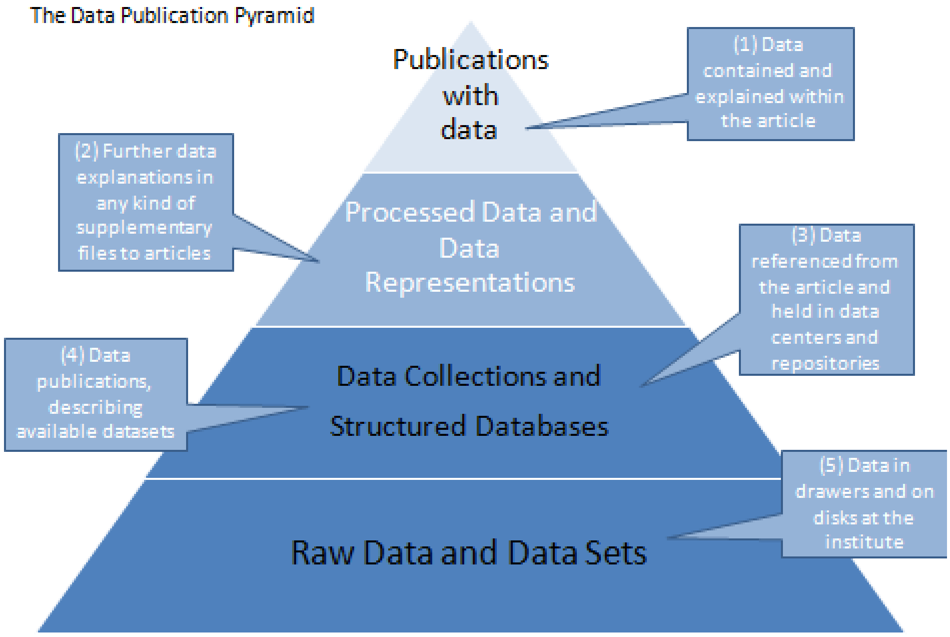
1. Researchers need somewhere to put data and make it safe for reuse
2. Researchers need to control its sharing and access
3. Researchers need the ability to integrate data and publication
4. Researchers need to get credit for data as a first class research object
5. Researchers need someone to pay for the costs of data availability and re-use

Publishers are beginning to embrace the opportunity to integrate data with publications but barriers to the sustainability of this practice include the sheer volume of data, the huge variety of data formats and a question mark over exactly what data should be made available within, be made supplemental to or be linked with the publication. Also the quality of the data and attached metadata may not be consistent, lacking peer review, or is not being made transparent.

The relationship between data and publications can be illustrated with a modified version of Jim Gray's e-science pyramid which in this report is presented as the Data Publication Pyramid, see the Graph 1 below. As we descend the pyramid the exclusive relationship between data and publication diminishes. At the top, for example, the journal (and author/researcher) takes full responsibility for the publication including the aggregated data embedded in it and the way it is presented. For data published in the second layer, as supplementary files to articles, the link to the published Record of Science remains strong, but it is not always clear at what level the data is curated and preserved and if the criteria for discoverability and re-usability are met. At the Data Collections and Structured Database layer, the publication includes a citation and links to the data, but the data resides in and is the responsibility of a separate repository. The publication of data becomes collaborative.

At the bottom layer of the pyramid, most datasets remain unpublished and hence unfindable and not re-usable.

As Jim Gray already made clear, the data published now within or with publications, is only the tip of the data iceberg.


The Data Publication Pyramid
Publications with data
Processed Data and Data Representations
Data Collections and Structured Databases
Raw Data and Data Sets
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute

Graph 1: The Data Publication Pyramid, developed on the basis of the Jim Gray pyramid, to express the different manifestation forms that research data can have in the publication process. See Chapter 1 for a full explanation.

As more publishers respond to increasing author demand to making research data available they are focused on:

1. establishing cross publisher best practice to make data available and retrievable in a persistent way
2. collaboration with publicly endorsed community archives to make data and publications interlinkable
3. presenting data in more sophisticated formats to increase reuse

Libraries and data centres have overlapping and complementary roles in terms of data integration. Barriers to integration of data include a lack of policies to address the concerns of researchers when it comes to making their data available, the lack of uniformity in data preservation and curation strategies and practices.

New publishing models linking data and articles require that libraries and data centres need to address particular concerns:

1. preservation and persistence of data to ensure continued access to linked data
2. making data findable and reusable though the use of metadata and integration into retrieval services
3. working closely with researchers to encourage data sharing and best practice in data management

In general, the need for action has been recognized in the library and data centre community. Noteworthy initiatives like the ones selected for description in this report

(DataCite, PANGAEA, DRYAD, and Dataverse) illustrate this as well as how libraries and data centres support data integration. However, the degree of preparedness to take on the challenge varies between disciplines and between individual institutions.

It is important that libraries and data centres act in conformance with the requirements of the research community, which they serve. They also need to be involved in the research process from the very beginning in order to ensure high data quality, which facilitates retrieval, usability, and preservation.

An examination of the research and noteworthy initiatives highlights that opportunities exist across the three disciplines. These opportunities exist particularly in the areas of availability, findability, interpretability, and re-usability.

Data Issue:	**Researchers opportunities (Chapter 2):**
Availability	Researchers demand their data be treated as first class research objects Researchers loosen control over data Define roles of responsibility and control
Findability	Agree convention to propose to publishers regarding data citation Use of persistent identifiers such as DOI's Ensure common metadata and citation practices
Interpretability	Recognize that data require metadata and work towards community best practice in metadata development
Re-usability	Be concerned about the long term ability for secondary use and consider or seek out responsible preservation actions
Citability	Agree a convention for data citation Follow metadata standards for datasets Use of persistent identifiers such as DOI's
Curation	Develop sustainable and realistic data management plans Collaboration with public data archives
Preservation	Develop sustainable realistic preservation plans Active engagement with public data archives

Table 1: Data Opportunities for Researchers

Data Issue:	**Publishers opportunities (Chapter 3):**
Availability	Articles with data provide richer content and higher usage Impose stricter editorial policies about availability of underlying data which is in line with general funder's trends Ensure data is stored in a safe place, preferably a public repository Be transparent about curation and preservation of submitted data
Findability	Ensure bi-directional links between data and publications Ensure common citation practices
Interpretability	Provide services around data such as viewer apps for underlying data from within the article or interactive graphs, tables and images Data Publications
Re-usability	Interactive data from within articles Links to the relevant datasets, not just to the database Data Publications
Citability	Establish uniform data citation standards Follow metadata standards for datasets Use of persistent identifiers such as DOI's Data Publications
Curation	Transparency about curation of submitted data Collaboration with public data archives
Preservation	Transparency about preservation of submitted data Collaboration with public data archives

Table 2: Data Opportunities for Publishers

Data Issue:	Libraries and data centres opportunities (Chapter 4):
Availability	Lower barriers to researchers to make their data available. Integrate data sets into retrieval services.
Findability	Support of persistent identifiers. Engage in developing common metadescription schemas and common citation practices. Promote use of common standards and tools among researchers
Interpretability	Support crosslinks between publications and datasets. Provide and help researchers understand metadescriptions of datasets. Establish and maintain knowledge base about data and their context.
Re-usability	Curate and preserve datasets. Archive software needed for re-analysis of data. Be transparent about conditions under which data sets can be re-used (expert knowledge needed, software needed).
Citability	Engage in establishing uniform data citation standards. Support and promote persistent identifiers.
Curation/ Preservation	Transparency about curation of submitted data. Promote good data management practice. Collaborate with data creators Instruct researchers on discipline specific best practices in data creation (preservation formats, documentation of experiment,…)

Table 3: Data Opportunities for Libraries and Researchers

0. INTRODUCTION

Science is changing. The massive volume and variety of data pouring out of publicly funded science are transforming the face of research. These data belong to everyone. If we manage these precious resources properly, we may tackle the Grand Challenges of our times – even as budgets become more restricted. It is easy to take for granted that data in the public domain will be protected and remain both available and accessible. Researchers, publishers, policymakers and funders – among many others – have started to appreciate that a robust, sustainably funded infrastructure is absolutely necessary to protect the hard-earned fruits of publicly funded research.

Opportunities for Data Exchange (ODE)[1], a project funded by the European Commission (FP7) with Grant Agreement number 261530, is gathering evidence to support and promote data sharing, re-use and preservation. ODE partners are members of the Alliance for Permanent Access (APA) and represent stakeholders with significant influence within their communities. ODE is identifying, collating, interpreting and delivering evidence for emerging best practice in sharing, re-using, safeguarding and citing data. ODE is also documenting drivers of change, and barriers to progress in this important area.

The transition from science to e-Science is happening: a data deluge emerges from publicly-funded research facilities; a massive investment of public funds into the potential answer to thegrand challenges of our times. This potential can only be realised by adding an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-Infrastructures. The importance of this layer, on top of emerging connectivity and computational layers, has not yet been addressed coherently at ERA or global level. All stakeholders in the scientific process must be involved in its design this layer: policy makers, funders, infrastructure operators, data centres, data providers and users, libraries and publishers. They need evidence to base their decisions and shape the design of this layer.

The ODE partners are:

- European Organization for Nuclear Research (CERN)
- Alliance for Permanent Access (APA)
- CSC – IT Centre for Science
- Helmholtz Association
- Science and Technology Facilities Council (STFC)
- The British Library
- Deutsche Nationalbibliothek (DNB)
- International Association of STM Publishers (STM)
- Stichting LIBER Foundation)

All of them are members of the Alliance for Permanent, which collectively represent all these stakeholder groups and have a significant sphere of influence within those communities. The project will identify, collate, interpret and deliver evidence of

[1] *www.ode-project.eu*

emerging best practices in sharing, re-using, preserving and citing data, the drivers for these changes and barriers impeding progress, in forms suited to each audience. ODE will:

- Enable operators, funders, designers and users of national and pan-european e-Infrastructures to compare their vision and explore shared opportunities
- Provide projections of potential data re-use within research and educational communities in and beyond the ERA, their needs and differences
- Demonstrate and improve understanding of best practices in the design of e-Infrastructures leading to more coherent national policies
- Document success stories in data sharing, visionary policies to enable data re-use, and the needs and opportunities for interoperability of data layers to fully enable e-Science
- Make that information available in readiness for FP8

This report sought to coalesce current though and opinions from numerous and diverse sources to reveal opportunities for supporting a more connected and integrated scholarly record. Four perspectives were considered, those of the Researcher who generates or reuses primary data, Publishers who provide the mechanisms to communicate research activities and Libraries & Data enters who maintain and preserve the evidence that underpins scholarly communication and the published record.

1. INTEGRATION OF DATA AND PUBLICATIONS: GENERAL

1.1. Introduction and summary

The web, the cloud and computational capabilities in general provide an ever growing infrastructure for scholarly communication that makes it much easier for researchers to share their research data with others. At the same time, and often driven by the same factors, nearly all scientific disciplines have a computational stream , generating ever more research data. We seem to be at the verge of a Data Deluge as a recent EU report also concluded. [2]

We know from previous research, carried out for the project PARSE.Insight[3] , that around 60 % of researchers would like to use the research data of others. A similar reluctance for sharing as has been apparent in the interviews of WP3 of this project ODE that built on PARSE.Insight, where over 40 % of researchers state to have real problems in sharing their own data. This is further elaborated in Chapter 2: Researcher's Perspective. In this sense, we may coin a new 60-40 rule: 60 % likes to get data from others but 40% have problems to give their own.

Andrew Treloar, Director of the Australian National Data Service gave a talk on 28 March 2011 at a JISC workshop in Birmingham on the management of research data[4] and distinguished several basic problems for research data. In the handling of research data, he described in a cascading way, how research data are often:

1. Unavailable, and if at all available:
2. Unfindable, and if available AND findable:
3. Uninterpretable.

And even if all these 3 obstacles have been overcome, the research data found may still prove to be:

4. Not re-usable.

In this report, created and delivered in the context of project ODE (Opportunities for Data Exchange) we investigate how integration of data and publications can help solve these 4 obstacles. The questions we try to answer in this report are: How do research data enter the stage of scholarly communication and what are the present practices and policies? Where can we find important improvements for the accessibility and re-usability of research data? What roles and responsibilities may we expect for different stakeholders in the scholarly information chain?

[2] John Wood, EU, 2010, Riding the Wave: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] Survey Report PARSE.Insight : http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[4] Andrew Treloar at JSC workshops see http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmevents/mrdinternationalworkshop.aspx

We have used 3 different perspectives to shed light on this:

1. A researchers' perspective, see Chapter 2
2. A publishers' perspective, see Chapter 3
3. A libraries' and data centres' perspective, see Chapter 4

The final chapter summarizes the outcomes of a workshop held with librarians in Barcelona on June 29, 2011[5] as part of the LIBER 2011 conference with a view to:

4. Report Epilogue: Mapping the Road ahead?, see Chapter 5.

In this general introduction, we summarize findings from previous research, such as PARSE.Insight and from newer desk research focused on the way research data are connected with publications. The hard data from the PARSE.Insight study were of high value for this project for 2 reasons: they span the responses from different stakeholder groups (researchers, publishers, librarians and data managers) and the survey was truly international and truly multidisciplinary. As the PARSE.Insight report points out, the representativity of the large spectrum of scientific disciplines was well covered, as well as the spread of respondents over different continents (with some dominance for US and Europe). This makes it one of the few studies that overarches local or mono-disciplinary inventories for current practices on the sharing of data.

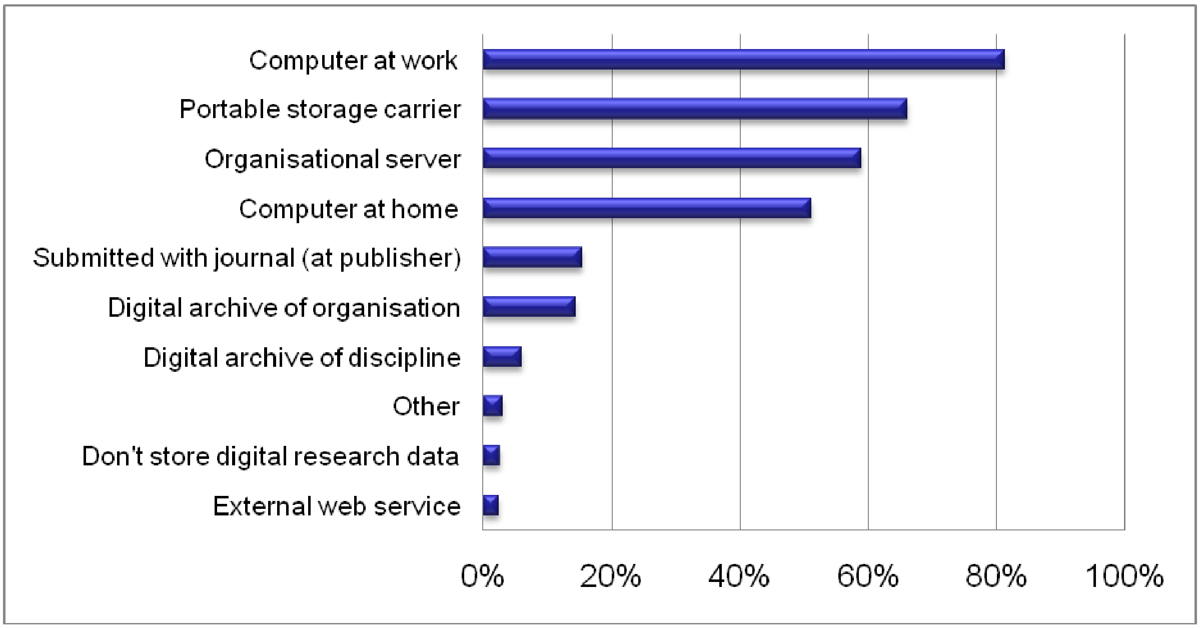
The value that this study attempts to add on the previous PARSE.Insight report by re-use of the data gathered there, is by means of a re-analysis of the survey responses from the perspective of data-sharing. We also included much unused data from PARSE.Insight that proved very relevant for this study.

1.2. Current Practice – some numbers relating research data with publications

In PARSE.Insight, the 2008 survey included a question: Where do you currently store your research data ? From the responses is clear that approximately 18 % of researchers submit data with the manuscripts of their publications, this is a slightly higher figure (but not much) than the number who deposit in archives, see the following diagram from Parse.

[5] Barcelona workshop: http://www.libereurope.eu/news/liber-annual-conference-and-first-european-project-workshops

Where do you currently store your research data ? (multiple answers possible)



Graph 2: Source: PARSE.Insight[2] survey, held among researchers internationally, N = 1202 researchers

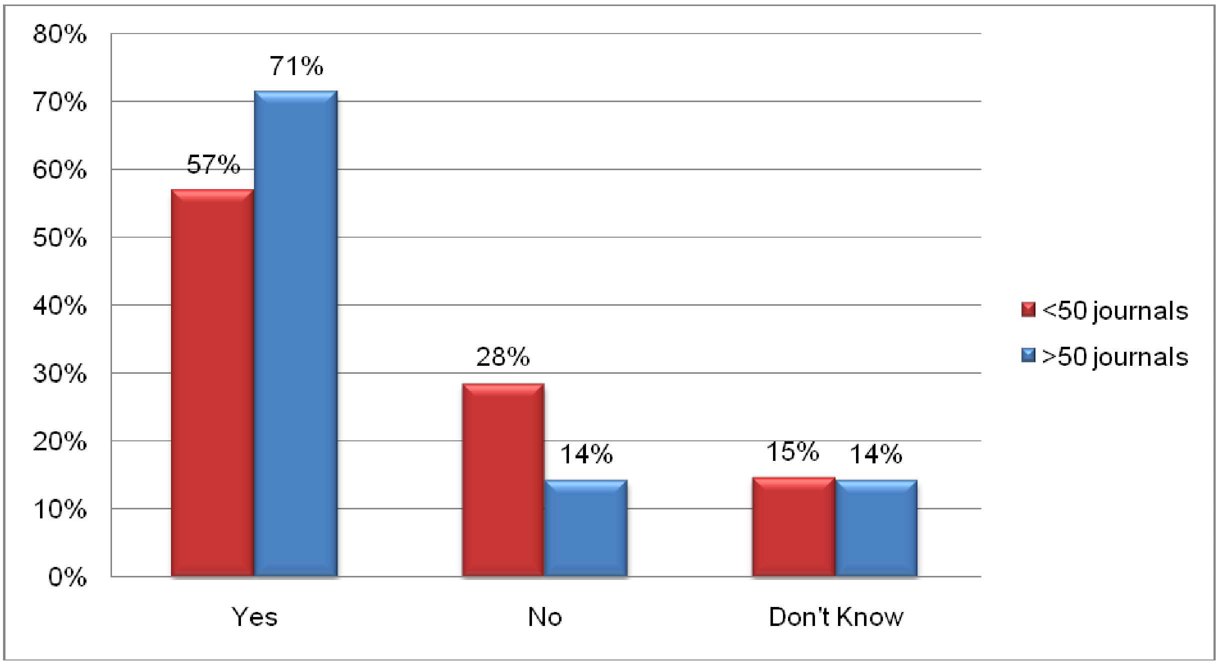
But this number is now growing fast, as the following evidence shows from a recent study into Medical Journals[6]. In that study it was found that for a sample of 28 journals in medicine, sampled from 138 high-impact medical journals, the number of articles carrying supplementary data files roughly doubled every 2 years and is now over 25 %. Compared to 2003, when 32 % of the journals provided the possibility to add web-based supplements to an article, this percentage had grown to 50 % in 2005, and to 64 % in 2009. The number of articles offering online supplements, increased in that same period for these journals from 7% to 14% to the already mentioned 25%, respectively. There was also an increase in the percentage of journals for which at least 20% of articles have online-only supplements (4% in 2003 to 11% in 2005 to 32% in 2009).

A marked increase in the number of video supplements was observed, while the largest share of the supplementary files is for data represented in supplementary graphs and tables. The number of articles with supplementary tables doubled every 2 years (10 to 22 to 55 to 100 from 2003 to 2009), as did the number of supplementary tables (29 to 57 to 149 to 317, the last two numbers referring to 2007 and 2009, respectively).

From the PARSE.Insight2 survey, we also know that > 90 % of journals accepts supplementary material . The graph here below presents the responses from different publishers (small versus large). If the number of journals is factored in, it appears that >90 % accepts supplementary files containing research data.

[6] See as source: http://www.annemergmed.com/article/S0196-0644(10)01648-3/fulltext#sec3

Can authors submit their <u>underlying digital research data</u> with their publication to you?



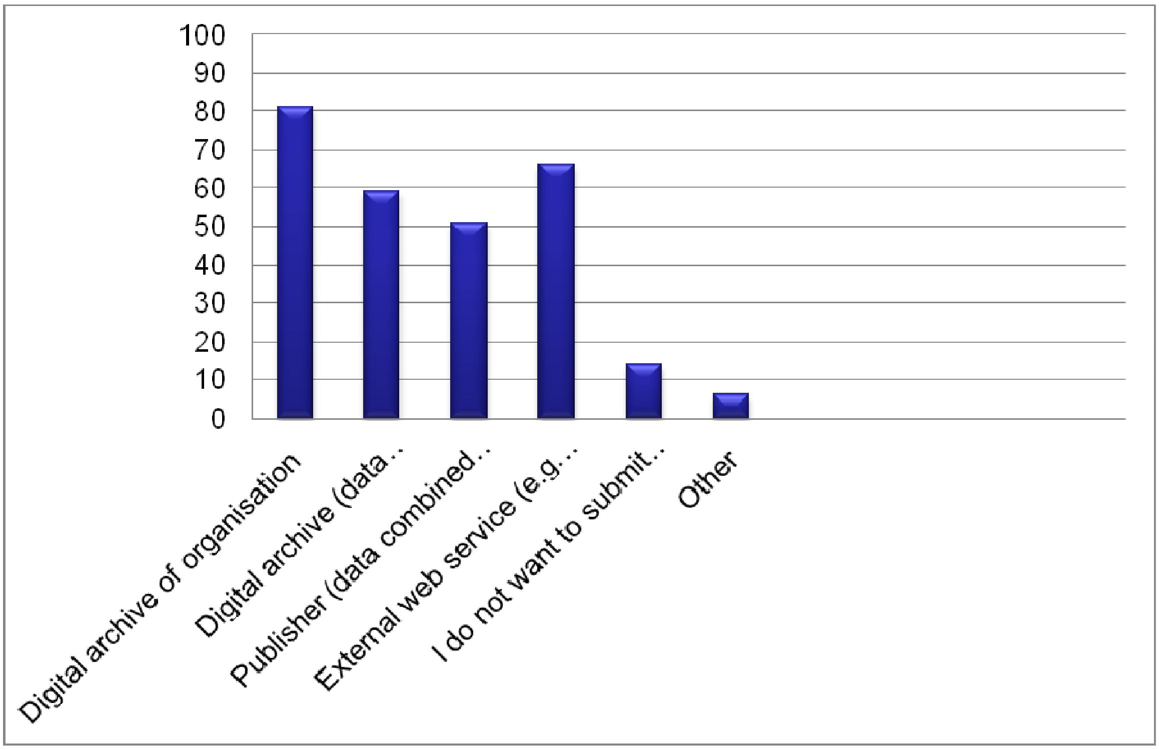
Graph 3, source PARSE.Insight[2], N= 134 publishers

1.3. Why the relationship between data and publications is so important

Whereas the data presented in the study quoted in the previous paragraph5 show the substantial amount of supplementary data added to journal articles, as well as its high growth rates, there is every reason to expect even more. The survey of PARSE.Insight[2] enquired where researchers would like to submit their research data and the responses gave the following picture (graph 4):

Next to the most popular category of Data Archives (of their organisation: 81 %, in their subject field: 60 %), more than half of the researchers (51 %) would like to submit their data to publishers. Of particular interest here is the marked difference between the most popular categories in terms of desired destinations, Graph 4, when compared to the actual destinations, Graph 2. Archives and publishers are the most favored, but remarkably the least used nowadays: in current practice, institutional archives score below 20 % , subject archives below 10%, publishers score a little closer to but still below 20 %.

Where would you be willing to submit your research data? (multiple answers)

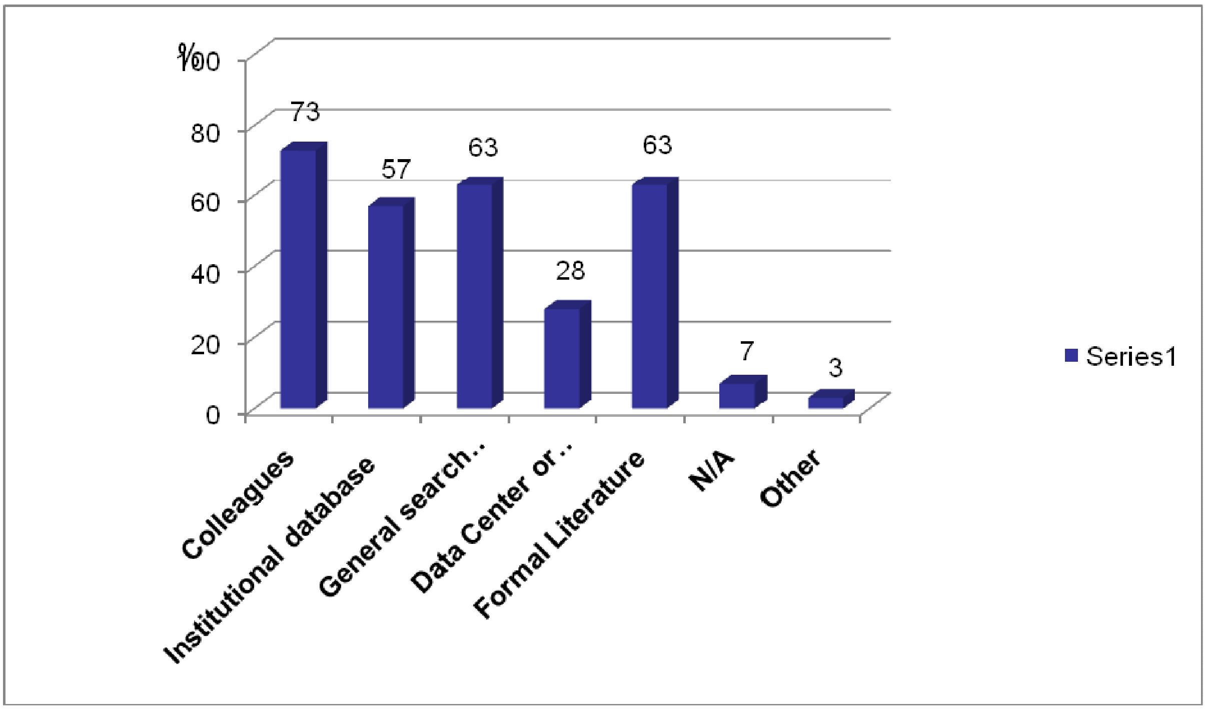


:

Graph 4, Source: PARSE.Insight[2]: > 50 % of researchers would like to submit their data to journals, N= 1202 researchers

The reason why researchers would like to submit their research data together with an article, probably relates to the way they find datasets. Again from PARSE.Insight2 we know that 63 % of researchers like to go to the formal literature to find and discover the existence of data (see also Graph 5). This option is ranked second, immediately after colleagues as a source (73 %) and equal to the use of general search engines (63 %):

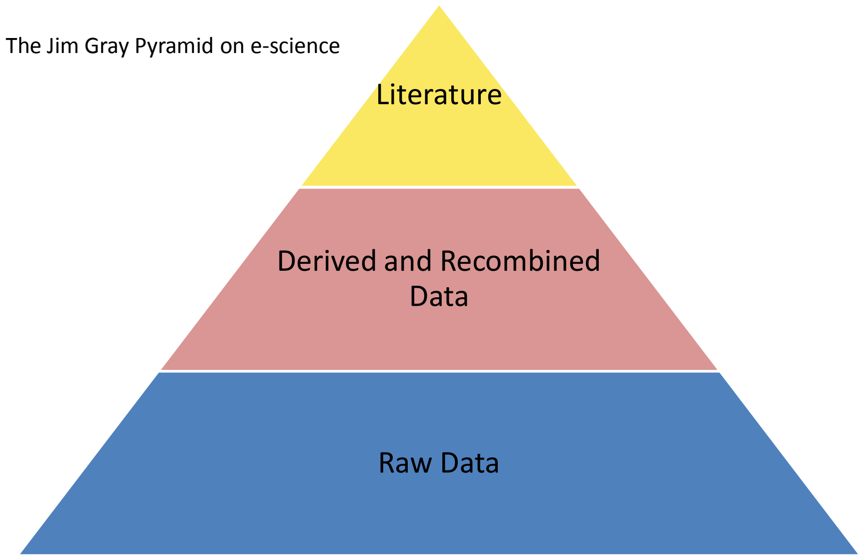
Where do you locate and access digital research data ? (multiple answers)



Graph 5, Source: PARSE.Insight [2], N=1202 researchers

1.4. The Data Publication Pyramid

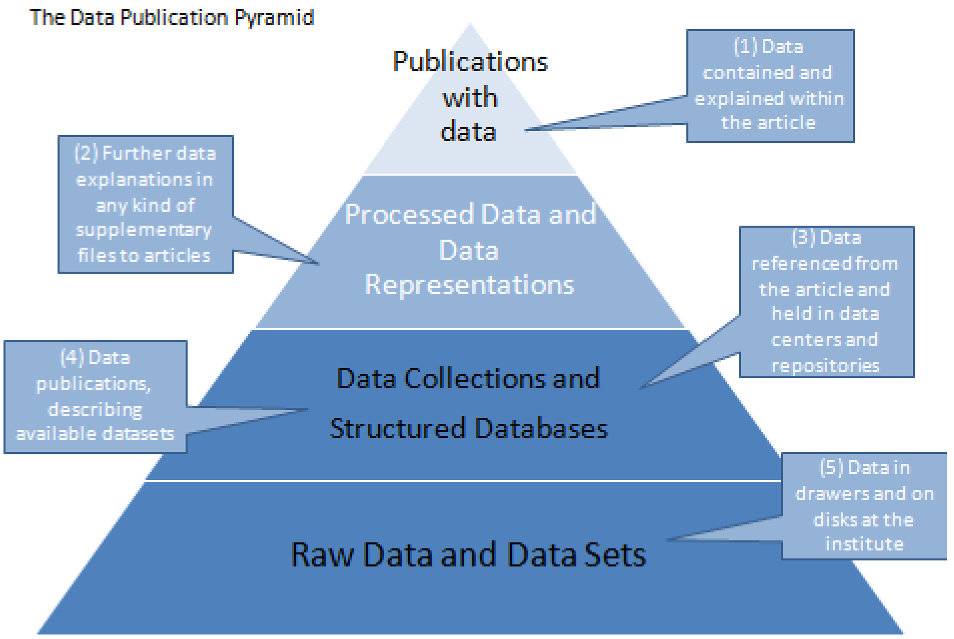
Microsoft researcher Jim Gray is an often quoted source on the way literature and data would become more interrelated. He foresaw ways (graph 5) in which the e-environment would make it so much easier to move from the literature to data and back:


The Jim Gray Pyramid on e-science
Literature
Derived and Recombined Data
Raw Data

Graph 6, Source Jim Gray on e-science[7]: Publications are only the tip of the iceberg

For the purpose of this report, we endeavor to adapt the Jim Gray Pyramid slightly and introduce our own so-called **Data Publication Pyramid**. Four years after Jim Gray expressed his thoughts, we see a new wave of practices emerge where the literature is in fact integrating literature and data or at least making its best attempts. This Data Publication Pyramid (graph 1) aims to show the different manifestations that data can undergo when published within or in the context of publications, or even when not published at all (but remaining in drawers and on disks of the institute):

[7] http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf


The Data Publication Pyramid
Publications with data
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
Processed Data and Data Representations
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
Data Collections and Structured Databases
(5) Data in drawers and on disks at the institute
Raw Data and Data Sets

Graph 1 repeated, The Data Publication Pyramid, based on the Jim Gray Pyramid.

We wish to emphasize at this point that this pyramid does not describe which stages of manifestation research data can go through in their evolution towards reusable data products. The main purpose of this pyramid is to explain the different manifestations research data can have in the context of their availability within, with, supplementary to or referenced from an official scholarly article as the main manifestation of the record of science.

In Chapter 3 (Integration of Data and Publications), in Chapter 4 (Data Centre and Library perspective) and Chapter 5 (Next Steps, the Road ahead) this Data Publication Pyramid is explained further and used to distinguish different best practices as they currently occur.

In short, we can summarize the benefits of good integration of articles and research data as follows, along the key issues raised by Andrew Treloar3 and tabulate obstacles and the solutions provided by integrating Data and Publication (Table 4)

- publications help the data to be better discoverable
- publications help the data to be better interpretable
- publications provide the author better credits for the data
- and reversely: the data add depth to the article and facilitate better understanding.

Obstacle	Integration of Data and Publication helps to
Data are unavailable	Indicate availability of data
Data are unfindable	Links from the publication help locate datasets
Data are uninterpretable	Publications will describe and explain datasets, links from data to publications help interpretability of datasets
Data are not re-usable	Description in article can improve usability or the article can provide access to a re-usable set, perhaps even offer an API to datasets ?

Table 4: Data Obstacles

As will be explained in the following chapters, there are 3 additional necessary conditions that need to be fulfilled for good accessibility and re-usability of research data. They are:

1. Citability
2. Curation
3. Preservation

These elements are also addressed in-depth in the following chapters and used as criteria for drivers and potential opportunities for different players in the scholarly communication landscape. Included in the following chapters are case studies for laudable initiatives, reported unsolved issues, and opinions and desires as expressed to the project team in email exchanges and interviews from the perspective of researchers, publishers and libraries/data centres.

2. RESEARCHER PERSPECTIVE OF DATA/PUBLISHING INTEGRATION:

2.1. Researchers are the source of data

Researchers are able to generate more data than ever before. This has partially been driven by technological advances that increase the accuracy, sensitivity and multiplicity of empirical data collection across disciplines8. But equally, researchers have been able to track, aggregate, abstract, transform and generally re-purpose existing data to drive forward data driven research. Facilities and infrastructures that increase our understanding of sub-atomic particles or extend our reach into the universe have lead to way in generating data; they have been quickly followed by social, environmental and biomedical disciplines that capture complex and uncertain data for modelling biological processes, social dynamics and environmental forecasting. Data collected in vivo or in situ and modelled in silico reveal promising new ways in understanding and intervening in all manner of human behaviour for benefit9'10. All these data are a fundamental component of scholarly communication and are the evidence that underpin scholarly publication. Their value extends beyond original use and many represent substantial, non-reproducible and valuable intellectual assets to many stakeholders1112. Our ability to manage and maintain such digital data have not always kept pace with our ability to generate it and presents a risk to modern research; we are in danger of losing the ability to link the evidence base that support scholarly publication and as a consequence break the cycle of scholarly communication.

From a researchers perspective the value of data is that of a first class research object which represents the basis of their research. Researchers discover and use data from others and analyse them to formulate new and testable hypothesis before extending the evidence base with empirical data. The implications of first class research objects are that they require preservation, recognition, validation, curation and dissemination; in doing so they become more available, discoverable, interpretable, re-usable and citable. This section of our report will review contemporary evidence of current practice from a researcher's perspective and investigate any need for change. From these needs we will

[8] Hanson B, Sugden A, Alberts B (2011). Making data maximally available. Science. 331(6018):649

9 Editorial: Crowdsourcing human mutations. Nature Genetics 2011, 43(4):279

10 Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, Maglott D, Singleton BK, Anstee DJ, Basak AN, Clark B, Costa FC, Faustino P, Fedosyuk H, Felice AE, Francina A, Galanello R, Gallivan MV, Georgitsi M, Gibbons RJ, Giordano PC, Harteveld CL, Hoyer JD, Jarvis M, Joly P, Kanavakis E, Kollia P, Menzel S, Miller W, Moradkhani K, Old J, Papachatzopoulou A, Papadakis MN, Papadopoulos P, Pavlovic S, et al.(2011). Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach.Nature Genetics 43(4):295-301

11 Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, van Mulligen E, Singh B, Hooft R, Roos M, Hammond J, Kiesel B, Giardine B, Velterop J, Groth P, Schultes E, (2011). The value of data. Nature Genetics 43(4):281-3

12 Curry A (2011). Rescue of old data offers lesson for particle physicists. Science. 331(6018):694-5

seek opportunities in the form of enablers, drivers and incentives to support such change.

2.2. What is the current practice

Data has always been part of scholarly communication though increasingly scholarly publication has been unable to maintain the join to this evidence base[13]. Integrating data and publications implies an outcome of data sharing. Thus a researcher perspective is primarily concerned with data sharing; the integration of scholarly communication is essentially an enabler of this. In order to understand why data and publications risk moving apart it is essential to understand what causes such divergence and whether these are barriers we can influence.

Illustrated in Graph 4 in the Introduction Chapter, the PARSE.Insight[14] survey provided compelling and insightful evidence regarding the perspective of researchers on current practice in data management and sharing. With support from additional and contemporary reports,[15], and publications a strong picture of researcher attitudes to data management and requirements begins to emerge.

> ***Researchers find data in the same way as other information (via literature, colleagues) and IT has added to these rather than supplanted any one common practice activity.***
>
> The methods used to disseminate and validate research seems tied to common research practices where colleagues and formal literature predominate. Search engines and institutional databases suggest that technology has made this easier than their analogue counterparts of catalogues and libraries. This is in agreement with the report of the UK's Research Information Network (RIN) in which technology was used to increase efficiency in behaviour, in the life sciences, rather than supplant it[16]. Data archives offer a significant source of discovery but it is difficult to determine if this was in addition to, or independent of the more common discovery through literature and colleagues.

[13] Aalbersberg, IJ, Kähler, O. (2011)Supporting Science through the Interoperability of Data and Articles , D-Lib. Volume 17, Number ½ doi:10.1045/january2011-aalbersberg

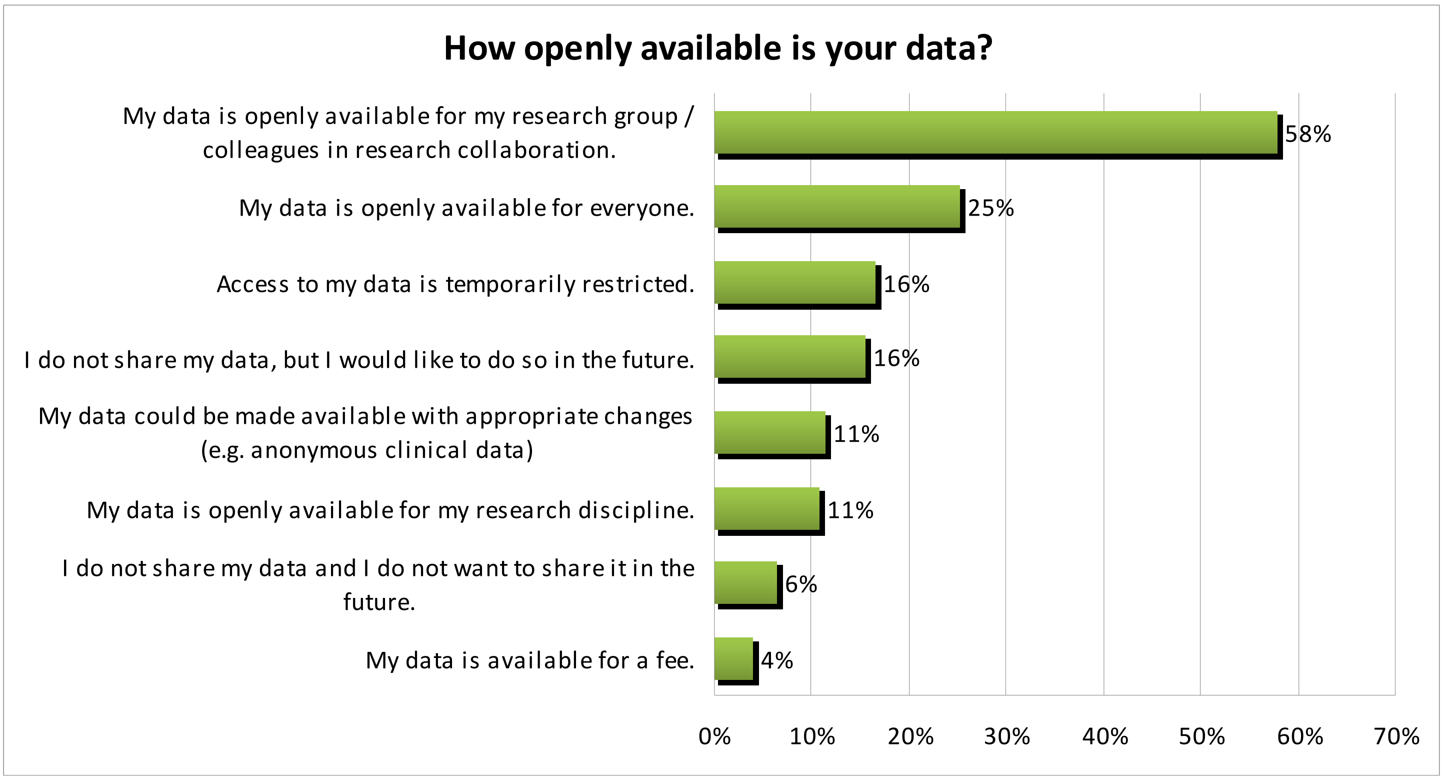
[14] http://www.parse-insight.eu/index.php

[15] If you build it they will come. How Researchers perceive and use web2.0. A Reseatrch Information Report, July 2010.

[16] Patterns of information use and exchange. Case studies of researchers in the life sciences. Research Information Network, November 2009.

Researchers want control over the extent and manner of sharing their data, consistent with the concept of data as a professional asset.

This particular finding of data largely being shared with colleagues but only for 25 % of the respondents sharing it openly with everyone and more than 20 % not sharing any, re-enforces the notion that researchers want to control their data and reject the concept of un-controlled open data sharing.



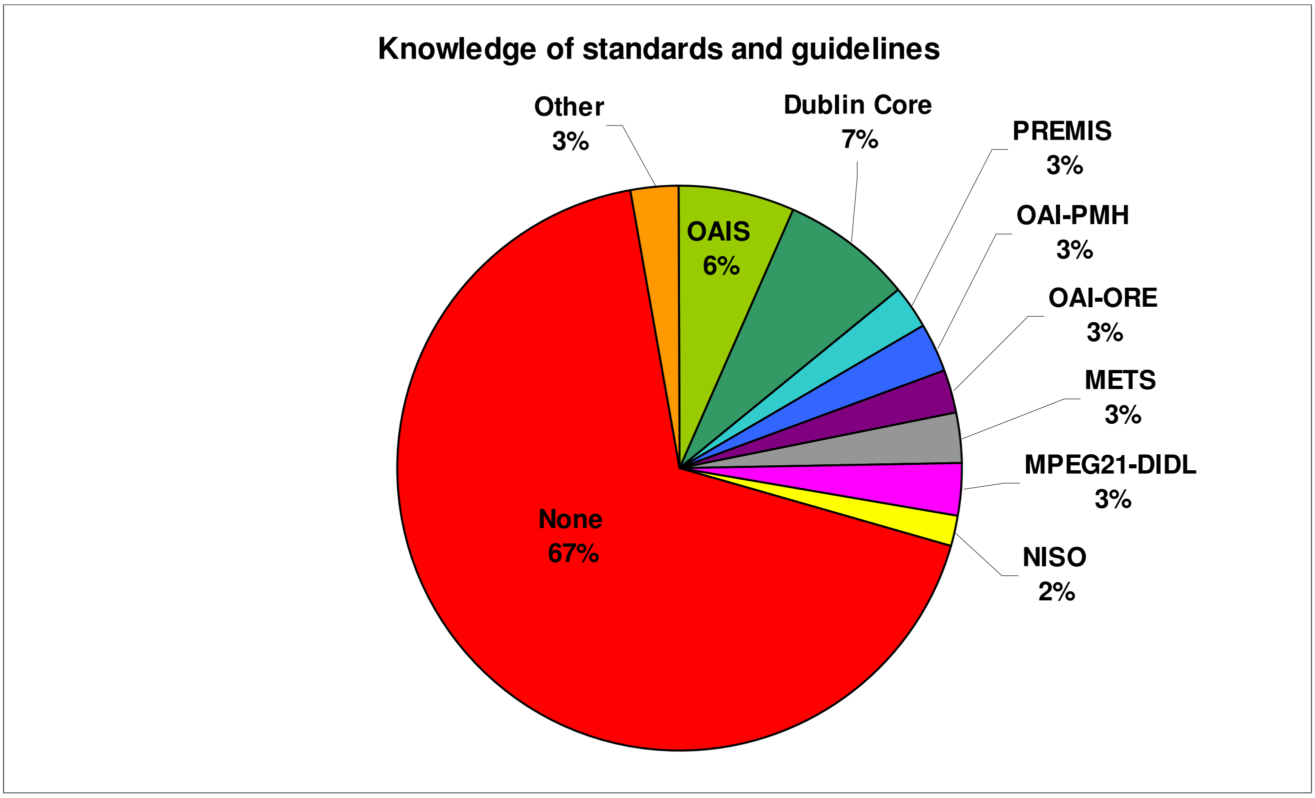
Graph 7: Which of the following applies to the digital research data of your current research N=1270 (PARSE.Insight[14])

They are open with their nearest colleagues and when their professional practice mandates it (e.g. publisher policy or good practice), though the extent of data sharing decreases rapidly as their potential control over the data decreases, i.e. they wish to choose when and with whom they share their data (Graph 7).

However the researchers' practice becomes more complex when questioned about their specific data needs, i.e. not what they do with their own data but rather what they require from others.

According to the responses to the PARSE.Insight survey, 63% of the researchers would like to make use of data gathered by other researchers in their discipline (N=430) while 70% of respondents already do (N=638). Interestingly, when asked about data gathered by researchers in other disciplines, still 40% would like to make use of it (N=689), and 46% state that they do so already (N=1264).

Some researchers have identified a requirement in skills training or resources to assist in the safe and appropriate maintenance processes of data management[17] and this is likely to be reflected across other disciplines. Several organisations are producing guides and tools to assist with this, but it is unclear if they are making as big an impact as they perceive. A key to this was the extent and use of data standards together with their effective uptake profile. A great majority of respondents claimed to have used no standards when offered a selection of those in current use (Graph 8).

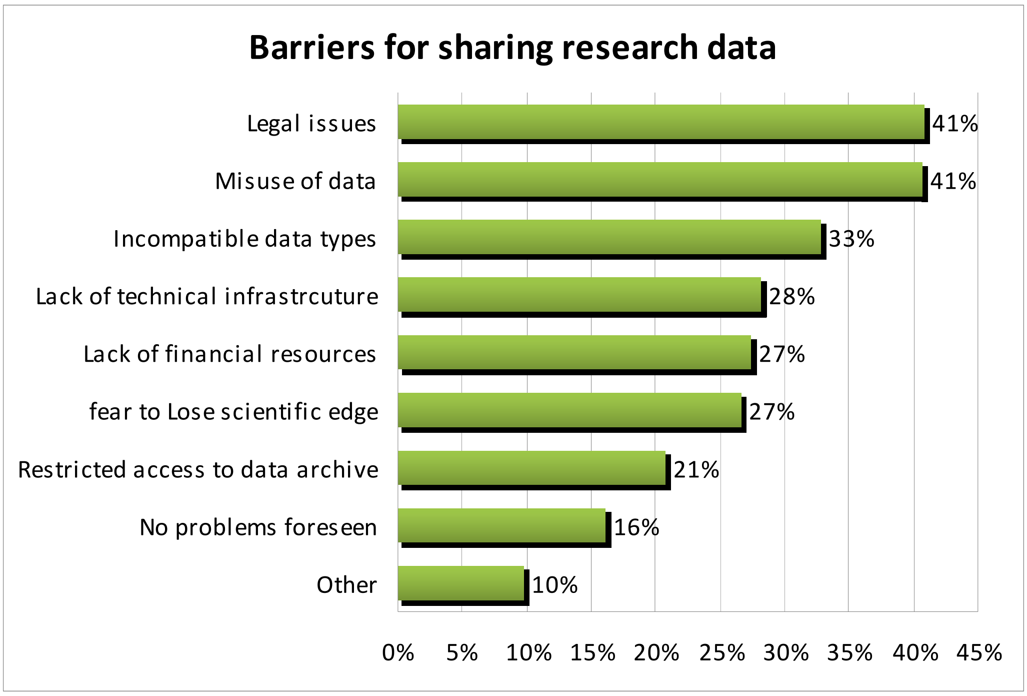


Graph 8: Which of the following standards or guidelines that are used in digital preservation are you familiar with? n=1202(PARSE.Insight[14])

It is well known that the development of community standards is a slow and demanding process. Agreement is based on function and often there are many opinions on pragmatic utility than can be accommodated in an easily implemented standard. There may be a case for supporting standards development as a community activity more actively, though how this could be achieved needs to be defined.

[17] Reichman OJ, Jones MB, Schildhauer MP (2011). Challenges and opportunities of open data in ecology. Science. 331(6018):703-5

Researchers perceive legal and professional reasons for not sharing their data



Graph 9: source PARSE.Insight: Do you experience or foresee any of the following problems in sharing you data? N=1270(PARSE.Insight[14])

Barriers to sharing data, whether real or perceived, were mixed between reasons for not sharing and reasons for not using certain data types or particular data sources. For example, Graph 9 suggests many respondents suggested there were legal barriers to their sharing, though it was unclear if these were fear of prosecution or responsibility for IPR particularly in biomedical sciences that involve human subjects or commercial potential[18]. Licensing research data is a recognised as a complex and time consuming activity[19] and there is a need to simplify and streamline the process by which researchers or those in control of research data assert control over their data assets, an opinion supported by the recent Hargreaves report in the UK[20]. The next most frequent barrier was a fear of misuse, which may include a validation threat to analyses that contradict the original findings, discovery of additional findings or exposure of the data creator to legal redress, thus strongly associated with the most common response of legal issues. For example repurposing data that leads to breaching data protection legislation could hold the data publisher or those responsible for the data vulnerable to criminal proceedings. Finally, incompatibilities between data and lack of a financial and technical infrastructure were cited as strong barriers to sharing, long recognised as a

[18] Mathews DJ, Graff GD, Saha K, Winickoff DE (2011) Access to stem cells and data: persons, property rights, and scientific progress. Science. 331(6018):725-7

[19] Alex Ball (DCC) 2011. How To Licence Research Data. A Digital Curation Centre and JISC Legal 'working level' guide.

[20] Digital Opportunity A Review of Intellectual Property and Growth, An Independent Report by Professor Ian Hargreaves May 2011

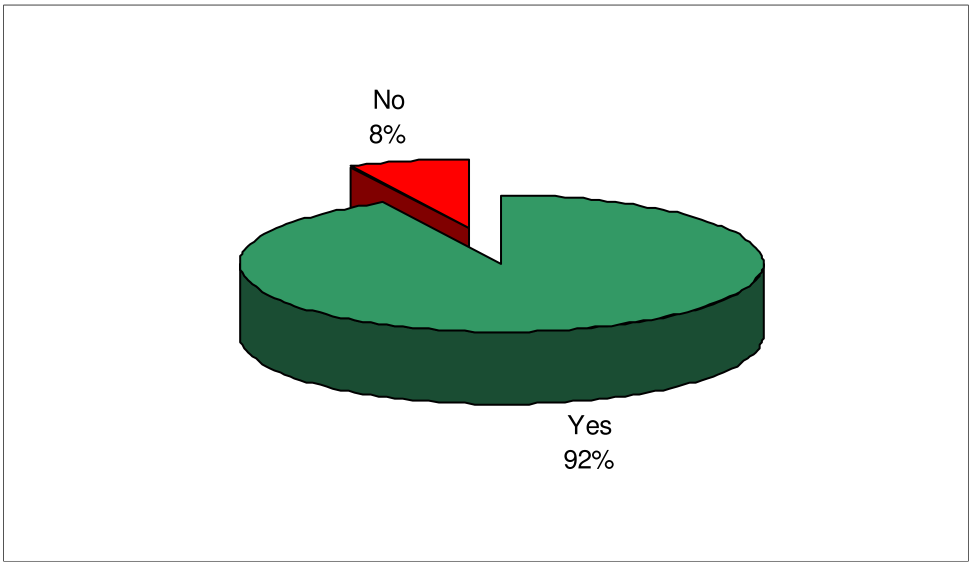
result of information technology and computational ability moving faster that our capability for data management and planning.

Taken together these results indicate researchers are willing to use others data providing they can validate them, but are wary about openly sharing their own. This need not be seen as a closed relationship. It could well indicate deeper levels of sharing enablers than simply altruistic motivation, e.g. attribution, provenance and reliability.

Researchers want credit

Even if data can be shared or published there was, as expected, almost universal recognition that a 'credit to the data creator' facility must exist (Graph 10). Good research practice requires recognition for intellectual contributions and these should include data. In the same way citation of traditional publications play for recognising individual intellectual work, a similar convention is required for data though no agreed convention exists[21,22].

Do you want to be credited when your underlying digital research data is used by others? (PARSE.Insight)



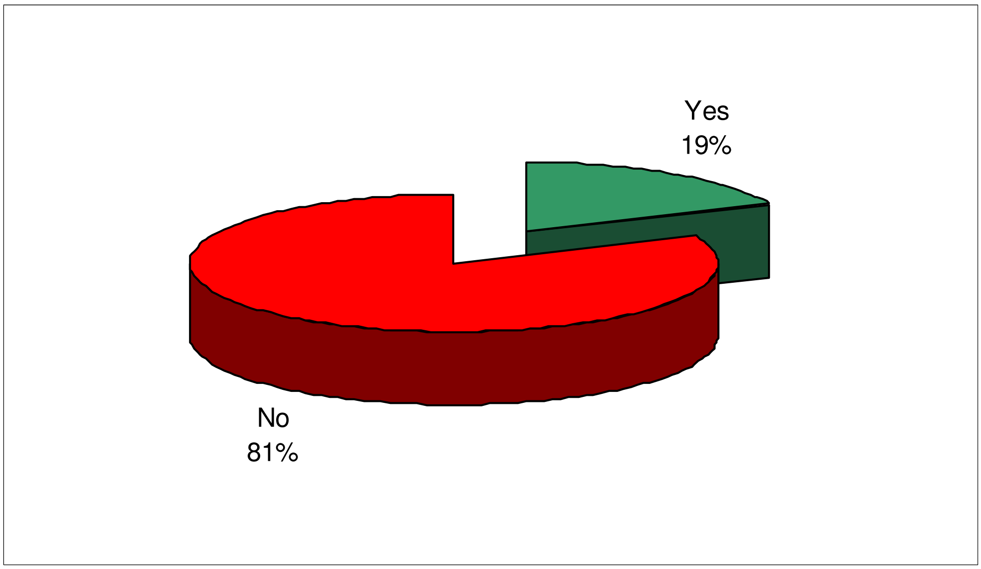
Graph 10: Source PARSE.Insight[14], N=1171

However, when it comes to clear guidelines about how data can be made citable and how this can be integrated with traditional publications, there are few examples of common standards or shared good practice. This is supported by the findings in PARSE.Insight, where the vast majority of researchers understand the benefits of a joined scholarly communication but were are unaware of publisher policies and these policies (see Graphs 11 & 12 and Chapter 3).

[21] Starr,J. Gastl, A (2011) isCitedBy: A Metadata Scheme for DataCite. D-Lib. Volume 17, Number ½ doi:10.1045/january2011-starr

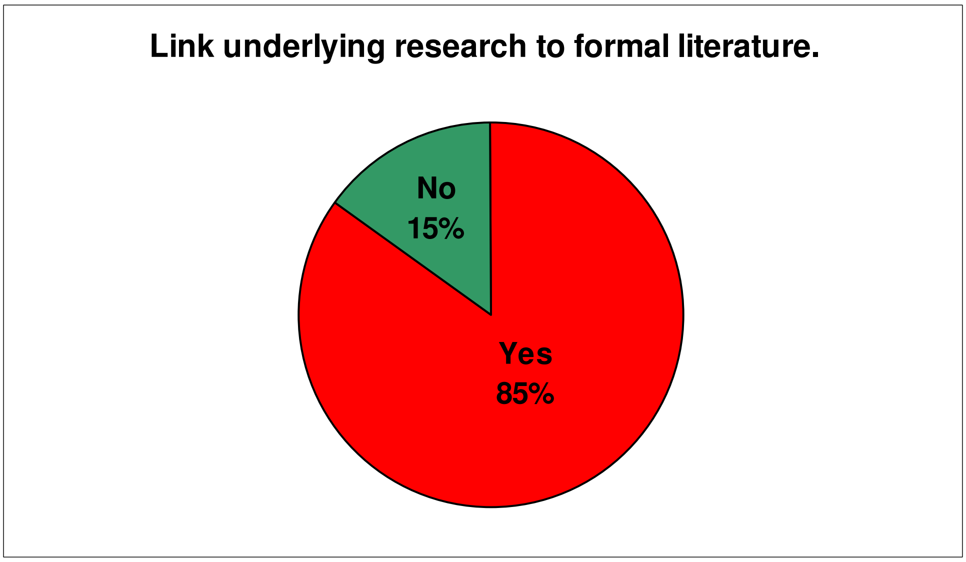
[22] Brase, J and Farquhar, A (2011). Access to Research Data . D-Lib. Volume 17, Number ½ doi:10.1045/january2011-brase

Do journals to which you typically submit your work require you to include relevant digital research data (i.e. data used to create tables, figures, etc.)? n=129(PARSE.Insight[14])



Graph 11: Source PARSE.Insight[14], N=1171

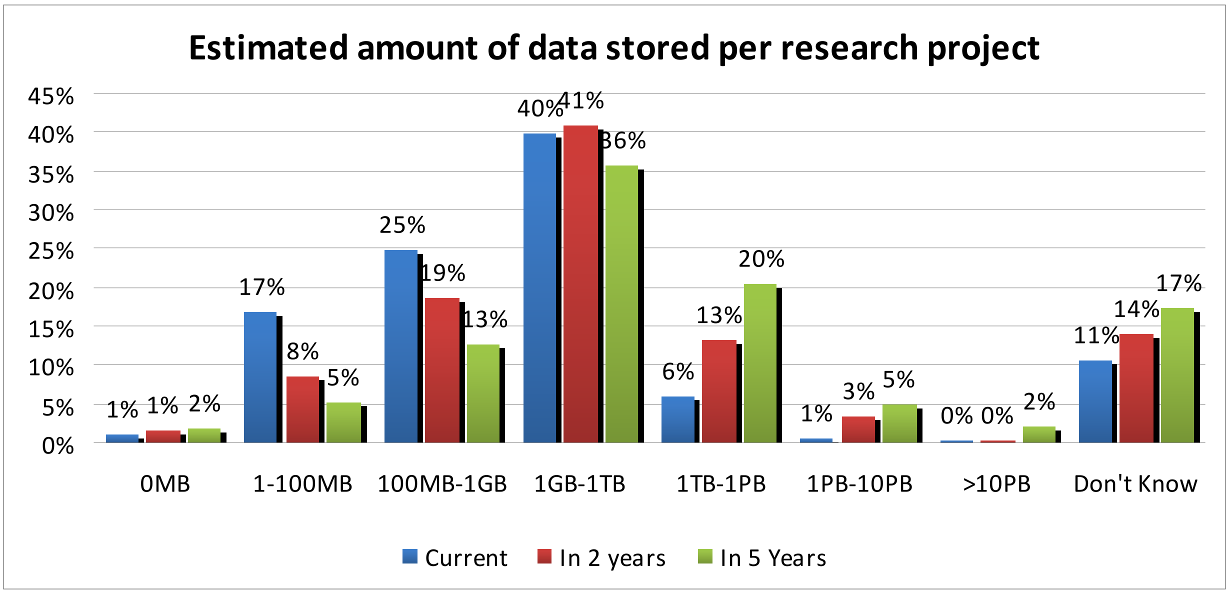
Do you think it is useful to link underlying research data with formal literature?n-2289(PARSE.Insight14)



Graph 12: Source PARSE.Insight[14], N=1171

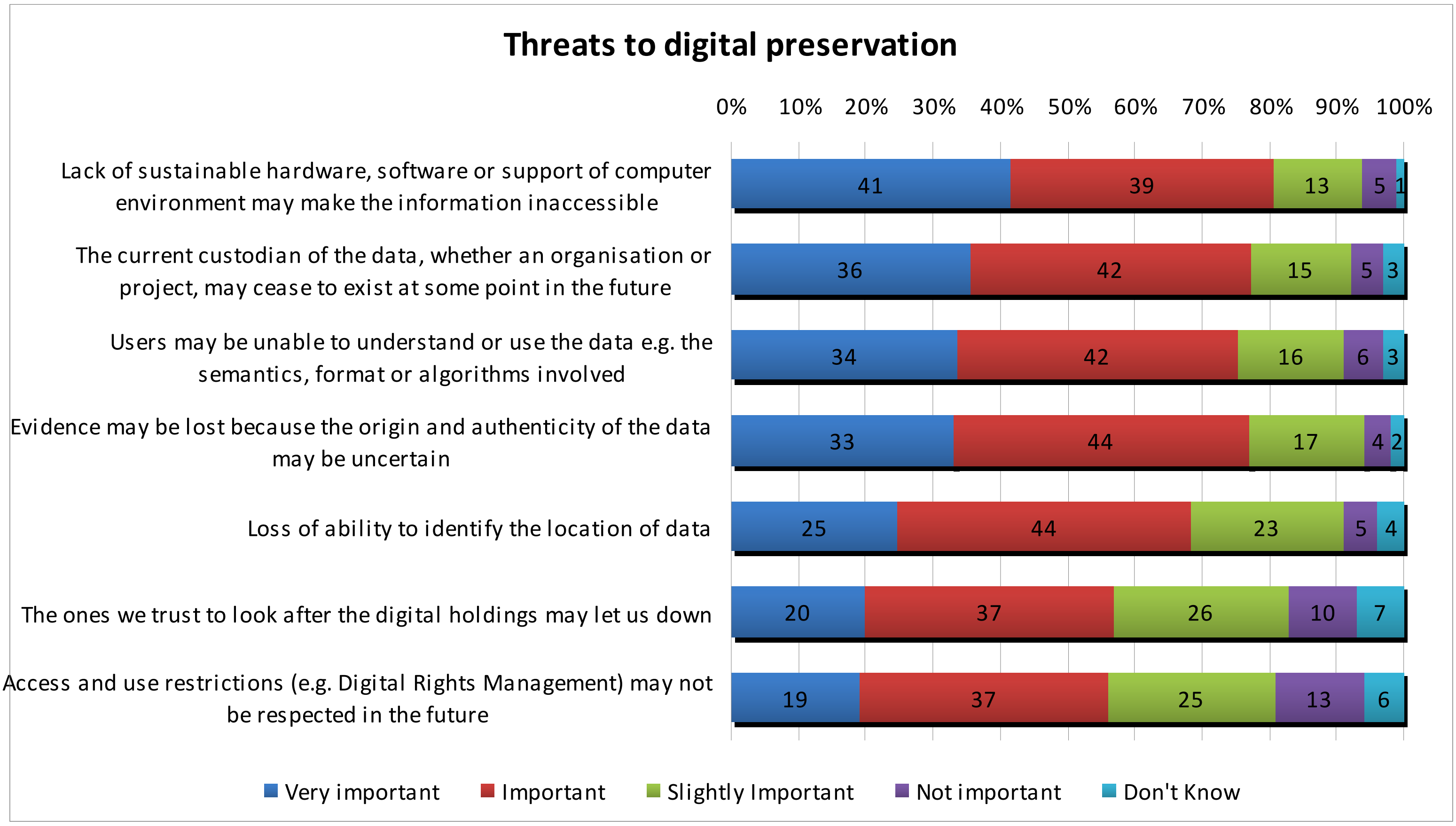
Many researchers see the problems with data getting worse

The problem of data volume is illustrated by a steady increase in expectancy from research together with an increasing 'don't know' cohort in the PARSE.Insight investigations of data volume expectations. The figures for data volumes below 1Gb-1Tb exhibit decreasing expectancy while those about 1Gb-1Tb are increasing expectancy (Graph 13).



Graph 13: Source PARSE.Insight[14], N=1202

Together with data volume expectancy there are well understood data preservation risks that include lack of infrastructure and custodian roles, indicated by data preservation issues being perceived as either important or very important by the majority of respondents (Graph14).

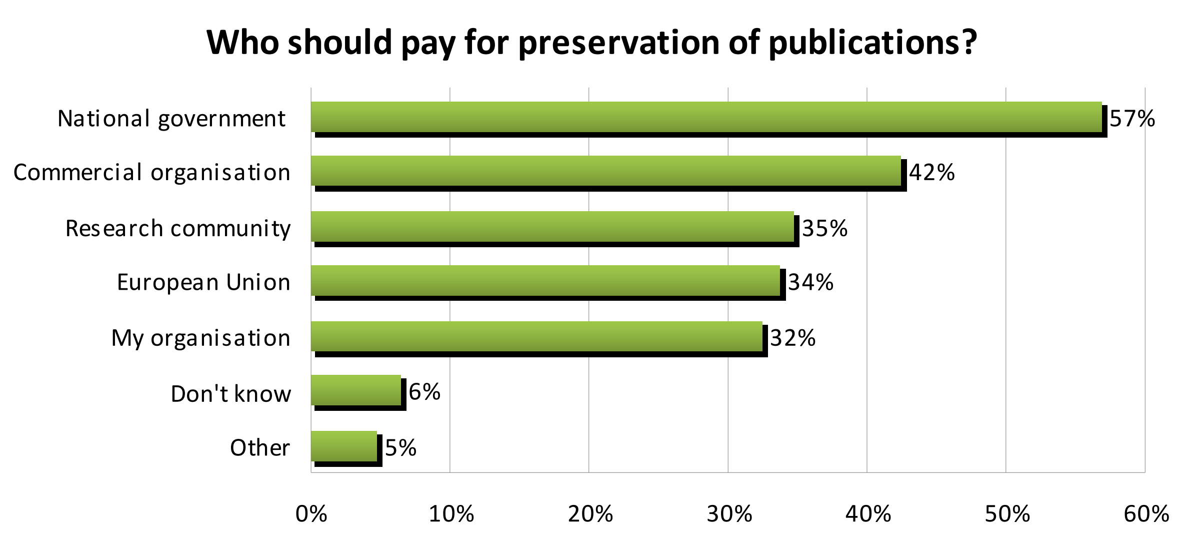


Researcher perceived threats to digital preservation(n=1201-1210) (PARSE.Insight[14])

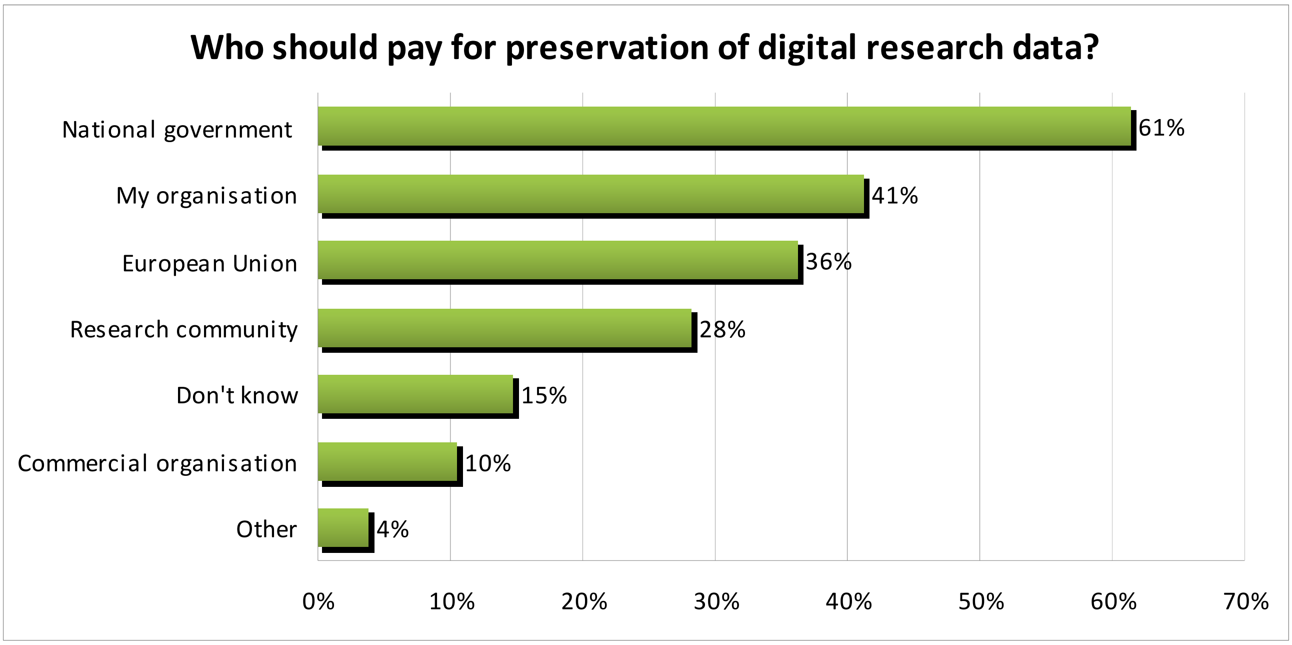
Graph 14: Source PARSE.Insight[14], N=1201-1210

Researchers believe others should pay

There are strong views, possibly reinforced by an appreciation of preservation and hardware considerations, that once the data have been generated and used, their preservation and archiving responsibilities should rest with other organisational structures, with the exception of a non-specific 'research community' being identified. These were firm views as few respondents either didn't know or selected 'other' (Graph 15 &16).



Graph 15: Source PARSE.Insight[14], N=1188: Who, in your opinion, should pay for the preservation of publications?, multiple answers possible)



Graph 16: Source PARSE.Insight[14], N=1188: Who, in your opinion, should pay for preservation of digital research data? (multiple answers possible)

2.3. Conclusions on current practice

Taken together there appears to be a complex relationship between the researcher and the data they collect and create. Researchers perceive and enforce their creator right over the data, choose when and with whom they share it and wish to maintain this control. This need for control appears based on perceived legal barriers and risk of misuse, or absence of a trust network common in other forms of scholarly communication; it may be a mixture of both. Researchers want somewhere safe to put their data while maintaining control in order to avoid legal redress and claims of misuse, but expect some central organisational structure to pay for these infrastructures. In a study of data sharing in the biomedical informatics domain, a training review indicated that many researchers recognise they lack sufficient skills to manage their data appropriately, but importantly are enthusiastic to change this situation23. Researchers would benefit in joining the publication with the data in a more formal and agreed convention, and recognition and credit mechanism for this can help as important drivers and incentives. They accept joining data to publication as good professional practice (see graph above) and agree that data supporting traditional publication should be available with the publication. Technology can reduce the latency to joining data to publications, but the policies of the publishers requiring the availability of data supporting publications are so far very much in a pioneering stage (see chapter 3).

2.4. Is there a need or case for change?

Data and publications belong together24 and researchers are the link between these two established intellectual objects. The evidence that supports scholarly discourse cannot be lost without severe consequences for scholarly communication. Distilled into statements, our desk research has revealed five abstract researcher requirements for integrating data and publication.

1. Researchers need somewhere to put data and make it safe for reuse
2. Researchers need to control its sharing and access
3. Researchers need the ability to integrate data and publication
4. Researchers need to get credit for data as a first class research object
5. Researchers need someone to pay for the costs of data availability for re-use

1. Where to put data and make it safe for reuse

Research data centres exist but they are fragmented and operate in different ways. Generally data centres are community or discipline focussed and where significant investment is available more large scale operations are established (as is evident in particle physics and astronomical disciplines). Such mature data archives are capable of taking full responsibility for the data they hold with clear preservation and access

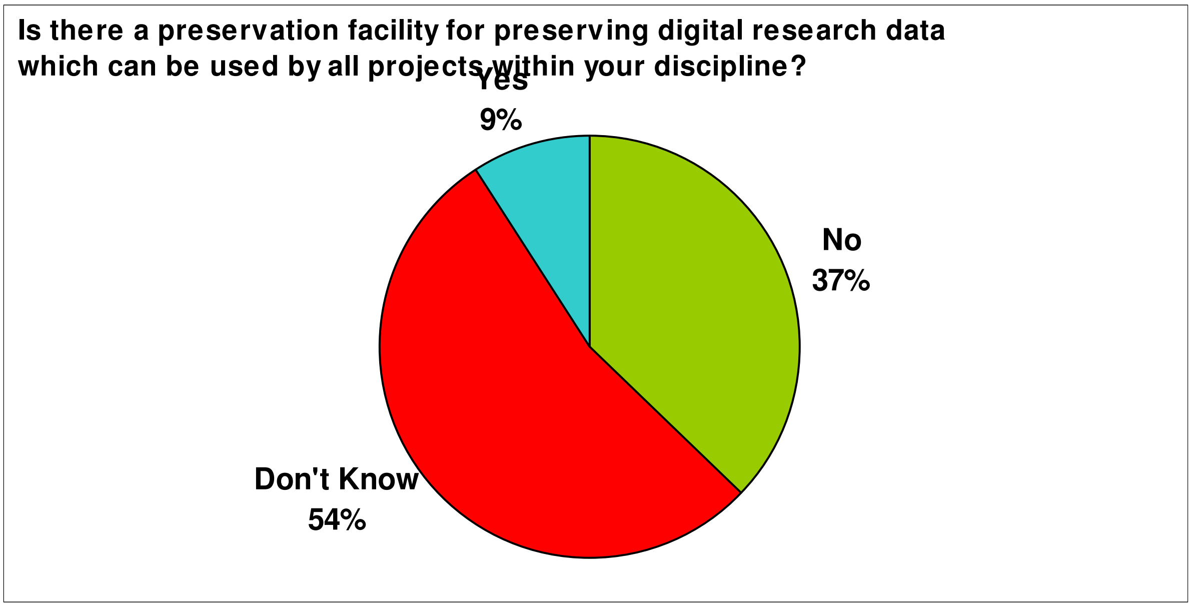
[23] http://www.cancerinformatics.org.uk/training.html

[24] Smit, E (2011), Abelard and Héloise: Why Data and Publications Belong Together D-Lib. Volume 17, Number ½ doi:10.1045/january2011-smit

policies, with some making use of community developed quality tools like the Data Seal of Approval[25]. These organisations can either be concerned with 'big data' in the case of particle physics and astronomical data centres, or complex social data, as in the case of UK Data Archive. In contrast there are numerous *ad hoc* collections of community activities borne out of immediate need and uncertain future. These so-called 'long tail' data centres in numerous and low level data generating disciplines, e.g. ecology, evolution etc have the potential to produce more volume and more complex data for which 'big data' solutions will not be appropriate. In turn these 'long tail' data centres will likely require much more resource to both establish and maintain.

Sustainability models for various data needs across disciplines would assist in determining how much resource is required for what type of data and for how long. A number of pilot and low scale projects have attempted to establish this alongside the well known and established initiatives for 'big' data[26]

In summary, the research data landscape is both large and complex. PARSE.Insight provides evidence that these examples are not enough: researchers are either unaware (54 %) or know there is no facility (37 %) to put and maintain their data.



Graph 17: Source PARSE.Insight[14], N=1198 Is there a preservation facility for preserving digital research data which can be used by all projects within your discipline?)

[25] Klump, J.(2011). Criteria for the Trustworthiness of Data Centres D-Lib. Volume 17, Number ½ doi:10.1045/january2011-klump

[26] www.datadryad.org

2. How to control data(?) sharing and access

Disciplines vary widely; the RIN report on patterns of information use and exchange provided compelling evidence from in depth interviews on the differences between sharing practices across life science disciplines[16]. There appeared many levels of control that were either encountered (while looking for data) or imposed (when the custodian of data). What this suggests is the perception of and need for data sharing are confounded by confusion over the risks of data sharing. With the PARSE.Insight data suggesting a fear of legal redress and misuse of data as the main concerns it would seem that ownership and responsibility for data have become the default stance over authority; without a clear accreditation or registration framework for data, like citation, or a open sharing environment with accountability, reluctance to share wins out over professional transparency. Researchers have ownership of data as a consequence of generating it, but transferring licence or ownership to use or share data is a confusing barrier and if there are any likelihoods of exposure to prosecution for either data misuse or unethical data collection then it is easier to simply close access to the data and severely limit any sharing. No clear authority or accountability role is available for research data outside its ownership.

So as data enters scholarly communication the issue arises; who will be responsible for it, securing the authority to persist, preserve, share it and take legal responsibility.

Some co-ordination and advice centres exist. The DCC[27] in the UK was established to support UK researchers engaged in digital research activities. It has since expanded into international partnerships and provides a rich resource for any researchers or institutional service (eg digital repository or similar) and offers advice on all manner of data curation issues including licensing data. In addition the UK Data Archive releases a series of best-practice guidelines for researchers from the social sciences that can be applied more widely**Error! Bookmark not defined.**.

3. The capability to integrate data and publications

Many researchers see the benefit of integrating the scholarly record but little best practice conventions exist. In contrast to the practicalities of what to do with data there are a growing number of examples of how to re-join data to the publications they support (see Chapter 3).

4. How to get credit

Almost all respondents in PARSE.Insight agreed that should their data be used they should receive credit for it. This is in line with the professional impact that researchers receive from publication activity. Presently there is no agreed framework for citation of data nor a capability to measure impact in a similar manner to the way traditional citation impact factors are created. Thus the question is raised, does data need a specific

[27] http://www.dcc.ac.uk/

citation/recognition framework or is the current framework sufficient to absorb the requirements of data citation? DataCite[28] believe data are first class research objects with separate requirements from scholarly publication and as such require a citation framework that can accommodate these. DataCite is an international association that is implementing such a framework independent of discipline to support data citation via the registration of persistent identifiers (DOI's) that enable linking to and from these data sets (see also Chapter 4).

5. Who pays for what

Researchers feel that while they recognise data preservation and archiving cost money, they are unable to pay for it. In fact the complex processes of data preservation are well understood in many disciplines outside research, especially those where severe and expensive legal requirements are imposed, e.g. financial institutions are required to keep data for many decades, nuclear installations are required to keep digital records and data for perhaps centuries. In the UK at least, the cost of both data preservation and data sharing is recognised in many policies from the Research Councils. Assisting in developing and promoting these policies with well documented and realistic scenarios and use cases for best practice data management throughout the research process would be advantageous to all .

2.5. Opportunities in data exchange relating to researchers

In summary, returning to the criteria over which we attempt to identify opportunities for data exchange, from a researcher's perspective the following have been identified.

Data Issue:	Researchers opportunity to help improve situation:
Availability	Researchers demand their data be treated as first class research objects Researchers loosen control over data Define roles of responsibility and control
Findability	Agree convention to propose to publishers regarding data citation Use of persistent identifiers such as DOI's Ensure common citation practices
Interpretability	Recognize that data require metadata and work towards community best practice in metadata development
Re-usability	Be concerned about the long term ability for secondary use and consider or seek out responsible preservation actions. Further, consider this as part of good research practice rather than as a closing activity.
Citability	Agree a convention for data citation Follow metadata standards for datasets Use of persistent identifiers such as DOI's

[28] http://www.datacite.org

Curation	Develop sustainable and realistic data management plans Collaboration with public data archives
Preservation	Develop sustainable realistic preservation plans Active engagement with public data archives

Table 1 repeated: Data Opportunities for Researchers

These have been further broken down to incentives, drivers and enablers.

Incentives

Joining functions that complete scholarly communication by integrating data and publications

Citation framework that encourages credit, attribution and re-use to remove hesitation on the researchers side

Review/validation process (e.g. data journal) that support trust

Drivers

Impact and re-use metrics that support incentives

Data management plans/Data sharing plans as part of start up activity

Disentangled responsibility between data creators and data custodians

Enablers

Clear and consistent IPR and other rights statements from stakeholders

Mandated infrastructures like data centres and data archives that can persist and preserve

Recognition frameworks that support data as first class research objects

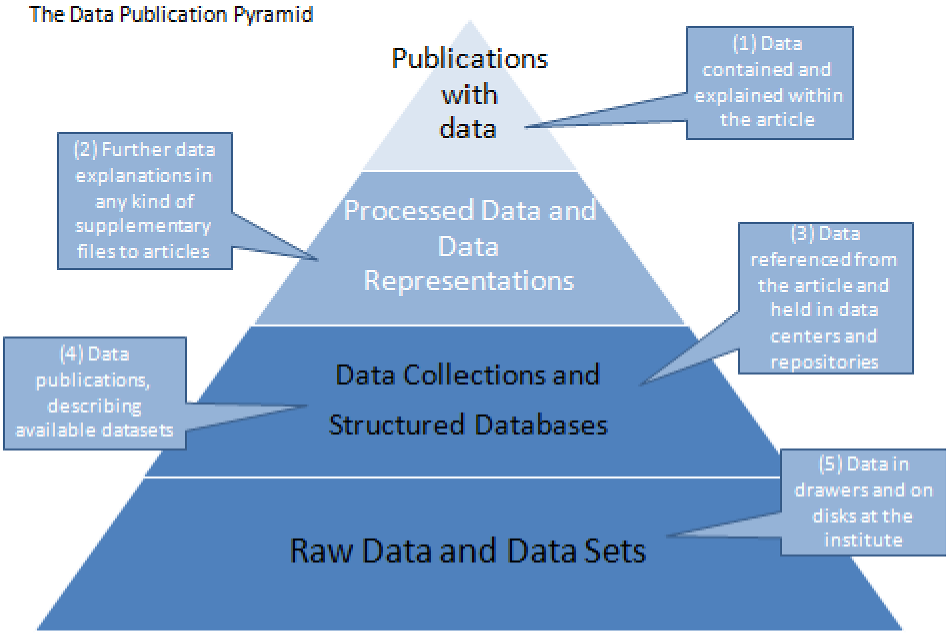
Submission processes that minimise overheads and effort

Embedded training activities and practices that focus on data management skills rather than simply data manipulation skills

3. INTEGRATION OF DATA AND PUBLICATIONS: THE PUBLISHERS' PERSPECTIVE

3.1. How scholarly journals handle the increasing amount of data alongside the article

Building on the previous chapters of this report, and well aware of the desire of researchers to publish their data in a citable way and to find data via the formal literature, we focus here on the ways publications and data are being integrated. In our pursuit of present practices and new initiatives in the field of STM journals, we have encountered (and will describe in this chapter) the 4 basic categories for different ways in which data and publications can be connected and/ or integrated. They follow largely the Data Publication Pyramid as presented in the Introduction Chapter and the first 4 categories listed there (while category 5: data in drawers and on disks, concerns the category of unpublished data, which is addressed further in Chapter 5):


The Data Publication Pyramid
Publications with data
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
Processed Data and Data Representations
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
Data Collections and Structured Databases
(5) Data in drawers and on disks at the institute
Raw Data and Data Sets

Graph 1 repeated: The Data Publication Pyramid, see chapter 1 for a full explanation.

This is a high level overview of each of these categories before they are described more in-depth in the remainder of the chapter, with several examples:

1. *Data contained within peer reviewed articles*

This is the traditional publishing model in which the researcher fully analyzes and processes the data and describes the conclusions derived from them in the scholarly article. The conclusions drawn from the data are illustrated by summarizing the relevant data (or data-outcomes) in tables, graphs and other illustrations, and, increasingly, also in multimedia applications.

Advantages are in the tight embedding and integration of the data into the scholarly record, citable and retrievable as such, available to all researchers and users. Authors get all credits for their article.

Limitations are that these present a high level of aggregation of the data, data are hard to find separately from the article and usually not in a re-use friendly way.

2. *Data resides in supplementary files added to the journal article*

Nearly all STM journals offer authors the service to add in supplementary files to their article any relevant material that is too big or that will not fit the traditional article format or its narrative, such as large datasets, multimedia files, large tables, animations, high resolution files, protocols, large bibliographies, etc. With the increased computational nature of many disciplines, the use of supplementary files has increased sharply recently (see also Chapter 1, Introduction).

Advantages of using supplementary files are that the volume of the data is no longer an issue and that the data are still closely tied to the official scholarly record and remain citable, while authors are no longer restricted by the article format. It makes optimal use of online facilities.

Limitations are that file size is usually not much larger than 10 GB and that, from an author's perspective, the curation and preservation of the supplementary files is not always clear. Few standards exist between journals on how to indicate the presence of supplementary files or where to find them. Only in few cases will the supplementary material be provided with a separate DOI or other persistent identifier to enable linking independent from the main article.

a. *Journal articles offer supplementary files with extra data, but restricted in volume and format*

A sub-category under Supplementary data files exists for those journals that have restricted the use of supplements for this purpose. The first examples of high standing journals which can no longer manage the sheer volume of materials in supplementary article files and as a consequence have put limitations on what can be included in supplementary files (Cell) or even who no longer accept supplementary material at all (Journal of Neuroscience) other than multimedia files that should be considered integral to the article content.

Advantages are that any supplements of a journal articles remain tied to the official scholarly record, and are part of its peer review process, but

Limitations are in the new restrictions posed; adding original research data is often no longer possible.

3. *Data resides in Community-endorsed Public Repository with bi-directional linking to and from articles*

In this model the data relating to a scholarly article are deposited in designated Public Repositories, best examples are GenBank and World Protein DataBank. The accession numbers of the data in those databases are added to the journal manuscript and referenced, often within the article text, as well as in the footnotes or reference list.

Advantages are that the data become part of larger datasets in the same area thereby serving the research community as a whole, and it is normalized, standardized, curated and preserved. The connection between data and publication are secured via the accession numbers that are embedded in the article. Even better are the examples (Pangaea, CCDC, PubChem) where the bidirectional linking between data and articles are secured, and likewise from the data to the articles. There are no restrictions on volume.

Limitations are that these databases tend to exist only in a few subject areas so far, mainly biology, life science, earth science and chemistry. The future of these databases often depends on government funding and may be threatened by budget cuts.

a. *Journals have set up an own storage facility for data*

A sub-category for data referenced from articles that reside in a special data-storage facility established by the publisher fits the example of Thieme publishers in Germany, as later described in paragraph 3.6. Thieme have recently instigated collaboration with the data facilities of FIZ Karlsruhe to offer their chemistry authors the possibility to store the raw and original research data alongside their articles.

Advantages lie in the strong link between the data and the scholarly record and the availability of the data for further examination and re-use to all the journal readers, editors and other users.

Limitations are the danger of creating new silos for data per journal or per publisher, which can be a barrier to discoverability and reuse.

b. *Journals offer dynamic data made interactive, data can reside with the article or in public repositories*

Another sub-category exists for journals who present the relevant data sitting in an official repository or data centre or anywhere else from within the article. This model emphasizes what readers of a journal article can do with the data rather than where the data actually resides. Via click throughs from graphs and tables, readers can play with the underlying data and their visualizations. As an example, the BioChemical Journal from Portland Press does this via dynamic pdf's. Elsevier has some examples of data viewers that work from within the article but using data in Genbank and PDB bank. This may increasingly emerge as a model for Linked Open Data and the emerging data web.

Advantages are clearly in reuse and in increased interpretability of underlying data. Data become re-usable within the context of the scholarly record.

Limitations are in the open availability of these data. Applications are usually only used for the data within the article, i.e. category 1 of this list.

4. *Journals dedicated to so-called Data Publications only*

In this model the journal publishes descriptive articles about datasets that are usually stored in a repository. The description of the data generation and its potential use allows the authors credits for their work while strongly promoting the interpretability and re/use. Examples are the Earth Science Systems Data Journal (ESSD) and the newly launched journal Gigascience. Other, already existing journals offer hybrid models in which they have opened up for descriptive data articles (Int Journal of Robotics) as a new article type next to the traditional research papers.

Advantages are in the credits for the author, the citability and the reuse.

Limitations are in the challenges for high quality peer review, very few peer review standards exist so far for datasets and their descriptions. The system depends on proper and persistent bi/directional linking.

We can summarize and compare these 4 categories (plus 3 sub-categories) in the following table. Subsequent paragraphs contain more extensive descriptions of current practices and policies.

Data (selections) underlying articles reside in:	**Advantages:**	**Limitations:**	**Examples**
Data contained within the peer reviewed article, in tables, graphs, plotting, etc	Analyzed data and relevant data selections are integral part of the Record of Science. Readers, users and peer reviewers can find and consult these data selections	Usually high level of aggregation of the data, more data summaries than full set of original data. Usually not findable or retrievable separately from the article. Not well reusable outside the context of the article.	All peer reviewed scholarly journals
Supplementary files to journal articles – whatever it contains with very few restrictions on size and format	Datasets and publications tightly connected, data is embedded in public record of science, managed and preserved as such, author gets full credits, reviewers and readers are able to access data in combination with the article	Volume is a limitation, usually datasets not bigger than 10MB. Curation sometimes unclear, preservation likely to remain limited to that for articles. Easy discovery and re-use hampered by fragmentation over journal silos. Not all supplements are linkable. Sometimes the files can only be accessed via the article and not independently.	Vast majority of STM journals and demand for this from authors is increasing rapidly lately. Some journals have made the availability of underlying research data (Nature, PLoS) a condition for publication. First examples appear of journals who can no longer handle the overload (Jnl of NeuroScience, Cell)
2.a Supplementary files to journal articles, with restrictions and tightened instructions to manage the proliferation of supplemented material	More clarity on the supplemental materials that journals can and will support. Better reassurance of curation and preservation and perpetual access. Journals will encourage authors to place unsupported materials in a reliable repository	Volume is usually further restricted to underlying tables or explanatory graphs and full data sets are not included.	Cell, Journal of Neuroscience, and a growing number of journals contemplating such restrictions as they find it hard to handle the growing volume and variety in formats (example: NISO/NFAIS working group)

Community-endorsed Public Repository with bi-directional linking to and from articles	Data resides in a place where proper curation and formatting is secured, as well as discoverability and reuse. The bidirectional linking ensures connections with the publications. Author credits are indirect.	Life science area is well covered with many initiatives, in other areas first initiatives emerging. Many disciplines still lack a common solution. Future of existing repositories sometimes under threat because of pending cuts in government funding.	Most journals in molecular biology and life sciences list these repositories and require authors to deposit there and submit the accession numbers of the databases. Strong supporters for this approach are PLoS, Science, Nature. Pangaea is an often cited example in earth sciences, collaborating with all Elsevier journals in this area.
3.a Journals have set up an own storage facility for data	Authors can be ensured that data is well curated and gets the right metadata attached for findability and check of provenance. Data remains closely connected to the article and becomes part of the public record of science.	Data will be spread over many different journals and may end up fragmented over silos, hampering reuse across different platforms.	Thieme, for its chemistry journals in collaboration with FIZ and TIB
3.b Journals offer dynamic data made interactive, data can reside with the article or in public repositories	Graphs that show the underlying data via an extra click add depth to a research article. Data remain in the context of the article and become reusable at the same time.	Usually only applied to data presentations in the article, not to (large) raw data set as such.	The Biochemical Journal by Portland Press, experimental special issues by the Optical Society in America (OSA), Elsevier for data in the World ProteinDataBase (PDB)
Journals dedicated to so-called Data Publications only	Data are described in-depth in these publications, facilitating findability, interpretability and re-use. Data remain in larger repositories and can be combined with other datasets. Data creators get the full credit of a public record of science. The data becomes citable.	Peer review of large sets of data is a challenge. Curation and preservation is in hands of the repository. The system requires persistent bidirectional linking to work well.	Earth Systems Science Data working with Pangaea.

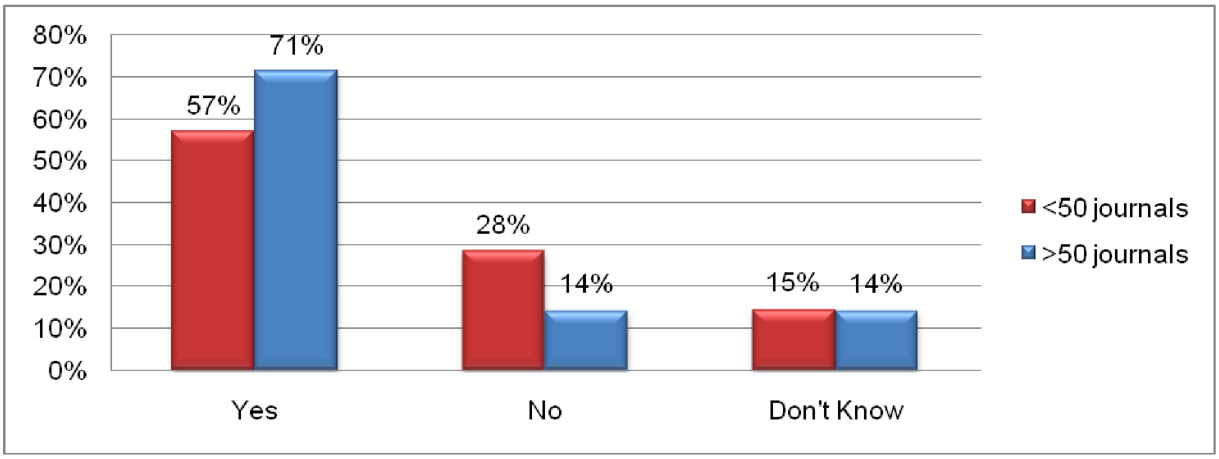
Table 5: Categories to publish research data.

In the following paragraphs, each of these categories will be further explained and illustrated with real life cases.

3.2. Common practice: Supplementary Material to Journal Articles

The large majority of journals accept research data in supplementary files. From the results of the PARSE.Insight[29] survey we know that most journals accept research data in supplementary files:

Can Authors submit their underlying research data with their publication?

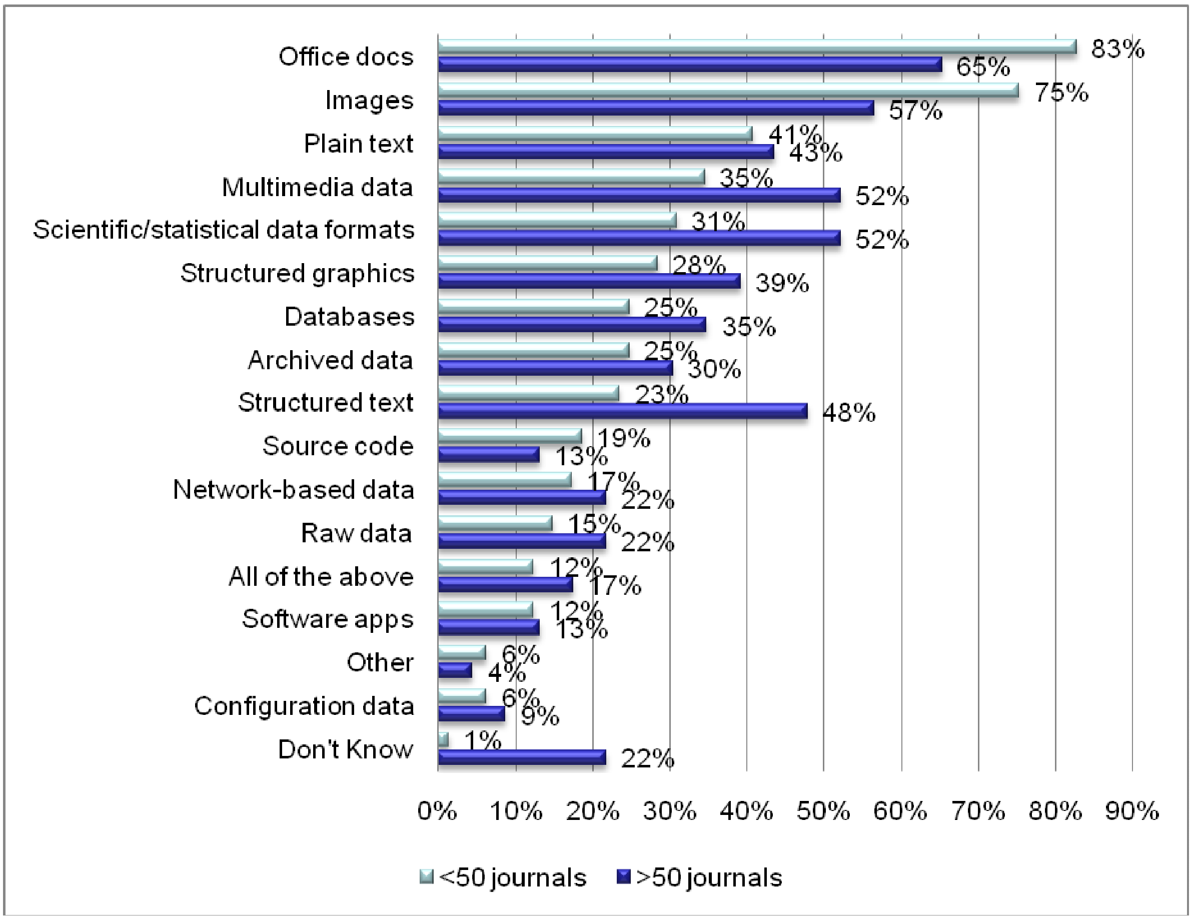


Graph 18 from PARSE.Insight[2] survey: N = 134 Publishers

If we weigh in the size of the publishers (see PARSE.Insight report; the 3 % largest publishers publish 70 % of all journal articles and they all accept supplementary files), then over 90 % of all journals accept supplementary files with research data.

Publishers generally accept a wide range of file and data formats (again: source PARSE.Insight):

[29] PARSE.Insight survey, see http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf



Graph19 from PARSE.Insight survey: What file formats does your journal accept ?, N=134 Publishers

As was shown in the introductory chapter (Chapter 1) of this report, we know that more than half of the researchers surveyed for PARSE.Insight would like to submit their data together with their manuscript to journals and publishers. This unveils a trend of likely growth in the submission of supplementary data files, because the present percentage who do so is below 20 % at the moment. Publishers confirm this growth trend as will become clear from several examples provided in this chapter.

The instructions that these researchers find for most of the journals are fairly straightforward. This is the general instruction to authors from a large publisher with more than 2000 journals (ic Springer, bold text by us)[30]:

> *<The Journal> accepts electronic supplementary material (animations, movies, audio, large original data, etc.) which will* ***be published in the online version*** *along with the article or a book chapter. This feature can add dimension to the article, as certain information cannot be printed or is more convenient in electronic form.*

[30] Springer author instructions, see http://www.springer.com/authors/manuscript+guidelines?SGWID=0-40162-12-339499-0

And from a similar size other publisher we find very similar instructions for most of their journals (bold typeface by us)[31]:

> *Elsevier accepts electronic supplementary material to support and enhance your scientific research. Supplementary files offer the author additional possibilities to publish supporting applications, high-resolution images, background datasets, sound clips and more.* ***Supplementary files supplied will be published online alongside the electronic version of your article*** *in Elsevier Web products, including ScienceDirect.*

With a similar preference for a certain file formats as found with other publishers.

Some journals have started to make the availability of underlying research data conditional for acceptance of the article by the journal. See for example the text by Nature in its authors instructions (underlining added by us)[32]:

> *An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims.* ***Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications in material transfer agreements****. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission.*

Nature says that supporting data must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript. But the journal does not say where authors should make their materials, data and protocols available for readers and users, except for a few exceptions regarding public repositories that will be mentioned in the next paragraph. The same is the case for the open access journal PLoS One[33]:

> *PLoS is committed to ensuring the availability of data and materials that underpin any articles published in PLoS journals. We believe the ideal is that* ***all data relevant to a given article and all readily replaceable materials be made immediately available without restrictions*** *(whilst not compromising confidentiality in the context of human-subject research). (....)*
> ***Failure to comply with this policy will be taken into account when publication decisions are made.*** *We encourage researchers to contact journal editors if they encounter difficulties in obtaining data or materials from published articles. PLoS reserves the right to post corrections on articles, to contact authors' institutions and funders, and in extreme cases to withdraw publication, if restrictions on data/materials access come to light after publication.*

How effective these mandates are is not known and anecdotal evidence points at a long way to go still. In an empirical study of data sharing by authors publishing in PLoS

[31] Elsevier author instructions, see an example at: http://www.elsevier.com/wps/find/journaldescription.cws_home/505772/authorinstructions#87000

[32] Nature instructions to authors on availability of data: http://www.nature.com/authors/policies/availability.html

[33] PLoS: http://www.plosone.org/static/policies.action#sharing

journals, it is reported that requests for data underlying the published articles were only successful in one case out of ten, in spite of the journals' clear policies (Savage and Vickers 2009)[34].

3.3. New limits on Supplemental Files to journal articles: restrictions to supplements

The notion is growing that scholarly journals struggle to handle the exponential growth in supplementary data files and the responsibility for them in terms of securing permanent access and long term preservation. As an effect, the Journal of Neuroscience has declared a new policy in 2010 to no longer accept supplementary materials at all. Instead, authors may host supplemental material on an external web site and include in their article a footnote with a URL pointing to that site and a brief description of its contents. But reviewers and editors will no longer evaluate the supplemental material, the article as submitted will be treated as a self contained entity.

At their journal site we can read the reasoning for this there is a clear indication that the exploding volume and the burden for peer review was becoming a real obstacle (underlining added by us)[35]:

> *Although The Journal has published electronically since 1996, supplemental material first appeared around 2003. Since then, the amount of material associated with a typical article* ***has grown dramatically*** *(….) The sheer volume of supplemental material is* ***adversely affecting peer review****.*

The related editorial even speaks of a 'proliferation among authors' adding more and more material in the article supplements.

In a similar context, we find a new policy by the journal Cell, implemented in October 2009 and again referring to the ever growing amount of supplementary material (underlining added by us)[36]:

> *Supplemental Information is a useful resource for presenting essential supporting materials online, and Cell Press is committed to the publication of these materials. However,* ***as the amount of Supplemental Information has grown, it has become increasingly difficult for authors, reviewers, and readers to navigate due to the volume of information and the lack of defined structure and limits****. To address these problems, we are introducing these guidelines, which we believe will make Supplemental Information more useful and accessible to readers.*

The restrictions that Cell puts in place concern a limit on volume and total number of supplementary items:

> *The total number of supplemental data items of all types (figures, tables, movies and other) per paper may not exceed two times the number of figures and tables*

[34] Savage and Vickers 2009 in PLoSOne: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0007078

[36] Cell editorial: http://www.cell.com/retrieve/pii/S0092867409011817

> *in the main paper. For example, a paper with 7 main figures can have up to 14 supplemental items total, of which up to 7 may be figures.*

Another restriction from ***Cell*** is that the supplemental material must bear a direct relationship to the main conclusions and content in the paper and must stay within the same scope:

> *Supplemental Information is limited to data and other materials that directly support the main conclusions of a paper but are considered additional or secondary support for the main conclusions, or cannot be included in the main paper for reasons such as space or file format restrictions. Supplemental Information should be within the conceptual scope of the main paper and not extend beyond it.*

In a quote in The Scientists (February 2011)[37], Emilie Marcus, Editor-in-Chief of Cell Press Journals, says "It had become a limitless bag of stuff." The publisher did not consider abolishing supplementary materials altogether because they have a diverse readership, with different levels of interest in a study's details, Marcus explained, but it was necessary to rein it in. "I do think there are different solutions for different journals," Marcus said. "Scientific communities and journals have probably not given enough thought to what to do with this capacity for supplemental materials. That needs to evolve."

In 2010 a working group was set up by NISO/NFAIS[38] with a charge to define best practice recommendations on how publishers can best treat supplementary information, in terms of inclusion, handling, display and preservation. These guidelines are expected to appear in the second half of 2011. They will cover best practices for supplementary files to journals and also roles and responsibilities for availability, findability, quality control and preservation.

3.4. How safe is data in supplementary journal article files? (or: quality and preservation of supplementary journal article files)

How do publishers treat the data in supplementary files and how safe is the data there? A complaint that publishers sometimes receive is about the lack of transparency over whether the supplementary files received from authors have been edited at all, peer reviewed or checked on format and general quality. Many publishers will just offer the author the service of posting the files in connection to the article exactly in the way they were received from the author. In some cases the files were part of the peer review process, in others they were supplied after acceptance of the article. Some publishers transfer the supplementary files into pdf's before posting, which does not serve the re-usability or further and deeper analysis of the data.

The ways in which the supplementary files are provided with metadata vary widely between publishers. Some leave this to authors entirely which does not add to

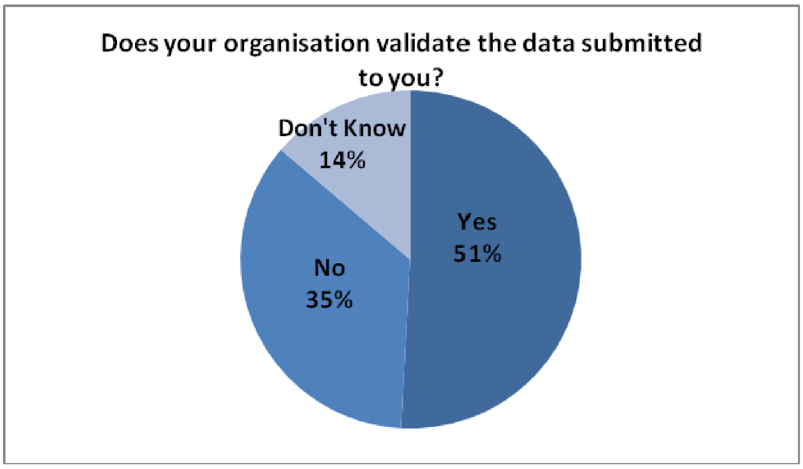
[37] The Scientist: Supplemental or detrimental? - The Scientist - Magazine of the Life Sciences http://www.the-scientist.com/news/display/58027/#ixzz1NRVKAV6F)

[38] NISO, see http://www.niso.org/workrooms/supplemental

consistency across data files. Also the mention of supplementary files in the article or the reference from the article to the supplements follows many different practices. Librarians find it sometimes a struggle to make sure that a literature search they carry out for research groups contains all related supplements (see Chapter 4). The NISO/NFAIS recommendations for Best Practice as mentioned earlier aim to advocate a clearer and a more common practice for this across publishers that help librarians find supplementary material when it exists.

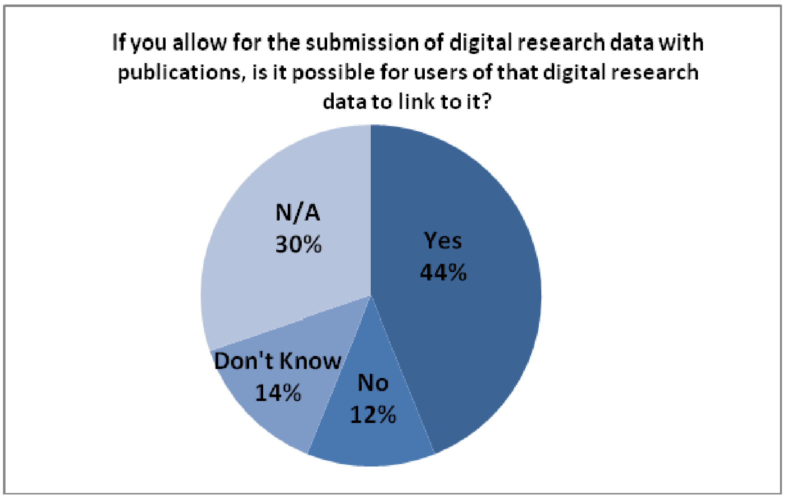
Again the findings from PARSE.Insight confirm this picture on many of its aspects. The 137 publishers responding to the PARSE survey say about research data submitted by the author:

- only 51 % of publishers validates the data submitted, mostly checking the file formats
- only 44 % facilitates direct links to it
- 39 % requires © transfer (against 57 % not)
- 70 % has no preservation measures in place for the supplemental data other than for the articles

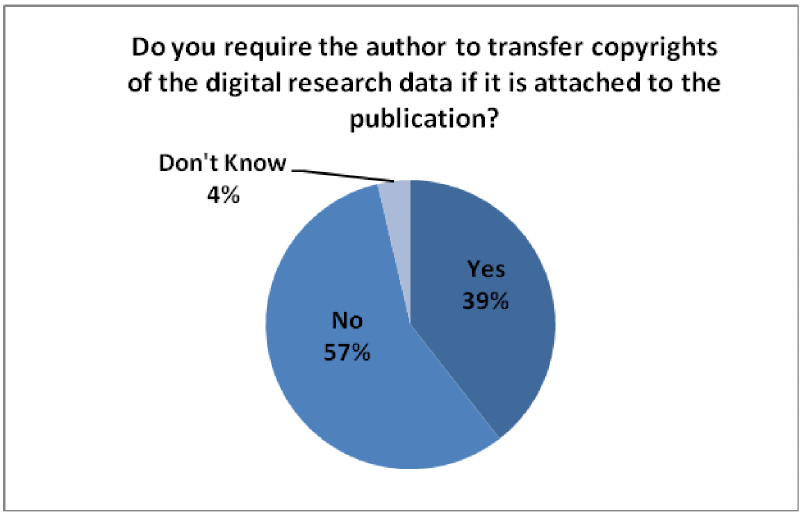


As shown in the following survey results :

Graph 20, source PARSE.Insight, N=137 publishers

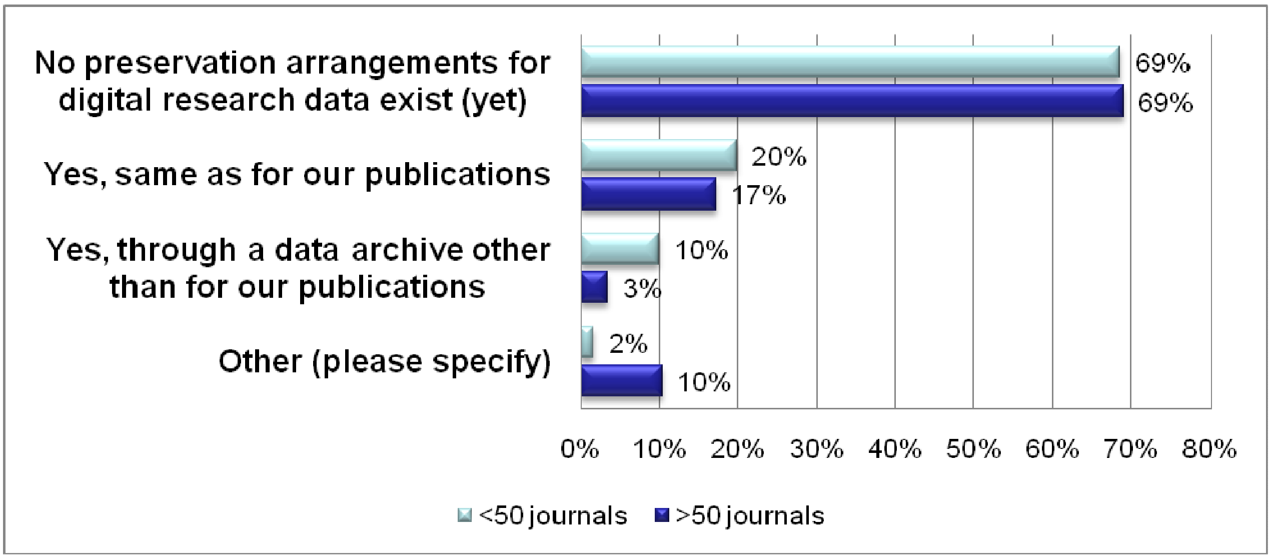


Graph 21, source PARSE.Insight, N=137 publishers



Graph 22, source PARSE.Insight, N=137 publishers

Do you have digital preservation arrangements for underlying research data



Graph 23 PARSE.Insight survey: (N= 137 publishers)

3.5. Data in community-endorsed public databases, linked to journal articles.

In a number of subject areas, community archives have emerged where researchers are requested to deposit their data. Some of these examples were already mentioned in Chapter 2 of this report. Publishers have adopted these practices and a large number of journals encourage the authors to deposit their data there, rather than sending it along with their article.

The advantages of this are manifold: the databases become more comprehensive, the data becomes better discoverable and can be used in combination with other data, and the connection with publications is ensured via bidirectional linking.

PLoS One has been advocating such a policy quite extensively[31]:

1. *Data for which public repositories have been established and are in general use should be deposited before publication, and the appropriate accession numbers or digital object identifiers published with the paper.*
2. *If an appropriate repository does not exist, data should be deposited as supporting information with the published paper. If this is not practical, data should be made freely available upon reasonable request.*
3. *The conclusions of a study must not be dependent solely on the analysis of proprietary data. If proprietary data were used to reach a conclusion, and the authors are unwilling or unable to make these data public, then the paper must include an analysis of public data that validates the conclusions so that others can reproduce the analysis and build on the findings.*

PLoS also adds about the ideal of data sharing:

> *We appreciate, however, that this ideal is not yet the norm in all fields. We are therefore currently collaborating with a number of subject-specific initiatives in order to develop relevant policies. In the meantime, authors must comply*

> *with current best practice in their discipline for the sharing of data via databases: for example, deposition of microarray data in ArrayExpress or GEO; deposition of gene sequences in GenBank or EMBL; and deposition of ecological data in DRYAD. We encourage all authors to comply with available field-specific standards for the preparation and recording of data.*

A similar policy encouraging authors to deposit their data in 'approved databases' is followed by the journal ***Science***. Their instructions say[39] :

> *Science supports the efforts of databases that aggregate published data for the use of the scientific community. Therefore, appropriate data sets (including microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for macromolecular structures, and climate data) must be deposited in an approved database, and an accession number or a specific access address must be included in the published paper.*

For those cases where such an approved appropriate repository does not exist, ***Science*** wishes to have the datasets on its own website as supplementary material or at least hold the material in escrow if the files are hosted on a institutional website by the author.

Further reading of the instructions seems to indicate that ***Science*** is also struggling with certain types of supplementary material in a careful balance to avoid having to absorb too much and still ensure that the underlying material can be examined by its readers. The journal asks authors to follow special procedures for making large or complex datasets available. See their special instructions for complex supporting data at their website.

While these well known titles are rather specific in their support for community endorsed databases, it has become an established custom in many areas for publishers to collaborate with the main data archives. Typical author instructions for data deposits in public archives, linked to the publications, are (source Elsevier[29]):

> *If your article contains relevant unique identifiers or accession numbers (bioinformatics)* ***linking to information on entities (genes, proteins, diseases, etc.) or structures deposited in public databases****, then please indicate those entities according to the standard explained below.*
> *Authors should explicitly mention the database abbreviation (as mentioned below) together with the actual database number, bearing in mind that an error in a letter or number can result in a dead link in the online version of the article. Please use the following format:* ***Database ID: xxxx***

And most publishers will specify the databases that allow links from (and increasingly to) the article:

[39] Science: see http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail and http://www.sciencemag.org/site/feature/contribinfo/prep/prep_online.dtl and http://www.sciencemag.org/site/feature/contribinfo/prep/prep_online_special.xhtml

> *Links can be provided in your online article to the following databases (examples of citations are given in parentheses):*
>
> - *GenBank: Genetic sequence database at the National Center for Biotechnical Information (NCBI) (GenBank ID: BA123456)*
> - *PDB: Worldwide Protein Data Bank (PDB ID: 1TUP)*
> - *CCDC: Cambridge Crystallographic Data Centre (CCDC ID: AI631510)*
> - *TAIR: The Arabidopsis Information Resource database (TAIR ID: AT1G01020)*
> - *NCT: ClinicalTrials.gov (NCT ID: NCT00222573)*
> - *OMIM: Online Mendelian Inheritance in Man (OMIM ID: 601240)*
> - *MINT: Molecular INTeractions database (MINT ID: 6166710)*
> - *MI: EMBL-EBI IntAct database for Molecular Interactions (MI ID: 0218)*
> - *UniProt: Universal Protein Resource Knowledgebase (UniProt ID: Q9H0H5)*

Findability, interpretability and re-usability are best served if the database ensures links back from the data(sets) to the articles about that data. Elsevier put this in place in 2010 via a collaboration with earth data archive PANGAEA. Datasets deposited at PANGAEA are automatically linked to corresponding articles in Elsevier journals on its electronic platform ScienceDirect and vice versa. A single click brings the user from the data to the ScienceDirect article, or reversely from ScienceDirect to the underlying data at PANGAEA, by means of DOI's, both for the article and the dataset, (see Elsevier/PANGAEA press release (Elsevier 2010)[40].

Elsevier summarizes the process in 5 simple steps:

1. Author submits article to publisher
2. Author submits data set to repository
3. At article publication, repository links article DOI to associated data set DOI, creating actual connection
4. User sees link to ScienceDirect from PANGAEA
5. User sees link to PANGAEA from ScienceDirect

A few other databases and data-archives are also capable of providing links from the data to the corresponding articles, for example CCDC (Cambridge Crystallographic Data Centre)[41] and the PubChem Database[42]. These initiatives will make it much easier for authors to deposit the data in public archives as they can be ensured that future users of the data can easily find the corresponding articles.

Important intermediary services have entered the field lately that help facilitate the workflow of authors and publishers for the parallel submission of data to a repository and the manuscript to the publisher, ensuring the bi-directional linking between publications and data in public repositories to be in place. DataCite and Dryad are very

[40] Elsevier press release: http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01616

[41] CCDC http://www.ccdc.cam.ac.uk/

[42] PubChem http://pubchem.ncbi.nlm.nih.gov/

good examples further being explained in the next Chapter about Libraries and Data Centres. For most publishers, any system that uses DOI's as persistent identifiers can be incorporated in the workflow.

3.6. Data storage as a service by the journal

We know of at least one example where the journal offers storage facilities for primary research data to the authors of their articles. In 2010 Thieme established, for its chemistry journals, an easy process by which the authors can submit primary chemistry data with their articles. Thieme, based in Stuttgart, works together with FIZ Karlsruhe (who also host their publications platform) and the TIB National Technical Library in Hannover. The process consists of 5 easy steps:

1. At the same time with the article the author submits the research data to Thieme.
2. Thieme hosts the research data in a data center (FIZ Karlsruhe).
3. TIB assigns a DOI to the data.
4. At the same time the article is published the primary data are published as independent entity but in connection with the article.
5. The article quotes the research data as reference items with the assigned DOI.

As their motivation for this new initiative, Thieme says in their press release[43] (link http://www.thieme.de/SID-4BE1BD47-99107897/connect-en/tc_oct_06_09.html):

> *In the field of chemistry, (...) data is accumulated by a variety of analytical, spectroscopic or computer simulation methods. Thus far, the vast amount of data lies scattered on the computers of scientists, who have produced the information. As no central repository exists, no archival storage is possible at the moment. Scientific results are solely published in journals – but not the primary data from which those results originate. Due to the missing credit that working up such data currently receives, primary data is often poorly documented, difficult to access and not saved for the long term.*

Dr. Susanne Haak, Managing Editor and responsible for the chemistry journals at Thieme explains, "*Access to primary data is a fundamental condition for research work, particularly in the natural sciences." Therefore, Thieme and experts from TIB have developed a uniform structure for publishing primary data. Through structuring and central data registration, a Germany-wide unique service of TIB, valuable knowledge will be harnessed.*

Since its inception at the end of 2009, the Chemistry journals of Thieme had 13 articles with data files added, per article roughly of the size of 5 – 10 MB. The data is not touched by Thieme, who simply collect them in ZIPfiles and check if the ordering of the subfiles is logical in the context of the article. The data files cannot be included in the

[43] Thieme press release http://www.thieme.de/SID-4BE1BD47-99107897/connect-en/tc_oct_06_09.html

(pdf) peer-review files and are not sent to the reviewers unless they request the Editor to see them.

Thieme does not have the ambition to build up a database with chemistry information. The main aim is to provide individual datasets that support better understanding of the related research article. As Dr Susanne Haak explains: “The philosophy of a database is quite different from the collection of new, raw data. Everything that is entered into a searchable database needs to be carefully verified (PubChem is the best example) whereas raw data from experiments in chemistry may turn out to just help a chemist to understand what happened during a reaction.” She adds: “ CIF files that might be submitted to us as part of primary data have usually also been submitted to CCDC before as this is since decades THE database for crystallographic structure information”.

3.7. Articles with interactive data

In previous chapters we have explained the data problem as not just about availability and findability. It is also about interpretability and re-usability (see Chapter 1). Several publishers have undertaken initiatives to offer services around research data underlying the article that improve interpretability and sometimes even re-usability.

The Biochemical Journal[44], published by Portland Press offers dynamic pdf’s for its articles in their so-called Utopia Documents. These provide extra features that turn graphs into tables and tables into graphs from which the users can start viewing and using the data in their own fashion. These are available in excel sheets, so can be reused in other calculations.

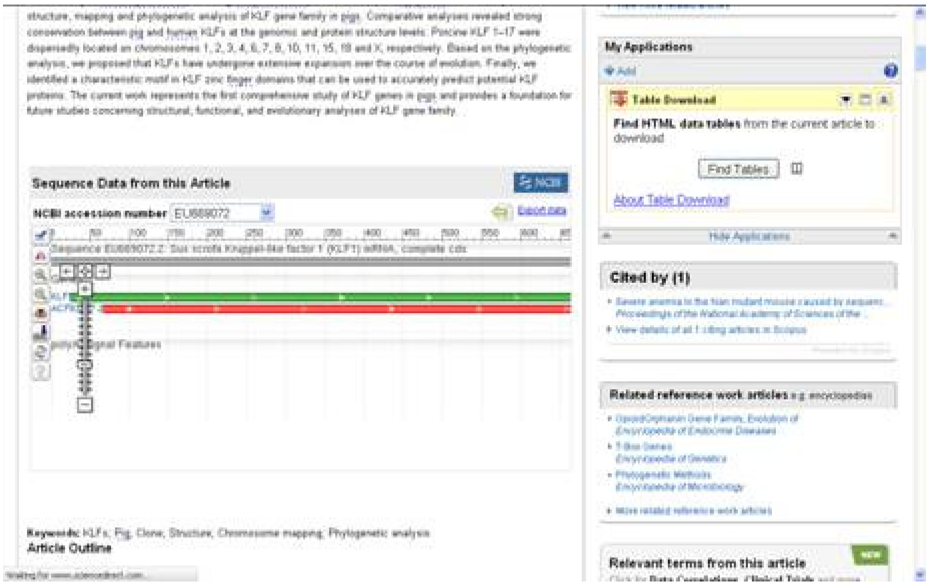
The **Optical Society of America** (OSA) together with the American NLM started a Interactive Science Publishing (ISP) project[45] in 2008 to enable authors to submit their data and figures to journals and to give editors and readers of their journals the possibility to view, analyze and interact with the source data connected with a scholarly article. Their main focus was the journal **Optics Express**. OSA supplies special software (from their ISP: Interactive Scholarly Publishing division[43]) that help the reader to apply all visualization features for the underlying material to the article. ISP allows authors to publish large 2-D and 3-D datasets with original source data that can be viewed and analyzed interactively by readers.

Elsevier provides in its articles a visualization and interaction applet for all related data that authors deposit in the PDB, the Worldwide Protein Databank. The app allows readers of the article to choose from several presentation formats to investigate the protein structures, in 2D or 3D, rotating or still. Elsevier emphasizes via this example that publishers need not be restrained in their offering of high-value added services for the analysis of data by the fact whether they store the data themselves. The data can just as easily be available in a public repository, available to all and available to applets to run on the data. In close collaboration with the NCBI, Elsevier offers a genome Viewer

[44] BioChemical Journal http://www.biochemj.org/bj/default.htm and http://www.biochemj.org/bj/424/bj4240317add.htm

[45] OSA-ISP project http://www.opticsinfobase.org/isp.cfm

for all gene sequence data deposited by the author in GenBank. This viewer can also be applied from within the Elsevier article[46]. See Illustration 1:



structure, mapping and phylogenetic analysis of KLF gene family in pigs. Comparative analyses revealed strong conservation between pig and human KLFs at the genomic and protein structure levels. Porcine KLF 1–17 were dispersedly located on chromosomes 1, 2, 3, 4, 6, 7, 8, 10, 11, 15, 18 and X, respectively. Based on the phylogenetic analysis, we proposed that KLFs have undergone extensive expansion over the course of evolution. Finally, we identified a characteristic motif in KLF zinc finger domains that can be used to accurately predict potential KLF proteins. The current work represents the first comprehensive study of KLF genes in pigs and provides a foundation for future studies concerning structural, functional, and evolutionary analyses of KLF gene family.

Sequence Data from this Article

NCBI accession number: EU669072

Sequence EU669072.2: Sus scrofa Kruppel-like factor 1 (KLF1) mRNA, complete cds

Keywords: KLFs; Pig; Clone; Structure; Chromosome mapping; Phylogenetic analysis

Article Outline

My Applications

Table Download

Find HTML data tables from the current article to download

Find Tables

About Table Download

Hide Applications

Cited by (1)

- Severe anemia in the Nan mutant mouse caused by sequence... Proceedings of the National Academy of Sciences of the...
- View details of all 1 citing articles in Scopus

Related reference work articles e.g. encyclopedias

- Opioid/Orphanin Gene Family, Evolution of
  Encyclopedia of Endocrine Diseases
- T-Box Genes
  Encyclopedia of Genetics
- Phylogenetic Methods
  Encyclopedia of Microbiology
- More related reference work articles

Relevant terms from this article

Illustration 1: Screen view of Gene-viewer on ScienceDirect, Elsevier[46]

3.8. Special Data Publications and Data Papers

With the advent of ever growing volumes of datasets, and the urge for more sharing and open availability of data, there have been numerous suggestions for a new phenomenon of journals specially dedicated to so-called data publications. We examine in this context one of the very first initiatives in this space, namely by the Journal of Earth Systems Science Data[47].

This journal, as it states on its website:

> *aims to establish a new subject of publication: to publish data according to the conventional fashion of publishing articles, applying the established principles of quality assessment through peer-review to datasets. The goals are to make datasets a reliable resource to build upon and to reward the authors by establishing priority and recognition through the impact of their articles.*

The journal sets as a very strict condition that the data described is deposited in a long term repository and lists several for which collaboration has been established.

Other criteria for the data sets described are:

- **Persistent Identifier**: The data sets have to have a unique and persistent identifiers, e.g. doi, ARK, etc.

[46] Example from the journal Genomics: http://dx.doi.org/10.1016/j.ygeno.2010.06.001

[47] ESSD: see http://www.earth-system-science-data.net/

- **Open Access**: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge.
- **Liberal Copyright**: Anyone must be free to copy, distribute, transmit and adapt the data sets as long as he/she is giving credit to the original authors (equivalent to the Creative Commons Attribution License).
- **Long-term Availability**: The repository has to meet the highest standards to guarantee a long-term availability of the data sets and a permanent access.

Since its launch in 2010, the journal has accepted around 30 articles in three volumes, most of which Special Issues. The papers describe the data, the planning, instrumentation and execution of experiments or collection of data. Any interpretation of data is outside the scope of its regular articles. Articles on methods describe nontrivial statistical and other methods employed, e.g. to filter, normalize or convert raw data to primary, published data, as well as nontrivial instrumentation or operational methods. Any comparison to other methods is out of scope of regular articles. The peer review is public and follows an open discussion format on their website.

The peer-review which checks on uniqueness, usefulness and completeness as well as quality, ensures that the data sets are:

- *at least plausible and contain no detectable problems;*
- *of sufficiently high quality and their limitations are clearly stated;*
- *open accessible (toll free), well annotated by standard metadata (e.g., ISO 19115) and available from a certified data center/repository;*
- *customary with regard to their format(s) and/or access protocol, however not proprietary ones (e.g., Open Geospatial Consortium standards), expected to be useable for the foreseeable future.*

The main aim of the journal appears to promote (re-)usability of research data:

> *The articles in this journal should enable the reviewer and the reader to review and use the data, respectively, with the least amount of effort. To this end, all necessary information should be presented through the article text and references in a concise manner and each article should publish as much data as possible. The aim is to minimize the overall workload of reviewers, e.g., by reviewing one instead of many articles, and to maximize the impact of each article.*
>
> *[...]It is clear that some of these quite abstract criteria may soon unfold to more (technically) specific ones, depending on the discipline or type of data. If necessary, the editors will try to make sure that more specific help for authors as well as for reviewers will be developed over time. (...)*
>
> *To help streamline the review process, a more formal list of criteria has been developed, which may serve as a checklist.*

In a personal interview conducted for this study, Hans Pfeiffenberger, one of the two Editors-in-Chief, explains how over 20 articles were part of special issues and so far less than 10 spontaneous submissions were received, mainly because authors are not yet accustomed to this new type of publication. From the peer review reports it is not completely clear how deep the reviewers have examined the data, but referees do tend to look at methodology and value of the data. The main task of the Editor is to check if the data are deposited in a safe repository and are persistently accessible. He thinks there is still an educational task to perform on the way peer review can help check and improve the quality of data.

Pfeiffenberger also sees a need for better standards on the citation of data, especially which version of data to be cited: in its raw form, cleaned up or as gritted data, or only the final data product ready for re-use ? He quotes Oxford scholar David Shotton who advocates that data should be cited as: ‘first described in" and then link to a paper.

The journal has an open peer review process, making the comments by the reviewers available for further transparency of the journal's policies.

As well as ESSD, which started in 2009, more initiatives in this area have now been launched. One of them is the journal GigaScience[48] published by BioMedCentral. The journal, which opened for submissions in the summer of 2011 and works together with the Beijing Genomics Institute BGI, aims according to its website: " to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish 'big-data' studies from the entire spectrum of life and biomedical sciences. To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources".

It is likely that other journals will follow this example, possibly also in a hybrid way, including data-articles as a new article type for existing journals, in the way the International Journal of Robotics Research accepts Data Papers[49]: "Data papers are short (circa 4 pages) submissions that support and summarize a substantial archival data set which has itself been peer reviewed with the same diligence that regular submissions receive. The contribution is expected to be in the quality and utility of the data to the robotics community".

Similar discussions have already appeared on the blogs around PLoS.

3.9. Gap analysis.

In the previous paragraphs we have provided an overview of emerging practices for the integration of data and publications. Descending the Data Publication Pyramid, we find that data have always been part of the traditional literature in a much aggregated way. In the recent decade new extended forms of data presentation have found their way into online supplements to journal articles. Initially the absence of volume restrictions for added data was a blessing. Their recent proliferation has caused a halt and new

[48] GigaScience Journal http://www.gigasciencejournal.com/

[49] Data Papers in IJRR http://www.uk.sagepub.com/journalsProdDesc.nav?prodId=Journal201324&crossRegion=eur#tabview=manuscriptSubmission

limitations are being put in place to keep the data added to journal manuscripts relevant and manageable.

At the same time, certain disciplines show the emergence of community endorsed data centres supported by scientific journals that include the archives' accession numbers for links from the publication, or even add interactive viewers within the article to study the data in the archive within the context of the article.

With more and more data deposited in archives, a new publication type has emerged, that of the Data Article.

Returning to to the particular problem of making research data conform to the list of 4 items of Chapter 1 in terms of:

- Availability
- Findability
- Interpretability
- Reusability

we can make a rough rating for the journal practices described here above against these 4 criteria (the greener the higher the overall rating is):

	Availability	Findability	Interpretability	Re-usability
Data presented within articles	+	+/-	++	-
Data in Journal Supplements, unrestricted	++	++	+++	+/-
Data in Journal Supplements, but restricted	+	++	+++	-
Data in public archives, linked to publications	+++	++++	+++	+++
Journals storing data	+	++	+++	++
Journals making data interactive	++	+++	++++	++++
Data Publications	++++	++++	++++	++++

Table 6: Rating the different ways to publish research data against the Treloar criteria

The most common current practice, which is to add data in supplementary files to journal articles, is clearly not the most ideal one if measured against these 4 criteria. Data deposited in public, community endorsed archives seems to have a better future, especially in terms of re-use and also for availability and findability. This is particularly true if they are accompanied by proper data publications that describe them and make them interpretable and re-usable. At the same time we know that only a few discipline areas have these public, community endorsed archives. Indeed, some of these archives appear to be threatened by cuts in government spending.

If we add the three additional criteria of Chapter 1, namely citability, curation and preservation, the ratings or each of the current practices are:

	Citability	Curation	Preservation
Data presented within articles	+++	+++	+++
Data in Journal Supplements, unrestricted	++++	+/-	+/-
Data in Journal Supplements, but restricted	++	++	+
Data in public archives, linked to publications	++++	++++	++++
Journals storing data	+++	+++	+++
Journals making data interactive	++++	+++	++++
Data Publications	++++	++++	++++

Table 7: Rating the different ways to publish research data against the criteria of citability, curation and preservation.

Curation and preservation of data in journal supplements is not always ensured by publishers as we see from the PARSE.Insight data and the specific requirements for specific formats of data, against an ever growing variety of data formats (including multimedia) asks for better solutions . Public repositories are better placed to deal with this issue. Bidirectional linking and data publications can ensure better citability of data. For the citability of data there are two aspects that play a role: the data as such should be citable (via DOI's, accession numbers of other persistent identifiers) but in addition, the people who created or generated the data deserve credit. An increasing number of data repositories provide citation means for the data itself (including the links to and from the data), but very few conventions exist for the way the people behind the data get citation counts and credits for the work. The best way to do that currently is following the traditional way people are cited, via a publication that describes the data or has the data included.

3.10. From raw data to processed data to data interpretations.

In 2007 a majority of the larger publishers in the STM arena undersigned the so-called Brussels declaration which states[50] :

> **Raw research data should be made freely available to all researchers.** Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars

In this context, it is important to emphasize that the statement concerns raw data, not processed data, or data presentations or the data-interpretations as presented in a journal article. Using the Data Publication Pyramid as presented in Chapter 1, the Brussels statement concerns the public posting of raw data, the base layers of the pyramid. Preferably this public posting is done in an aggregated way in community endorsed archives for specific subject fields. It can be expected that many scientific disciplines will see a growing need to have such common solutions available that allow

[50] http://www.stm-assoc.org/brussels-declaration/

interlinking with journal publications. Project Dryad is one such example, the establishment of DataCite in 2009 another one (both described in Chapter 4).

For the middle area in the pyramid, that of Data Selections, Processed Data and Data Representations, several options exist. These can be cared for in the context of journals or in the context of archives and databases, depending on the level of processing of the data. Traditionally these kinds of data have been included in supplementary files to journals and this custom is expected to grow further. Increasingly, we can expect more and more journal policies to raise the level of required data selection and processing (example: Cell) and no longer accept anything and everything in the supplemental materials but instead pose restrictions on volume and format along the criteria of relevancy and manageability.

For the apex in this pyramid, data interpretations are included in a publication; there are no indications of paradigm shifts. But it is likely that data and publications will integrate further, at different levels and in more novel ways. There will be new innovations in the way the data are displayed and presented that confirm certain analyses and conclusions. Examples given in this chapter concern data made interactive from within the article via protein viewers and genome viewers.

3.11. Diverging and Converging Trends.

We see diverging trends taking place as well as converging trends in the way publishers are handling the increasing amount of data alongside articles. This is probably a strong indication that this area is in transition at the moment.

A clear example of diverging trends can be found in:

- Journals have more and more data submitted in supplementary files, and most of the journals accommodate this, also for a growing variety of file formats. At the same time, the first few examples now exist where journals could no longer handle this flow in view of the sheer volume and hence have stopped accepting supplementary files or have put limitations for them.

Whereas a converging trend is emerging in these areas:

- More journals support the principles of data sharing and data availability and press (or even mandate) authors to deposit data in public archives and to follow the conventions for this in their subject area.
- More publishers collaborate with community endorsed, public archives to make data and publications inter-linkable and citable, thereby endorsing the Brussels declaration in practice, and with positive effects on the integration of data and publications, their findability, discoverability, interpretability and re-use.
- A growing number of publishers offer services to present data in more sophisticated and even interactive ways, that increase their interpretability and hence their re-use further.

3.12. Opportunities for publishers in Data Exchange

From the practices and laudable initiatives gathered and analyzed in this research study, we can summarize the following elements as important opportunities for publishers to further improve the integration of data and publications.

- *Require availability of underlying research material as an editorial policy (example: Nature, PLoS)*
- *More careful treatment of digital research data submitted to journals and ensure it is stored, curated and preserved in trustworthy places (several examples of collaboration with community endorsed repositories)*
- *Ensure (bi-directional) links and persistent identifiers (examples for listed public archives, DataCite, Dryad)*
- *Establish uniform citation practices (examples Elsevier-PANGAEA, ESSD, DataCite, Dryad, Thieme)*
- *Establish common practice for peer review of data (example ESSD)*
- *Develop data-publications and quality standards (example ESSD, GigaScience, IJ Robotics Research)*

In order to offset these points against the listed issues around data we create the following table:

Data Issue:	Publishers opportunity to help improve situation:
Availability	Articles with data provide richer content and higher usage Impose stricter editorial policies about availability of underlying data which is in line with general funder's trends Ensure data is stored in a safe place, preferably a public repository Be transparent about curation and preservation of submitted data
Findability	Ensure bi-directional links between data and publications Use of persistent identifiers such as DOI's Ensure common citation practices
Interpretability	Provide services around data such as viewer apps for underlying data within the article or interactive graphs, tables and images Data Publications
Re-usability	Interactive data from within articles Links to the relevant datasets, not just the database Data Publications
Citability	Establish uniform data citation standards Follow metadata standards for datasets Use of persistent identifiers such as DOI's Data Publications
Curation	Transparency about curation of submitted data Collaboration with public data archives
Preservation	Transparency about preservation of submitted data Collaboration with public data archives

Table 2 repeated: Data Opportunities for Publishers.

In general, we can say that publishers will tend to follow their authors' wishes. With the trend clearly towards researchers who share more and more data, funders who make this conditional, and libraries and archives working towards better accessibility and retrievability of data, publishers can play an important role in the integration of data and publications for the sake of better discoverability, interpretability and re-sue of research data.

4. DATA CENTRE AND LIBRARY PERSPECTIVE

This chapter describes how libraries and data centres respond to the increasing amount of data that is produced and available and how they support availability, findability, interpretability, and re-usability of data. As such we assume that libraries and data centres deal, or have to deal, with any level of data that researchers and/or publishers want to make available: selected data representations, data collections and structured databases, raw data and original data sets. Based on desk research, we describe the current practice and rationale for action in libraries and data centres. We elaborate on the implications of increasing data integration in publication workflows, and present exemplary data initiatives, in which libraries and data centres are involved. Contact persons of each initiative were addressed with key questions, and their responses informed the analysis. The chapter ends by highlighting gaps as well as opportunities.

4.1. Libraries and data centres as custodians of data

Libraries and data centres are traditionally positioned at opposite ends of the research lifecycle: Data centres help researchers collect and process their data, and libraries deal with the publications that result from research projects. Libraries also help arrange the input at the start a new research cycle: the search for publications as the basis for new research. With the convergence of data and publications, and interdependencies between data and journal publications, such traditional roles become blurred. Reports like "Riding the Wave"[51] recognise that the requirements of e-science and enhanced scientific publishing necessitate a comprehensive infrastructure for scientific information. Libraries and data centres have important, partly overlapping, but mostly complementary roles to fulfil.

To create an infrastructure that systematically supports such data publication scenarios, libraries and data centres must align, or create new common conventions in data description and identification, and balance the relation between disciplinary particularities and large-scale interoperability. In this process, libraries and data centres complement each other:

Research data centres can be best considered in this context as the experts in their respective research disciplines and in handling their discipline specific data. They are set up to support data creation and access. They provide research teams with storage space, and services around data creation as well as preservation, and they provide academics and other users with access to data files and with training and advice on how to use them. They are familiar with data protection and privacy regulations and with research ethic issues. They are well positioned to adopt new and emerging types of data in their discipline and increasingly sophisticated methods of record-linking and statistical matching. They are aware of data quality issues and existing disciplinary standards.

[51] E.g., 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 (2010). http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Libraries have been keepers of knowledge for hundreds and thousands (in the case of Library of Alexandria) of years. They are experts in categorizing fields of knowledge and in recording and cataloguing all relevant information about a particular publication, including provenance and information about its author. They have been collecting, organizing, describing, preserving and making available knowledge and information manifested in printed books and articles and they are ready to transfer this experience to new forms of (digital) collections.

Several symposia and publications bear witness that the libraries community is in a transition process, rethinking their role in an increasingly digital environment in general and in e-research in particular:

A study commissioned by the Research Information Network (RIN) in 2007 asked if data management was a job for academic librarians. The survey data provided mixed messages: "Many librarians see data curation as a natural extension of their current role, but there is also evidence of caution in terms of the curation of large-scale datasets linked to e-research."[52]

A symposium organized by the US Council on Library and Information Resources (CLIR) explored functions of the research library in a changing information landscape in 2008.[53] It came to the conclusion that libraries needed to engage in data management and data curation to reflect basic changes in how scholars work – in collaboration with faculty and publishers. Rick Luce suggested in the same publication that traditional library roles must be augmented by new capabilities, centred on collaborative, data-intensive information resources.[54]

The Association of European Research Libraries, LIBER, made "Scholarly Communication" one of its 5 strategic priorities 2009-2012.[55] The corresponding working group on e-science recently organised the workshop "Libraries and research data: exploring alternatives for services and partnerships" that met large interest (presently not published).

While the library system has evolved over the centuries, the data centre landscape is a relatively young one, with the first scientific data centres dating back to the mid 20th century (ex: US National Climatic Data Centre: 1951, World Data Centre: 1957/8). The data centre landscape is more fragmented than the library system. There are well-established disciplinary data centres in some data intensive research domains (see for example CESSDA member organisations for the Social Sciences, CERN computer centre for Particle Physics, the members of the International Virtual Observatory Alliance in

[52] RIN 2007: Researchers' Use of Academic Libraries and their Services. A report commissioned by the Research Information Network and the Consortium of Research Libraries (2007). http://www.rin.ac.uk/system/files/attachments/Researchers-libraries-services-report.pdf

[53] CLIR 2008: No Brief Candle: Reconceiving Research Libraries for the 21st Century. Council on Library and Information Resources (2008). http://www.clir.org/pubs/abstract/pub142abst.html

[54] Richard E. Luce: A New Value Equation Challenge: The Emergence of eResearch and Roles for Research Libraries. In: No Brief Candle: Reconceiving Research Libraries for the 21 st Century (2008). http://www.clir.org/pubs/reports/pub142/pub142.pdf

[55] http://www.libereurope.eu/committee/scholarly-communication

Astronomy or the said World Data Centres for geophysical data), but with few exceptions barely visible centres in the Humanities.

One noteworthy example of a publicly funded, national disciplinary data archive, with a vast collection of digital data in the Social Sciences and Humanities is the UK Data Archive (UKDA). It was established in 1967 by the UK Social Science Research Council, which has so far provided the long-term commitment of funds.

The Life Science community is characterized by numerous specialized data centres, e.g., GenBank at the US National Center for Biotechnical Information, the Worldwide Protein data Bank, Cambridge Crystallographic Data Centre and many more.

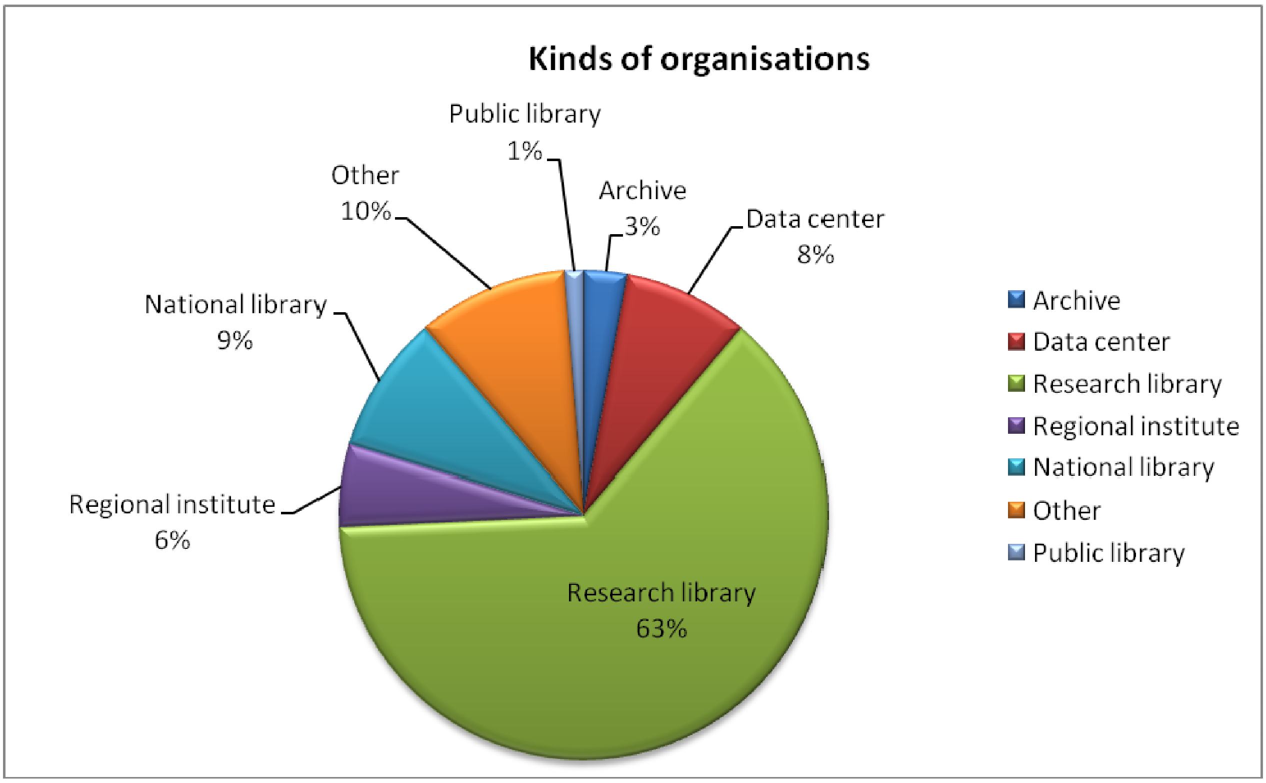
4.2. Common practice and rationale for action

Libraries and data centres serve the needs of the research community and in that role, must react to increasingly demanding user needs (see chapter 2 on researchers' perspective) and support increasingly sophisticated and complex publishers' products (see chapter 3 on publishers' perspective). In a society where information is available abundantly and often for free on the internet, libraries and data centres are under pressure to strengthen their role as professional information suppliers.

Another influencing factor is institutional requirements: Libraries and data centres are increasingly confronted with data management requirements from their funding bodies, involving them in data creation during the research workflow. Many academic institutes in a growing number of countries[56] have adopted Open Access policies which require their data centres to provide publishing support to the university's research groups. More and more funders oblige their grant recipients to make their data (openly) available after the end of a project (see also "Who pays for what" in chapter 2). There is a natural expectation that libraries and data centres will support the principal investigators with data management plans and to provide secure storage space for the created data.

A large evidence base on the common practice as well as on trends in the scientific infrastructure is available from the PARSE.Insight project. As far as libraries and data centres are concerned, however, the underlying figures must be interpreted carefully. The PARSE.Insight report never consistently defined the difference between data centres and libraries as we try to do for this report. The stakeholder group "data management" of the PARSE.Insight survey was composed of 7 archives, 20 data centres, 152 research libraries, 13 regional institutes, 24 national libraries and 3 institutions that identified themselves as "other" (see Graph 24).

[56] See ROARMAP, the Registry of Open Access Repositories Mandatory Archiving Policies: http://roarmap.eprints.org/



Graph 24, Source PARSE.Insight, N=241, background of 'Data Management' respondents

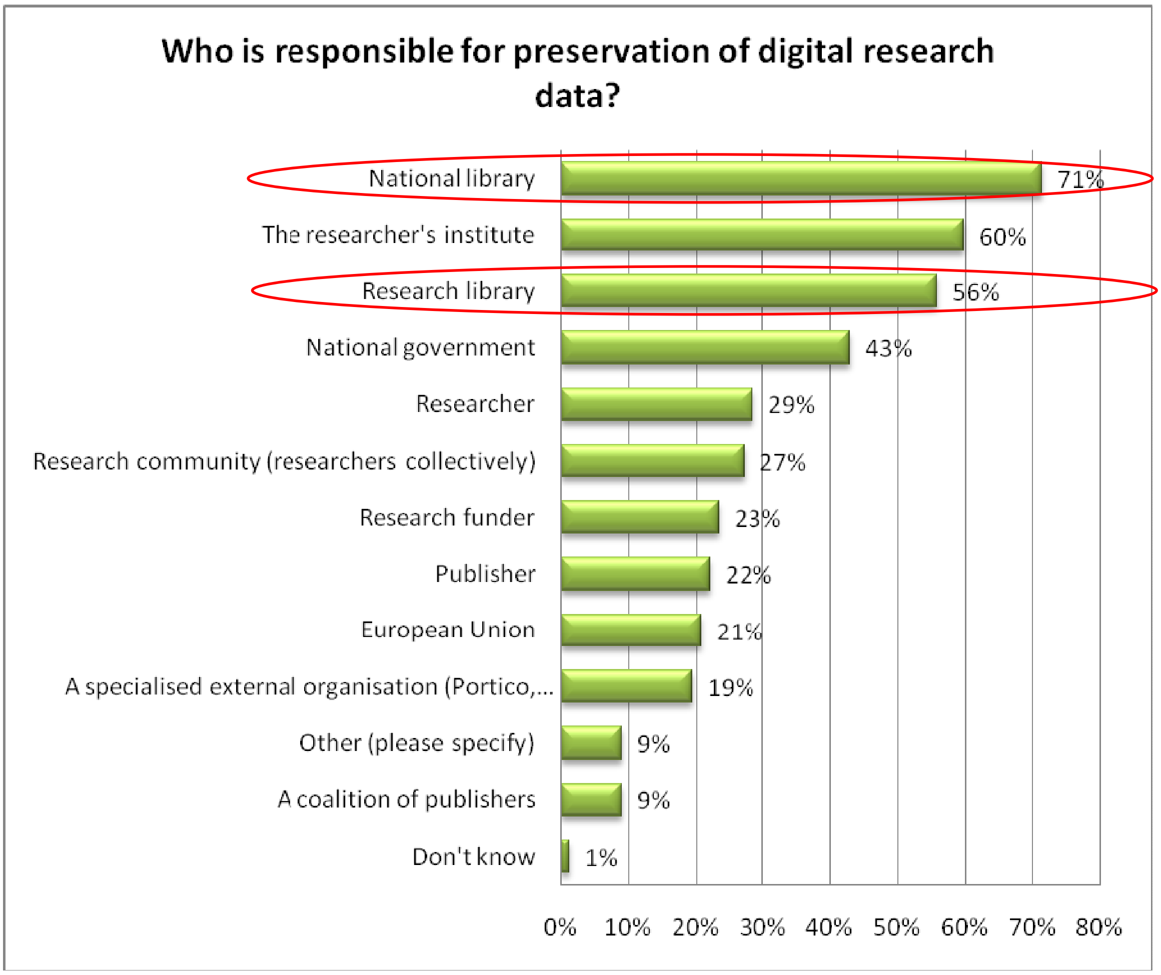
The questionnaire distributed in the stakeholder group "data management" is available online, as are Graphs of most survey results.[57] Here, a special selection of the findings most relevant to data sharing are reanalyzed. The participating libraries and data centres agreed predominantly that data preservation was important or very important for the following reasons:

- Publicly funded research output should be properly preserved (98%)
- Preserved data stimulates the advancement of science (96%)
- It allows for re-analysis of existing data (95%)

"Researchers want someone else to pay for data preservation!"

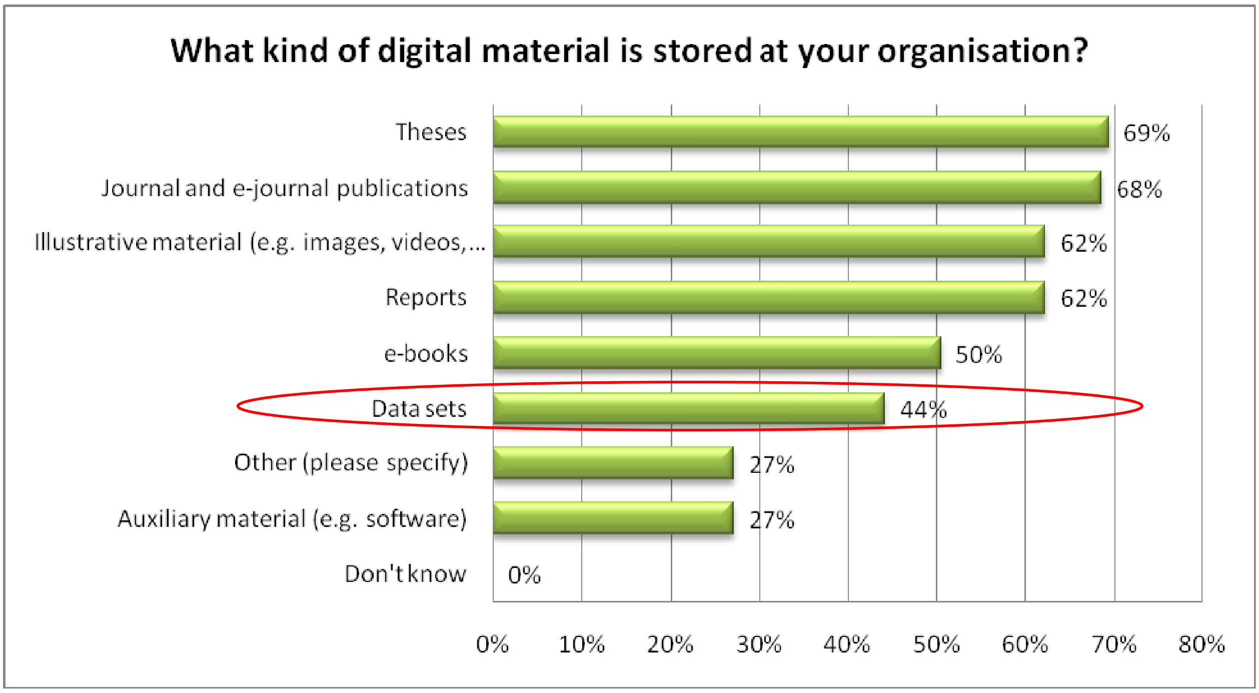
The PARSE.Insight survey showed that researchers look for an organisational structure to invest in data curation (see Graph 25, also chapter 2). In agreement with the high awareness for the importance of data preservation, especially libraries consider themselves responsible to fulfil this role:

[57] http://www.parse-insight.eu/downloads/PARSE-insight_survey_questions_datamanagement.pdf and https://www.swivel.com/people/1015959-PARSE-insight/group_assets/public



Graph 25, Source PARSE.Insight, N=241, 'Data Management' respondents, N = 77

However, in practice, only 44% of the responding institutions accept research data for storage and preservation (Graph 26).



Graph 26, Source PARSE.Insight, N=241, 'Data Management' respondents, N = 111

It is likely that the percentage is unevenly distributed between the group of participating data centres and that of the libraries. If 100% of the participating data centres and perhaps some or most of the archives accept research data for storage and preservation and only a very small proportion of the libraries actually accept research data, this result can either indicate a large gap on the side of the libraries or indicate a strategy of specialization and division of labour.

It is likely that the percentage is unevenly distributed between the group of participating data centres and that of the libraries. If 100% of the participating data centres and perhaps some or most of the archives accept research data for storage and preservation and only a very small proportion of the libraries actually accept research data, this result can either indicate a large gap on the side of the libraries or indicate a strategy of specialization and division of labour.

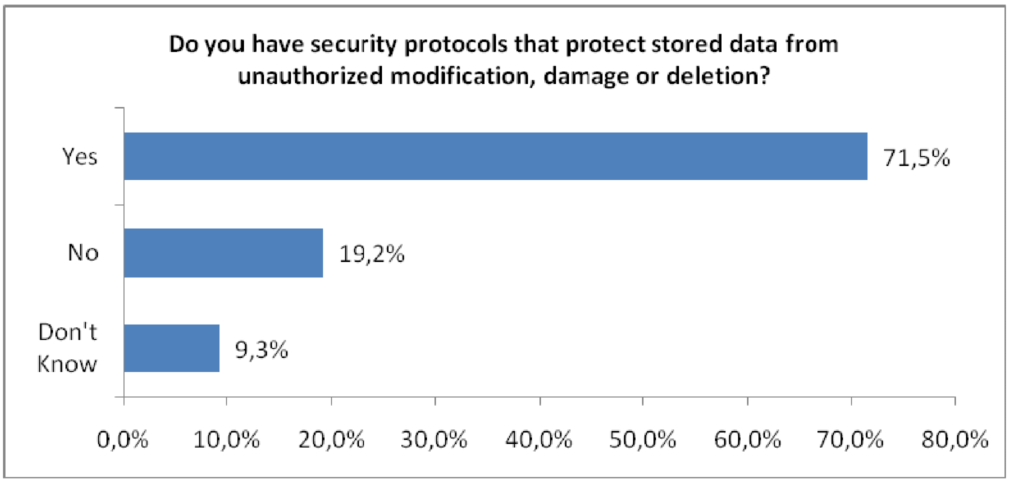
"Researchers want to be in control of their data!"

According to the findings of PARSE.Insight (see chapter 2) and to findings of the SURF foundation[58], it is of paramount importance for authors that they keep control of their data: "In all cases, when the data is transferred to another party, researchers wish to remain in control of their data." [59] Consequently, libraries and data centres, as keepers of authors' data must make sure that they respect this wish.

[58] SURF 2010: Martin Feijen: What researchers want. A review of literature describing what researchers want with regard to storage of and access to research data. Commissioned by the SURF Foundation (2010). http://www.surffoundation.nl/en/publicaties/Pages/Whatresearcherswant.aspx

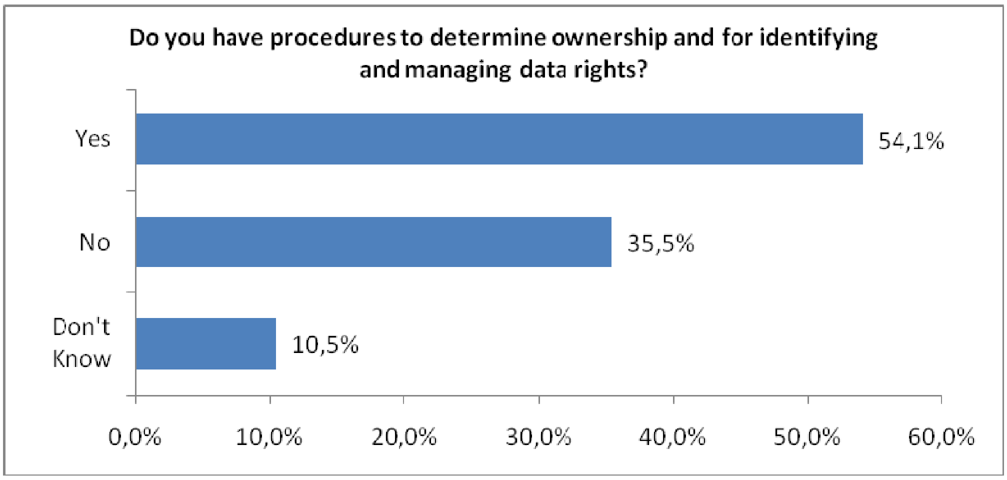
[59] SURF 2010

71.5% of the PARSE.Insight data management survey participants stated that they had security protocols in place that protect stored data from unauthorized modification, damage or deletion. In 19.2% of the participating organisations, action remains to be taken (Graph 27).



Graph 27, Source PARSE.Insight, N=241, 'Data Management' respondents, N = 172

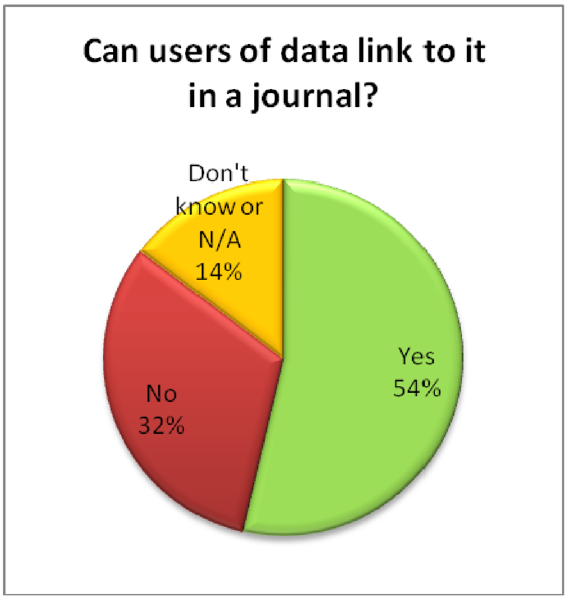
However, only 54,1% confirmed that they had procedures to determine ownership and for identifying and managing data rights – an important criteria for researchers to entrust an organisation with their data (Graph 28).



Graph 28, Source PARSE.Insight, N=241, 'Data Management' respondents, N = 77

"Researchers want credit for sharing their data!"

Another key topic for researchers is that *if* they make their data available, it should be visible and it should be possible to receive credit for it.



Graph 29, Source PARSE.Insight, N=241, 'Data Management' respondents, N = 164

At present, only 54% of the libraries and data centres support linking to stored data from journal articles (Graph 29). However, efforts are under way to facilitate exactly this. DataCite, for example, undertakes efforts to make research data citable and accessible in an internationally harmonized way[28].

The statistical findings suggest already that there is a high awareness in libraries and data centres, but not yet a comprehensive preparedness to take on the challenge.

4.3. Implications of data integration for libraries and data centres

There are good reasons for fostering the integration of research data and publications. As shown in the introduction of this report (see Chapter 1) integration of publications and research data has the potential to facilitate findability and re-usability of data, and to provide authors with better credits for their data. It also adds value and background to the publications.

Publishers offer several ways in which data and publications can be integrated already (see chapter 3). Data centres are, to a high degree, part of new publishing models, supporting data creation in the first place, or when publishers require authors to deposit underlying research data in public data archives and link to it. After all, many manifestations of data, as illustrated by the Data Publication Pyramid in the introductory chapter of this report (Graph 1) have an impact on libraries and data centres:

1. Data contained and explained within the article
   Implication for libraries/data centres: Prepare for adequate preservation strategies. The preservation of enriched articles may be more demanding than preservation of traditional articles. Novel ways of embedding data within the article (clickable graphs that provide underlying tables) will require more sophisticated preservation means.

2. Data published in supplementary files to articles
   Implication for libraries/data centres: Ensure that article and supplementary files stay together and that presentation and preservation mechanisms for supplementary files are in place.
3. Datasets referenced from the articles and held in data centres and repositories
   Implication for libraries/data centres: Distributed responsibility between article holder (publisher, library) and data holder (data centre or publisher). Datasets must be citable. The link between article and referenced data must be persistent. Presentation and preservation mechanisms must be ensured. If the data resides in a publisher's storage facility, perpetual access and eventually hand-off mechanisms to either libraries or data centres must be developed.
4. Data published independently from written publications, e.g. in databases or special data journals ("data publication")
   Implication for libraries/data centres: Support publication processes. Datasets need curation and special treatment that considers the granularity and dynamics of research data. Add metadata to datasets for documentation, to support re-use, and to facilitate search and retrieval of data.
5. Data in drawers and on disks at the institute
   Implication for libraries/data centres: Support scientists in preparing data management plans at an early stage of the research process to avoid "isolated data holdings" in the first place. Develop user friendly, low-threshold data publication or data deposit services.

To summarise, libraries and data centres must support data publishing as a prerequisite for data availability, including persistent identification/citation of datasets, and solutions for data description, documentation and retrieval, which together facilitate findability. They must also ensure long-term data archiving including data curation and preservation as a condition for data interpretability and re-usability. Libraries and data centres have started to enter into new alliances (as will be described in more detail in the next section) to develop new strategies together or with other actors such as publishers or research institutions. They can be involved on several levels, e.g. as active service operator, as provider of a specific sub-service (e.g. assignment of persistent identifiers), or as custodian of the results of such services.

4.4. Libraries and data centres engagement in new services and alliances

Durign the course of our research we found several new flagship projects and initiatives where libraries and data centres are probing the integration of data and publications on different levels. We present examples for persistent identifier and linking initiatives (findability – DataCite), data publication and data management support (Availability & interpretability – Dryad, Dataverse), and data archiving (re-usability – Pangaea).

In **DataCite**, libraries and data centres have allied to establish easier access to scientific research data online.[60] The goal of the international consortium is to make research data

[60] http://datacite.org/index.html

citable and accessible in a harmonized, interoperable and persistent way. 15 members from 10 countries – among them 6 Data centres or data services and 9 libraries – create and maintain an infrastructure to register data sets and assign unique persistent identifiers to them.

By using DOI names for data registration DataCite offers a simple and well known solution for citing data from publications. Users – researchers, publishers, libraries – can use the same technical infrastructure for datasets that they already use for research articles.

The focus of DataCite is the registration of data sets and assigning of persistent identifiers, not on storage and preservation of research data. The responsibility for the research data, including access, remains with the data centres or other trusted institution. The content holders are held responsible for quality assurance, metadata creation, storage and access. DataCite however provides supports, e.g. the DataCite Metadata Scheme[61], created by a working group of DataCite members. Also, a DataCite working group is defining criteria for trustworthy data centres with the rationale that stable descriptions of the duties and technical requirements for data centres which are using DOI names are needed.[62] With such criteria, DataCite would create an instrument that helps publishers and researchers trust that research data is stored in a reliable and persistent way.

One of the greatest achievements of DataCite is the inclusion of all relevant players in this arena: data centres, libraries, and publishers. DataCite acts on the assumption that progress in integrating data and publication can only be achieved by a joint effort of these stakeholders, combining strengths and influence of each of them. As the DataCite initiators confirm to us, one of their main short-term goals is to raise awareness in the editorial boards to allow referencing of datasets in publications. In the DataCite concept, libraries are not expected to act as data centres themselves, but continue to be a source of information for researchers. They should open their catalogues to scientific data and other content types and mediate access to data in data centres as remote content.

DataCite addresses data from science and technology alike and in principle spans all disciplines. The German Leibniz Institute for the Social Sciences (GESIS) runs a pilot DOI registration agency for social science research data based on the DataCite infrastructure.

DryadError! Bookmark not defined. is an example for an international repository that is committed to the "long tail" of data from the more decentralised biosciences, where data is not necessarily kept in large-scale repositories, or from under-financed fields such as the humanities:

Dryad is designed to preserve the underlying data reported in a paper at the time of publication, when there is the greatest incentive and the ability for authors to share their data. This is particularly important in the case of data for which a specialized

[61] http://datacite.org/schema/DataCite-MetadataKernel_v2.0.pdf

[62] Klump 2011: Jens Klump: Criteria for the Trustworthiness of Data Centres. D-Lib Magazine, Volume 17, Number 1/2 (2011). http://www.dlib.org/dlib/january11/klump/01klump.html

repository does not exist. Datasets are assigned with persistent identifiers to enable data citations.

Dryad was developed in the US by the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research Center, in coordination with journals and societies in evolutionary biology and ecology. The initiative resulted from a workshop dedicated to "Data Preservation, Sharing, and Discovery: Challenges for Small Science in the Digital Era" in May 2007. To journals and societies in these "smaller" sciences, Dryad offers a shared solution for data publication and archiving, thus relieving them from developing solutions on their own which would only lead to a fragmented landscape.

In the UK, there are attempts to integrate Dryad as a building block in the national scientific infrastructure. The initiative "Dryad UK"[63] is a 12 month JISC funded project and run by The British Library and Oxford University with in partnership with several associate organisations. The project aims at moving Dray to a sustainable business proposition and establishing a UK mirror of the Dryad repository. In order to reach this goal, a business models are being developed and publisher expansion is being promoted[64]. Cost recovery is a realistic and achievable goal but establishing the most appropriate model for this is the challenge of Dryad. For example, a authors may fund Drayd submissions from their research grants.

The Dryad UK initiators confirm to us that they have begun discussing potential business models together with publishers and funders. They point out the beginning integration of the Dryad repository with large publishers, e.g. PLoS and BiomedCentral for the first time, and new publisher workflow integration, allowing for peer review of the data behind an academic publication. They also acknowledge that initial resistance from some large commercial publishers, who are wary of imposing a system on all of their journals, or who still plan to explore commercial opportunities in the field themselves, could be overcome.

While DataCite and Dryad deal with readily produced datasets at the moment of publication, **Dataverse** is an example for an initiative offering data management support throughout the research life cycle, thereby preventing that data gets lost in disks and drawers in the first place:

The Dataverse Network[65] is an application to publish, share, reference, extract and analyze research data. It started as collaboration between the Harvard-MIT Data Center (now Institute for Quantitative Social Science) and the Harvard University Library and is presently implemented in Social Science disciplines. However, Dataverse collaborates with researchers and archives to expand the Dataverse Network as a data management and publishing framework beyond social science.

[63] http://datadryad.org/dryaduk

[64] See Beagrie et al: Business Models and Cost Estimation: Dryad Repository Case Study. Proceedings of iPres 2010. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf

[65] http://thedata.org/home

The open source Dataverse Network software allows for implementation of individual virtual archives, so called "dataverses". An institution can implement several of such dataverses and in this way distribute ownership between multiple researchers or research groups. That way, Dataverse addresses the desire of scientists to maintain control of their data, in particular to keep sovereignty over restrictions for their data sets (important for interpretablility and for re-use). At the same time, Dataverse offers a central repository infrastructure with support for professional archiving services such as back ups, recovery, and persistent identification.

A dataverse is designed to contain data organized in studies. Each study comprises the actual data files, complementary files, and metadata. Persistent identifiers are allocated to studies and can be used in publications to point to the respective evidence base. By allowing owners of the data to create persistent identifiers and a citation for their datasets even before public release of the data set, Dataverse accommodates the fear of authors that their data may be (mis-)interpreted by others before their own analyses are published. The owner of the Dataverse data can still use the persistent reference in an article, and then release the dataset once the article is published.

The Dataverse representatives point the data citation mechanism out as a major achievement of their initiative. The Dataverse Network software generates automatically a data citation for each data set published in a dataverse. Before the project, citations of data were inconsistent or nonexistent in many publications, which made data retrieval highly uncertain. Dataverse facilitates referencing data sets from publications in a standardized and persistent way.

Barriers are seen in lacking knowledge among researchers of the services that allow them to share data, in insufficient recognition and incentives to researchers for publishing their data sets, and in inconsistency between requirements from funding agencies or journals on publishing data, and providing insufficient funding to continue implementing and maintaining framework solutions.

Pangaea can be considered a representative initiative in the area of disciplinary data publishing and curation:

Pangaea[66] is an information service run by the German Alfred Wegener Institute for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM), also located in Germany. It focuses on archiving, publishing and distributing georeferenced data from earth system research.

Various international research projects use Pangaea as their data repository. Thereby, Pangaea addresses a need of publicly funded projects, which are often required to store primary data for a defined period after the end of the project. The German regulation, for example, sets forth that "Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin."[67] Because not all

[66] http://www.pangaea.de/

[67] Recommendations of the Commission on Professional Self Regulation in Science: Proposals for Safeguarding Good Scientific Practice (1998)

institutions can provide adequate storage capacity, let alone an infrastructure supporting description, persistent identification, search and retrieval of data, this task can be fulfilled by disciplinary data services.

Pangaea has developed and offers a range of archiving and publishing services. For example, it acts as publishing service and long term archive for the World Data Center for Marine Environmental Sciences. Pangaea aims at firmly establishing the concept of data publishing in the Marine Environmental Community within the next couple of years. Linking data to journal articles is an important part in this process and it happens in a bi-directional way: articles link to the data in Pangaea, and the Pangaea data point and link to the articles that use the data. Pangaea is already a designated archive for the Earth Science journals of Elsevier (see chapter 4) and is also the 'home' of the new data journal Earth System Science Data (ESSD).

The people responsible for Pangaea consider it the most important achievement of Pangaea that it has established collaborations with science publishers for cross referencing of science data and articles. DataCite DOIs are used for persistent identification of datasets. The Pangaea data publication process follows established publication processes with submission including a metadescription, formatting rules, abstract, archiving with lead time for proof-read, defining a citation, registration and final publication. Search and retrieval of the data sets via library catalogues is ensured through cooperation with the German National Library of Science and Technology

4.5. Gaps and dilemmas

The findings of the PARSE.Insight project, the initiatives presented in this chapter, and the sheer number of workshops, conferences and publications related to data management, data sharing, (open) access to data suggest that the need for action has been recognised in the library and data centres communities alike. They are actively involved in developing persistent identifier systems for research data, data citation standards, and solutions for data description, documentation and retrieval, as well as in data curation and preservation.

Availability

In terms of available infrastructure, plenty of solutions and possibilities are already available for the often mentioned problem of making research data available. There are vast possibilities for researchers to make their data available via institutional or disciplinary repositories, and increasingly together with publications. A challenge is that not all researchers are aware of the services available to them. Another challenge may be that the multitude of possibilities may create a fragmented landscape. Here, especially research libraries as information suppliers have an important role to play: They should engage with researchers to raise awareness for good data management

http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/self_regulation_98.pdf practices, the benefits of data sharing, and the options available in the different disciplines, but also acceptance for the best and most reliable services available for a specific discipline. As Martin Feijen highlights in a report by SURF 2010: "Researchers must be in control of what happens to their data, who has access to it, and under what conditions."[68]

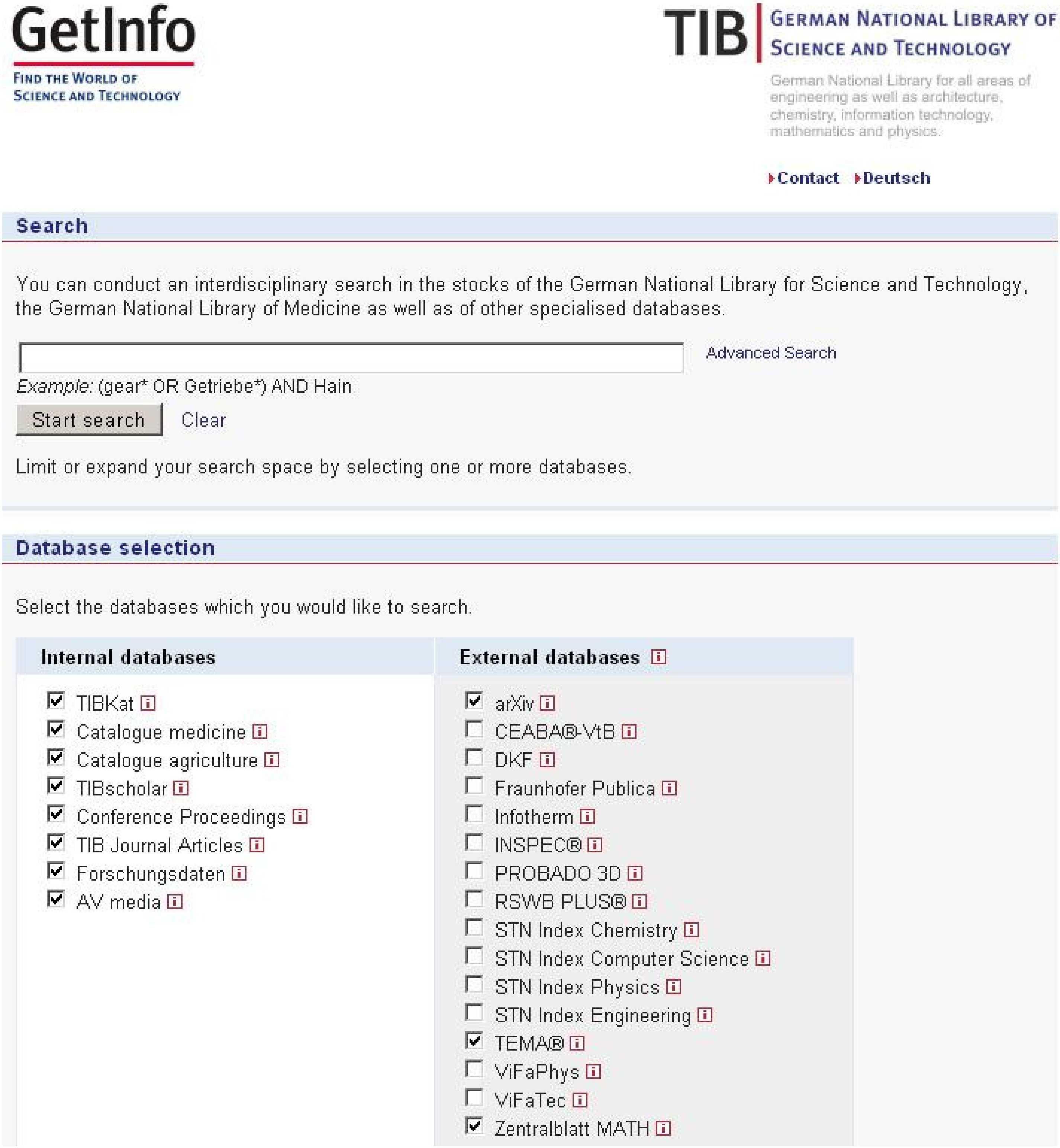
When addressing researchers as both data users and data creators, a division of roles and responsibilities seems suitable between libraries and data centres with libraries addressing researchers in general, of all professional levels and all kinds of disciplines, and data centres advising the professional users. On the service level, it seems suitable that libraries focus on data retrieval, cataloguing, registering ("retrieval focus"), and data centres focus on ensuring the long term availability of published data, including data storage, backups and replication in multiple locations ("data management focus").

Findability

A precondition for proper data retrieval is that good metadescriptions are added to the datasets and that links that lead to datasets either from metadescriptions or from publications are persistent. Metadata schemes must reflect the granularity of research data, the relationship between data sets, and the often frequent updating of datasets. With persistent identifier initiatives like DataCite or the Data Document Initiative in the Social Sciences[69], some research communities are on a good track for common solutions. The challenge lies in aligning progress in all scientific disciplines – not necessarily between disciplines, but at least within disciplines. Retrieval services only make sense if researchers and institutions within an institution adhere to certain common conventions. Most data centres provide data retrieval services for their own data base. Libraries that integrate data material in their catalogues to facilitate findability and access are still a rare exception. One of the rare examples is the "GetInfo" service of the German National Library of Science and Technology, where different internal and external databases can be searched at one time (Illustration 2).

[68] SURF 2010

[69] http://www.ddialliance.org/



GetInfo

FIND THE WORLD OF SCIENCE AND TECHNOLOGY

TIB | GERMAN NATIONAL LIBRARY OF SCIENCE AND TECHNOLOGY

German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics.

Search

You can conduct an interdisciplinary search in the stocks of the German National Library for Science and Technology, the German National Library of Medicine as well as of other specialised databases.

Example: (gear* OR Getriebe*) AND Hain

Limit or expand your search space by selecting one or more databases.

Database selection

Select the databases which you would like to search.

Internal databases

- ☑ TIBKat
- ☑ Catalogue medicine
- ☑ Catalogue agriculture
- ☑ TIBscholar
- ☑ Conference Proceedings
- ☑ TIB Journal Articles
- ☑ Forschungsdaten
- ☑ AV media

External databases

- ☑ arXiv
- □ CEABA®-VtB
- □ DKF
- □ Fraunhofer Publica
- □ Infotherm
- □ INSPEC®
- □ PROBADO 3D
- □ RSWB PLUS®
- □ STN Index Chemistry
- □ STN Index Computer Science
- □ STN Index Physics
- □ STN Index Engineering
- ☑ TEMA®
- □ ViFaPhys
- □ ViFaTec
- ☑ Zentralblatt MATH

Illustration 2: Screen view of the GetInfo service by TIB, Hannover, Germany.

Interpretability

To interpret research data created by others, good descriptions and documentation must be available (as for findability). The more documentation available, the easier it is for researchers to interpret other researchers' data. Documentation can range from descriptions of the data to so-called data publications all the way to (linking) the full publication using the data. Services like Pangaea that require researchers to submit metadescriptions with their data and adhere to certain formatting conventions (so that all datasets can be interpreted in a similar way) are a solid beginning. Crosslinks between articles and data are another means to support interpretability, because verbalized interpretation of the dataset in a publication helps the understanding of the original dataset. While links from articles to data become increasingly common, the other way around from data to articles is not yet so widely used, but good examples exist: e.g., Pangaea, PubChem and the Cambridge Crystallographic Database Centre. From a technical viewpoint, the interpretability of datasets can be ensured by separating them from vulnerable data carriers like CD-ROMs or DVDs and storing them on hard drives, including backups, forward migration and replications. Data centres seem to be best equipped to take on this challenge. In disciplines where there are no established data centres (yet), the universities institutional data centre, well equipped libraries, or library federations or initiatives like Dryad UK should stand in, although this may perpetuate the risk of fragmentation.

Re-usability

Ensuring re-usability is the most difficult goal of data management in a data centre and library setting. In addition to all the preconditions needed to ensure interpretability, re-usability often requires software to be available for analysing the datasets. The researcher who wants to re-use another researcher's dataset does not only need intellectual, discipline specific understanding of the available datasets, but also the skills to operate the appropriate software. Besides constant monitoring of the data holdings, libraries and data centres need to maintain format and software registries to plan for data preservation actions. First approaches to preservation of scientific data were for example, developed in the CASPAR project[70], and are followed up in the APARSEN network of excellence[71], but continued research is needed.

General dilemmas

Altogether, the many new initiatives in the area of data integration are promising. However, against the expected explosion of research data (see chapter 1 and 2) they are still more or less exceptional cases. There are a couple of pioneering libraries, often embedded in big and capable universities and involved in several initiatives at one time. The danger is that a few actors master the transition to a data-intensive scholarly information infrastructure well, and that the majority of stakeholders follow in a passive manner.

[70] http://www.casparpreserves.eu/

[71] http://www.alliancepermanentaccess.org/current-projects/aparsen

The results of the PARSE.Insight Gap Analysis of the "Scientific Libraries" community confirm this danger.[72] Although it was focused on preservation of digital research data, its tendency can easily be transferred to the wider area of data integration.

Overall, the PARSE.Insight Gap Analysis indicated a gap between better and less prepared libraries. This gap could be found in almost all analyzed areas:

The more data a library has to store, the better it considers itself prepared and responsible for digital preservation – to the point that funding for digital preservation will be an important issue for the libraries. Another relation was shown between the amount of data stored at a library and the implementation of data and access management strategies. Libraries with preservation and selection policies in place have smaller preservation gaps (in terms of preservation strategies implemented) than those who have not. Effectively, the PARSE.Insight Gap Analysis found those libraries which were slower in addressing the problem, were far behind the "early starters" in most categories.

Another, similar gap becomes apparent between disciplines: Influential, policy setting data centres do not exist in all scientific communities. A step in the right direction are two large programs implemented by the European Commission, CLARIN in the in the Linguistics and Humanities, and DARIAH in the Arts and Humanities.

Substantial and sustained funding is required to develop and market new services but often there is a tendency to address novel challenges in time-limited projects. While it is absolutely necessary to initiate a first step into action that way, it puts at the same time even promising results need to progress to sustainable services and this needs to be directed by libraries and data centres in partnership with their funding bodies[73].

4.6. Opportunities for libraries and data centres

The issue of data availability and data re-usability gets a lot of attention from users, funders, and decision makers[51,74,]. For libraries and data centres, this opens up the possibility to re-position themselves as complementary professional information providers in this field. In order to enable data availability, findability, interpretability and re-useability, libraries and data centres' data managers need to be involved from the very beginning of the research process in order to ensure high data quality.[75]

Datasets differ in many important ways from publications (e.g. granularity, iteration rate, data can be dynamic) and libraries must adjust to these new requirements as part of their new role. The immediate future for libraries will likely be characterised by

[72] PARSE 2010: PARSE.Insight. Deliverable D4.3. Gap Analysis Final Report (2010). http://www.parse-insight.eu/downloads/PARSE-Insight_D4-3_GapAnalysisFinalReport.pdf

[73] Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010).

[74] Prepublication data sharing", "Post-publication sharing of data and tools", Nature: Vol 461/10 September 2009

[75] Richard E. Luce (2008)

emerging and new roles in librarianship professions. Considering the initiatives presented in 4.4, it seems obvious that there is a need for “data librarians”. The implementation of the “Datasets Programme” from The British Library[76] illustrates action in this area. Libraries in the US like the University of North Carolina at Chapel Hill Library or Penn State University Library have already developed data management toolkits to support researchers at the proposal stage.[77]

In all their actions, libraries and data centres as institutions serve the research communities. Many studies point out that researchers prefer local and discipline specific data management support and want to retain control of the data until research is published.[78],[79] This balanced with the need to make data easily available and searchable suggesting there may be a role for libraries to act as an intermediary between researchers and larger data centres and in disciplines where there are no large data centres available, between researchers and the institutional repository. Overall, dialogue and interaction between the stakeholders is crucial and platforms to systematically enable it are desirable.

[76] http://www.bl.uk/datasets

[77] http://www.lib.unc.edu/reference/data_services/researchdatatoolkit/index.html, http://www.libraries.psu.edu/psul/scholar/datamanagement.html

[78] SURF 2010

[79] PARSE.Insight. Deliverable D3.6. Insight Report (2010). http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

Data Issue:	**Libraries opportunity to help improve situation:**	**Data centres opportunity to help improve situation:**
Availability	Accept datasets for storage at library and/or open up library catalogue to research data sets to allow access to data at least as remote content.	Be open and transparent and lower barriers to researchers to make their data available via discipline or institutional data centre.
Findability	Support of persistent identifiers. Engage in developing common metadescription schemas and common citation practices. Promote use of common standards and tools among researchers	
Interpretability	Provide metadescriptions to datasets. Support crosslinks between publications and datasets.	Help researchers understand metadescriptions of datasets. Establish and maintain knowledge base about data and their context.
Re-usability	Be transparent about conditions under which the data sets can be re-used (expert knowledge needed, software needed).	Curate and preserve datasets. Archive software needed for re-analysis of data.
Citability	Engage in establishing uniform data citation standards. Support and promote persistent identifiers.	
Curation/Preservation	Transparency about curation of submitted data. Collaboration with data creators and data centres. Promote good data management practice.	Transparency about curation of submitted data. Collaboration with data creators and libraries. Instruct researchers on discipline specific best practices in data creation (preservation formats, documentation of experiment,…)

Table 3 repeated: Data Opportunities for Libraries and Data Centers.

5. REPORT EPILOGUE: MAPPING THE ROAD AHEAD

5.1. Can Libraries and Data centres fill the missing link?

This report has shown that the key to ensuring the long term success of integrating data and publications is to ensure that the data are managed and preserved in such as way that they remain:

- Available
- Findable
- Interpretable
- Reusable
- Citable

These criteria are meaningful to, and act as incentives for, both researchers and publishers to engage in linking data to publications but neither group believe they can fulfil this role on their own. No one group has responsibility across the whole communication chain or has the resources to satisfy all of these criteria. So a key question we wish to raise here is; can librarians and data centres help fill this missing link ? To what extent do they have the existing relationships with researchers and/ or the related knowledge (including skills) to ensure that many of these criteria are met?

We have established that our three stakeholder groups of researchers, publishers and libraries and data centres have much to gain from embracing the integration of data and publications. There are several opportunities to be grasped for researchers, publishers, and libraries and data centres. It is clear from the researcher's perspective that making data available needs to be incentivised. Funding is one incentive, but it is important that researchers are credited for their data and that making data available will increase the author's visibility, in other words the data should be easily citable.

It is also clear that, whilst publishers, in principle, are open to further integration of data and publications, there are challenges associated with ensuring the quality and longevity of the data submitted. Accepting and storing data submitted in supplementary files can be hugely demanding on publishers' resources. There is also the question over what level of data, as illustrated in the 'Gray's Pyramid, see Graph 1, a publication should accept and in what manifestation.

There are exciting developments in publishing in relation to data, such as data-publications whose main aim is to describe available datasets. Publishers are investing in developing services to enrich publications with data and are doing so in collaboration with public archive services such as libraries and data centres.

Data centres play an important role in the long term storage of data and it can be surmised that libraries have a supporting role in this landscape, whether that be in supporting researchers in storing their data or ensuring that data remains available to and is discoverable by the end user when starting up new research projects. It is clear that, for libraries, a priority is to ensure that quality data can always be accessed easily by their users.

What is unclear is a definition of the roles that libraries should fill and what the incentives and barriers exist for researchers to work with libraries on data management. This lack of clarity calls for a dialogue to take place about why and how libraries should play a role in the integration of data and publications. There are gaps that need to be addressed as the preceding chapter identifies. By drawing on data from previous research and on real life examples it presents some arguments and outlines opportunities but a more complete picture may be drawn by engaging library professionals and researchers themselves in this dialogue.

5.2. What does the Data Publication Pyramid mean for roles and responsibilities ?

The Data Publication Pyramid has appeared throughout this report (see Graph 1). Every layer of the Data Publication Pyramid presents different challenges and calls for a variety of approaches to improve data exchange. As we descend the layers of the data pyramid the division of roles and responsibilities becomes less clear.

The top layer, publications with data, is well established and already an integral part of the record of science and the systems around it for discoverability, access and retrieval are well in place. The associated roles and responsibilities between researchers, libraries and publishers are well defined and delineated. But it only presents a tip of the iceberg of available data.

At this stage of the pyramid, there is limited potential for the reuse of data as the data is usually embedded in the article in an aggregated form. While this potential for re-use increases in the layer below that: Processed Data and Data representations, often presented in supplementary files to journal articles, the criteria of discoverability and longevity are less sure there. As a result, an increasing number of publishers encourage authors to submit their data to the third layer, that of Data Collections and Structured Databases, from where and to which the publications can link.

At this level, the Data Collections and Structured Databases layer, boundaries begin to blur and roles and responsibilities of libraries and data centres need to be better defined. This layer also offers the most potential for data exchange as, at this level, the data are at their most findable, reusable, and well curated and preserved.

The 'long tail' of data at the bottom of the pyramid, where data tends to remain in drawers and on disks of the institute, has been wholly the domain and responsibility of the researcher. That this data meets none of our criteria for data exchange is both a problem and an opportunity. As discussed in chapters 2 and 4, there are several reasons why this data has not been integrated. Lack of incentives, lack of local or discipline specific repositories, and fears over losing control over or credit for the data all contribute to this data remaining locked in their silos. Collaborative projects such as DataCite, Dryad and Dataverse are addressing these issues but more needs to be done to encourage researchers to start embracing data management and to make their data available.

5.3. In dialogue with Research Librarians: LIBER Workshop 2011

As research libraries are one of the key stakeholders in this report, ODE ran a workshop at the LIBER 2011 annual conference drawing on a draft of this report. Five speakers representing various backgrounds, including publishing, data centres, libraries, and research, were shown a copy of the draft report and asked to prepare a provocative statement related to their reading of the report.

Speakers and Provocations

"We need proper data citations in the reference section." Merce Crosas

Merce Crosas, Project Director of the Dataverse Network at Harvard, reminded the audience that, for the most part, we all agree that making data available is important, that we have important reasons for it (advancement of science, verifyability of research results etc.), and that nowadays, we have the technologies to realize it. So, she asked, shouldn't it really be only a small problem that we can easily solve? In her opinion, upgrading data citations in scholarly articles from in-text citations to full citations in the reference sections would be a big step forward in the right direction.

"Libraries must tackle the long tail data problem." Brian Hole

Brian Hole from DryadUK at The British Library stated that a major barrier to data exchange is the "long tail of data", which means that a large proportion of research data sits on researchers computers and doesn't get into a repository. These data are not made available for reuse and are almost inevitably lost. This is particularly true of Humanities data and Hole is convinced that there is a massive potential for re-use of these data.

Hole argued against the prescription that smaller libraries should leave data curation to large data centres. When no specific subject repositories for the long tail data, which is particularly true of the humanities, it simply isn't exploited, maintained or preserved. In Hole's opinion, (research) libraries are very well placed to bridge this gap because they are already placed within the researchers' workflow. At the very least they can educate researchers and provide data management plans. Beyond that, they may create and maintain their own repositories or act as advocates for the establishment of data repositories within their institutions.

“Clear and easy to understand citation metrics for datasets and automatics mechanisms to count them are urgently needed“ Maurits Van der Graaf

Maurits Van der Graaf, author of a SURF study on the quality of research data[80], advocates establishing a Dataset Impact Factor as an incentive for researchers to publish and properly cite datasets, and a Data Archive Impact Factor as a means to further professionalize data archive management. The Data Archive Impact Factor could help data archives to measure their relevance to scientific research, and help funders to evaluate the effectiveness of their support. Van der Graaf expects that this would influence publishers in developing additional services linking to high impact data archives.

“No Publication without Data – no data without publication” Eefke Smit

Eefke Smit, Director of Standards and Technology at the International Association of STM Publishers, referred to the Data Publication Pyramid of this report to illustrate that currently only a very small fraction of all data created gets ever published.

Smit argues for a change in practices because publication has a large potential to make data visible and re-usable. And in return data publication enhances the traditional publication by providing supporting evidence and background to the official Record of Science. Also, via citations it serves as a credit systems for the people behind the data.

Moreover, Smit calls for constructive collaboration throughout the information chain, where data is not necessarily kept at the publisher, but rather securely stored and preserved in certified, reliable data repositories, and via persistent identifiers remain linked to publications – bidirectionally.

“The research community needs to establish a common format for data acquisition and interpretation” Rick Luce

Rick Luce, Vice Provost and Director of Libraries for Emory University, drew the audience’s attention to the complexity of roles and responsibilities in the data landscape. No institution can soundly manage data over time when it doesn’t know what is allowed to do with it or how to treat the data correctly. Therefore, it is essential to clarify the question of ownership at the instant of data handovers.

Thereby, three principles should be observed:

- The data integrity principle: Ensure the integrity of research data to implement trust in research processes, enable researchers to verify published research results.

[80] http://www.dlib.org/dlib/january11/waaijers/01waaijers.html and http://www.surffoundation.nl/nl/publicaties/Documents/SURFshare_Over_kwaliteit_van_onderzoeksdata_dec2010DEF.pdf

- The data access and sharing principle: Make data that is integral to publicly reported results publicly available.
- The data stewardship principle: Provide proper data documentation, curation, and long term preservation to enable re-use.

5.4. Emerging Issues

Feedback from the workshop was recorded and five main issues or concerns emerged:

- **Citation Metrics**
- **Roles**
- **Why libraries?**
- **Approaches to data publishing**
- **Incentives**

Much of what came out of the workshop reflected the speakers' provocations and reaffirmed the relevance of the issues that have been highlighted throughout this report. The following are some the issues that were emphasised by participants throughout the session. Due to the nature of the workshop, some opinions are in conflict and reflect the lively debate in the workshop.

Citation Metrics

Measuring the impact of published data may be central to the success of data publications and measuring the impact of publications is a preoccupation of research libraries.

This is an area which requires further analysis. Journal Impact Factors, although established, have drawbacks; it may not be ideal to connect the impact of data to the impact of a journal, further it is important that whatever metrics are developed, they are simple and easy to interpret.

Roles

Unsurprisingly for a workshop with library professionals, the definition of roles was a major concern and drew many comments from the audience. Comments ranged from what roles Libraries should play in data management and exchange, to what new skill set library professionals need to meet the new challenges implied in these changing roles.

It was acknowledged that libraries need to align themselves with what is a dynamic research life cycle, becoming more project-oriented rather than providing services on an as needed basis. This reflects Rick Luce's (2008) assertion that libraries need to reposition themselves to become involved in research at the beginning of the research life cycle.

The majority of participants believed that libraries should have a role as repositories of data. Some traditional library skills are applicable to data management, such as collection management and information retrieval skills and the libraries established relationship with researchers as teachers of information literacy skills means that they are well placed to provide guidance to researchers in the creation of data management plans.

If libraries are to become more involved in the integration of data and publications then there may well be a necessity to develop software development skills to support the exploitation of data as well as expertise in the area of intellectual property (IP) in order to ensure that the rights of the creators of the data are protected.

The increase in the use of integrated data means that boundaries between libraries, data centres and publishing are blurring, possibly, as one participant stated, becoming one intellectual unit, and this has perhaps the greatest implication for future roles.

Why Libraries?

As a further provocation participants were asked why libraries should have a role in data exchange? This question not only served to draw out the rational for the repositioning of libraries but also helped highlight some of the barriers to library involvement in data exchange.

Reflective of an issue raised in chapter 2, researchers are not interested in archiving or curating data themselves and there is a real danger of fragmentation if these activities are left up to individual researchers. Libraries already have a track record in supporting researchers in their work.

This could present an argument for institutional repositories but on the other hand, there may be a danger of fragmentation if libraries become repositories of data where discipline specific international resources such as PANGAEA or GenBank are available and a preferable solution for researchers. Also, libraries' existing strengths lie in creating structure rather than storage.

Approaches to data publishing

Much of the dialogue in the workshop focused on the approach that should be taken to encourage, sustain and support data publishing and data integration. First of all, there was a general sentiment that there is a need for a definition of metadata for research data.

There is also a need to focus on interoperability rather than specific subject domains, regions or institutions. This could pose a problem for libraries as existing library information infrastructures are not always easy to make interoperable.

At the same time it is important to work with researchers and research groups at local and project level. As one participant put it "local effort can result in global solutions".

Incentives

Incentives for researchers to publish their data were discussed in chapter 2. From the research library perspective it is important that there is an incentive for researchers to work with libraries in making this data available.

Without a doubt citation and impact is a major incentive for researchers (hence the importance of citation metrics). In order for data to be widely cited it must be findable, reusable and citable. Librarians and libraries have the skills and resources to make this achievable.

Another incentive that libraries may provide (if they act as repositories) is the promise of sustainability; the fact that an institution is taking care of the data, preserving it and curating it. This is not just a boon to researchers but a valuable addition to research proposals.

5.5. The Next Step: Survey to Document Current and Project Future Roles

What is clear from the workshop analysis is that there are many questions that need to be answered. Research Libraries are keen to engage in a dialogue about data exchange and there exists awareness that there is a need to reposition library institutions in this changing landscape. The workshop also served to validate the issues raised in this report.

The report highlights an opportunity for collaboration. With the increasing demand from funders for researchers to make their data publicly available and the ensuing need for support in data management, and publishers supporting the principle of data sharing by signing up to the Brussels Declaration, libraries are faced with the opportunity to reposition themselves to become embedded in the research process.

In order to mine the potential of the bottom layer of the data pyramid it is important to understand how researchers can be supported in making data more available than is presently the case.

This report has provided evidence of the impact that data sharing and reuse has and can have on the scholarly communication chain and how important the integration of data and publications is.

The survey among research libraries and researchers within their institutes that follows this report during the fall of 2011, aims to clarify further the roles of stakeholders concerned by measuring their awareness and readiness for more responsibility in research data. It draws from the findings of this report and from what we have learned from the LIBER workshop. It aims to reveal how stakeholders' roles are changing. Perhaps more importantly, it presents information and guidance for the likely evolution of these roles -- to ensure the ongoing integrity of the scholarly record and for the creation of incentives for stakeholders to support this.

Better insight in existing strengths versus weaknesses and opportunities versus threats should help create the conditions to increase data publication activity and, ultimately, data sharing and reuse.