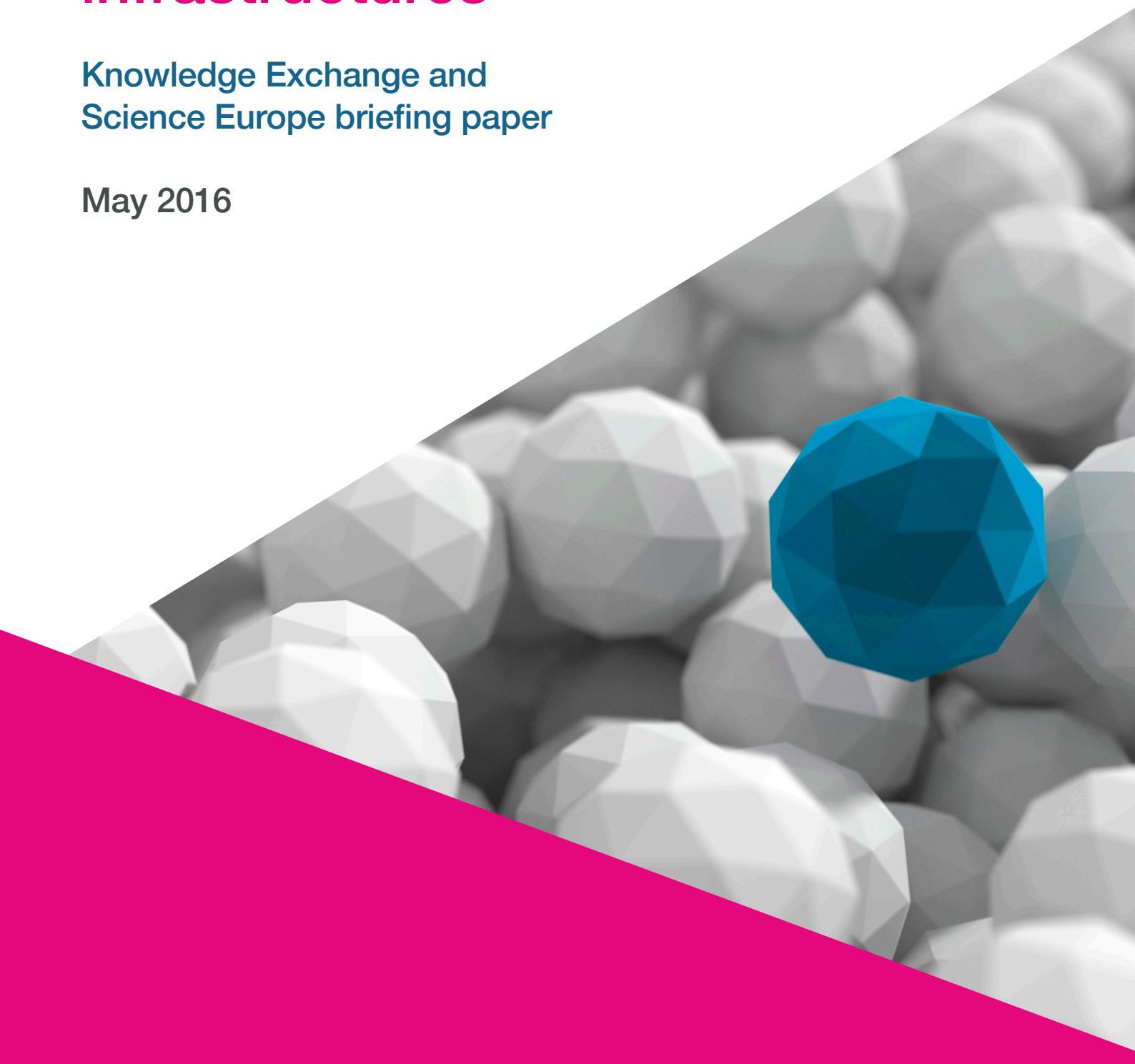
Knowledge Exchange

SCIENCE EUROPE
Shaping the future of research

# Funding research data management and related infrastructures

## Knowledge Exchange and Science Europe briefing paper

May 2016



**“Funding research data management and related infrastructures”**

**May 2016**

**Authors: Knowledge Exchange Research Data Expert Group and Science Europe Working Group on Research Data.**

**Editors and contributors:**

**For Knowledge Exchange:**
Magchiel Bijsterbosch (SURF, the Netherlands), Bas Cordewener (Jisc, UK), Daniela Duca (Jisc, UK), Matthias Katerbow (DFG, Germany), Irina Kupiainen (CSC, Finland), Ingrid Dillo (DANS, the Netherlands).

**For Science Europe:**
Peter Doorn (NWO/DANS, Netherlands), Harry Enke (Leibniz Association, Germany), Jesus Eugenio Marco de Lucas (CSIC, Spain).



# Contents

# Introduction

Responsible Research Data Management (RDM)[1] is a pillar of quality research. In practice good RDM requires the support of a well-functioning Research Data Infrastructure (RDI). One of the challenges the research community is facing is how to fund the management of research data and the required infrastructure.

The Science Europe Roadmap (2013) calls for the establishment of an ecosystem of research data infrastructures and for the design of appropriate funding structures adapted to national and organisational capabilities.[2] The Knowledge Exchange Annual Plan of 2014 prioritises work on sustainable business models to support the vision to realise an openly available layer of scholarly information – including aspects such as storage, preservation and curation, and not limited to scientific and scholarly publications, but also including research data, research tools, and related information (authority lists, identifiers, etc.).

Knowledge Exchange and Science Europe both defined activities to explore how RDM/RDI are, or can be, funded. Independently they each planned to survey users and providers of data services. On becoming aware of the similar objectives and approaches, the Science Europe Working Group on Research Data and the Knowledge Exchange Research Data expert group joined forces and devised a joint activity to collect information and produce a report to inform the discussion on the funding of RDM/RDI in Europe, to help raise awareness of the current challenges, and subsequently to communicate opportunities for coordinated action to relevant stakeholders. This briefing paper presents the results of the joint activity, detailing the approach and main outcomes of the study.

# Executive summary

1. Research Funding Organisations (RFO) and Research Performing Organisations (RPO) throughout Europe are well aware that science and scholarship increasingly depend on infrastructures supporting sustainable Research Data Management (RDM)[3]

2. In two complementary surveys, the Science Europe Working Group on Research Data and the Knowledge Exchange Research Data Expert Group explored how organisations funding and performing research think and act with respect to the funding of RDM and the related infrastructures. The resulting report illustrates the diversity of the funding landscape with respect to research data in Europe and the critical challenges that this presents. The funding of RDI, enabling RDM, comes from a great variety of sources and institutions that have different responsibilities and that operate at local, national and international levels. Significant parts of the funding have particular disciplinary dimensions. The funding actors, levels and disciplines are not part of a coordinated structure. This situation presents a huge challenge to the sustainability of RDM

3. Some RFOs and most RPOs contribute to the funding of specialised data infrastructure providers, which play key roles in providing RDI and in supporting RDM. Especially among RFOs there is no generally-accepted view on who should be responsible for the sustained funding of such providers; however, providers funded by RPOs tend to focus on servicing their own organisation. As a consequence, the infrastructure providers have different perspectives on their own and others' roles and responsibilities, which is a hindrance for effective (inter-)national and (inter-)disciplinary coordination. The many RDM services that these organisations fund, offer and use represent a wide variety, and all of these come in many flavours: local, national, international; discipline specific; and with all types of different, sometimes overlapping, beneficiaries

4. RDM, although recognised as important, is generally not (yet) regarded as a fundable part of the standard research process. The specifics of RDM and the budget scope for funding RDI are usually not clearly defined. The funding is not well connected to specific RDM requirements at different stages in the research process/data life cycle

5. Other studies and projects carried out during the past six years have identified that costs and funding of RDI/RDM need to be better defined and coordinated. Many principles and recommendations have been formulated. The fact that this survey and report reaches similar conclusions indicates that the problem persists

6. There are differences in the ways in which the various actors perceive their own and others' roles with regard to RDM and RDI. The funding mechanisms do not yet seem to be adapted to the shifting demands that are being made concerning the management, preservation and sharing of research data across borders, disciplines and beyond a particular organisation's interest

7. Sustainability of RDI/RDM is at risk as long as funding is project-based. Funding of existing RDI/RDM infrastructure needs to be reconsidered, business models for sustainable entities need to be developed, and responsibility for maintaining the data produced during projects (operations around curation, storage, archiving, sharing) needs to be defined and assigned. This requires more coordination, involving many actors, levels and disciplines. There is especially room for improving the coordination of funding mechanisms for RDI between the national and the European level

8. Irrespective of the business model and funding channels chosen, ultimately the money for data infrastructure originates from two sources: government funding and (depending on the discipline) private sector research funding, where the latter is rarely the core funding. An optimal balance between public and private funding sources, with agreed distribution of responsibilities and acknowledgement of common as well as specific ambitions, is urgently needed

9. When formulating policies with respect to the funding of RDM facilities, it makes sense to take into account the full research cycle and data lifecycle, and not just the phases of the actual research project. The challenge is mainly in the sustainability of the results *after* the funding of a project has ended

10. Given the diversity in Europe, a common vision, strategy and funding practice is not easy to accomplish. The increasing shift to an Open Science approach offers a good starting point for the layout of a layered, component-based RDI with complementary RDM support functions at various levels: international/national/local and mono/inter/multi-disciplinary, offering various types of RDI services (computing, storage, network, data, research support, training and education). There is a growing awareness that funding budgets need to be adapted to this situation, for instance by dedicating a certain percentage to RDI/RDM

11. Examples from outside Europe (for example, the National Science Foundation cyber infrastructure programme of Data Infrastructure Building Blocks in the US) may serve as an inspiration to make progress towards an RDI/RDM environment that optimally suits European Open Science ambitions

# 1 The joint Knowledge Exchange – Science Europe study

## 1.1 Background

Over the past decade, Research Data Management (RDM) policy and practice have been and still are on the rise. In a context where technological advancements facilitate data-intensive research, and where open access to and reproducibility of research results are increasingly called for, researchers are accountable for how data is treated before, during and after the research process. Research Data Management is undeniably part of good scientific practice; RDM implies specific tasks and responsibilities which require adequate support and provision to be properly undertaken by researchers.

The sustainability of RDM represents a challenge within the existing funding structures. At the core of this particular challenge lie issues related to the eligibility for funding of RDM activities during the project phase, and how the curation and long-term preservation of data after a project and its funding have ended can be paid for. Open Science implies, among other things, the optimal accessibility and sharing of research data, but without sustained Research Data Infrastructures (RDI) and services these activities are hardly feasible.

The roles and responsibilities for tackling this challenge are shared among various actors of the research system, ranging from Research Funding Organisations (RFOs), Research Performing Organisations (RPOs), universities, data infrastructures and services providers. These stakeholders are distributed along local, national and international dimensions. Moreover, the variety in the use, integration, combination and preservation of research data along disciplinary dividing lines and in different phases of the research cycle also demands to be taken into consideration.

## 1.2 Approach and methodology

The data in this study have been gathered via two online questionnaires (including both free text and multiple choice questions), follow-up interviews and four case studies.[4]

Key contacts in Science Europe Member Organisations (MOs) were invited to take part in a first survey and to provide the overarching view of their organisation. A total of 21 responses were submitted by 27 Science Europe MOs (the seven UK research councils submitted one single set of responses) from 17 countries.

In a second step, a complementary survey was circulated to 150 RPOs, universities, and research service and/or infrastructure providers, all based in the same 17 countries covered in the first survey. A total of 57 responses were received, of which 20% came from service and/or infrastructure providers.

On the limitations of the study, it should be noted that:

- The survey was addressed to pre-selected organisations and not to a random sample of organisations[5]
- Several statements represent the view of the surveyed or interviewed individuals, rather than stating the position of the individuals' organisations; in cases where people gave personal views it was not clear if they had an organisational strategy or not
- The geographical scope of the overall study is limited to the 17 European countries which were initially represented in the results of the first survey. Moreover, the number of responding organisations per country is imbalanced. The respondents were not a random sample, meaning that statistical generalisations cannot be made

- The matter of funding RDM and research data infrastructures and services is closely intertwined with the overall research (funding) processes. Many more actors than those organisations participating in the surveys and interviews play a role in these overall (funding) processes

- Finally, the variety of requirements concerning data infrastructure and services among disciplines makes it hard to formulate general statements that do justice to this variety

These limitations make it difficult to interpret and generalise the findings of this study. The conclusions of this report are therefore necessarily tentative. Nevertheless the results are a clear sign of the complexity of the RDM funding landscape in Europe and demonstrate that the long-term funding of data infrastructure, increasingly vital for science and scholarship, is by no means guaranteed.

# 2 Diversity of actors and services

## 2.1 Key actors and their (actual or perceived) roles

Local, national and cross-national organisations assume different roles within the process of funding and delivering RDM infrastructure and services:

- Several RFOs cover a number of eligible costs that are related to RDM, via research grants
- Some RFOs have specific budgets to fund (elements of) research (data) infrastructure; examples are the National Financing Initiative for Research Infrastructure (INFRASTRUKTUR) in Norway, or the fund for medium- and large-scale research infrastructure of the Hercules Foundation in Flanders, Belgium
- A number of RPOs and universities allocate part of their budgets for developing in-house data services or internal research data infrastructures; alternatively they outsource these to external service/ infrastructure providers
- National governments often provide indirect funding for research infrastructure activities, for instance via a national roadmap, often connected to the Roadmap of the European Strategic Forum for Research Infrastructures (ESFRI). These infrastructures are usually funded on a project basis (with a cycle of about five years)
- The European Commission runs schemes that allow the funding of cross-border activities in the RDM domain. Like national funding, this is usually on a project basis and related to the EU Framework Programme for Research and Innovation

The perception that each of these types of organisation tends to have about its own role and the roles of others in the RDM funding landscape can be described as follows:

- **Research Funding Organisations** contribute to policy development, and half of the responding RFOs implement measures to ensure that RDM-related goals are adhered to by relevant stakeholders. Most of these are 'soft', but some are strong measures, where funding allocation is conditioned by the compliance with the RDM policies. Fewer than half of the responding RFOs provide incentives for data sharing, and none of these incentives are financial. Most of the RFOs stress that responsibility for RDM lies with the researchers and their institutions
- **Research Performing Organisations** and universities ensure that curation and long-term preservation of the data and the metadata continues beyond the project funding period. Many RPOs set up their own RDM policies to comply with RFOs requirements (or benefit from RFOs incentives). RPOs would then allocate their budget in accordance in order to develop local data infrastructure or make use of existing (inter)national facilities
- **Research Data Infrastructure providers** develop and offer RDM-related services targeted to research groups or institutions. These providers can be discipline-specific or national, private or publicly-funded organisations, or they can be part of research libraries and research offices within universities. They tend to think that the responsibility for RDM lies with the researchers and their institutions

Various roles for different organisations are emerging within the RDM funding landscape. There is no effective coordination: neither among a given type of organisation nor across types; neither at national nor at local level. The international, disciplinary research infrastructures (such as CESSDA, the Consortium of European Social Science Data Archives)[6] display perhaps the highest degree of coordination, although the degree of coordination varies from discipline to discipline. Situations also vary considerably between one country and another and also between one organisation and another in a given country, but in general progress towards the establishment of national or regional research data infrastructure is slow.

## 2.2 Diversity of services, providers and beneficiaries

The providers of RDM-related services and infrastructures are diverse; they can be e-infrastructure providers (such as computing centres), libraries, archives or repositories, higher education institutions (universities) and other research institutions.

The range of provided services is wide; these can include data curation, long-term preservation, data storage and computing facilities. Only in very few cases do providers offer a broad portfolio of RDM and RDI services; more common is the provision of a combination of a small number of data services.

Training of and support for RDM is a service to researchers that is offered by a number of providers, sometimes in the context of projects, sometimes on a paid basis. The provision of training and support for RDM is a major challenge. It is probably necessary to have a combination of regular courses integrated into the curriculum (for students) and specific training courses (for staff) provided by data intermediaries (including university libraries) and data service providers.

Services are provided at different levels: many services focus on a particular field, others cover several disciplines or are more generic in nature. Moreover, there are local services for individual institutions national bodies and also international data services, in the context of pan-European research infrastructures.

Service providers either support all areas of research or specific fields, such as arts and humanities, health and social care, science, technology engineering and maths (STEM), and social sciences.

The beneficiaries of the services provided are mainly universities (as intermediaries) or the researchers themselves directly; sometimes they include heritage institutions such as museums or state archaeological services.

There is no clear insight into existing discipline-specific curation and archiving services. This may need further exploration, as many disciplinary research communities play an active role in research data activities (including among others: astronomy and astrophysics, genetics, biodiversity, high energy physics, Earth observation).

# 3 Core findings of the study

This chapter aims to illustrate the diversity in the responses received by highlighting the most important findings.

## 3.1 General challenges related to the funding of RDM and RDI

Although RDM is increasingly accepted as an essential part of good science/scholarly practice, and although the availability of high quality and sustainable research data infrastructure and services is generally acknowledged to be a condition *sine qua non* for all fields of scientific research, the overall and most prominent findings that the responses to the survey indicate are as follows.

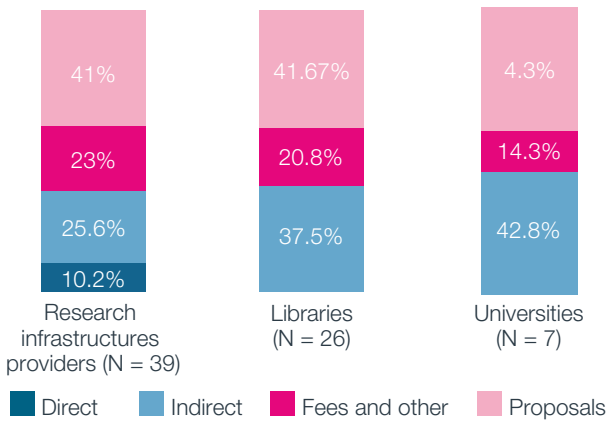
### 3.1.1 Challenge 1: Funding of RDM

RDM funding is generally not (yet) seen as a part of the standard research process, nor is it part of the normal research budget, and the specifics of RDM and the budget scope for funding data facilities are usually not clearly defined. The variety in the survey responses seems to indicate that this general situation is shifting, but the pace of this shift is highly variable across institutions and countries. There are initial movements in a few organisations to make RDM fundable explicitly. Few organisations were able to provide concrete budget figures for RDM and RDI, and insofar as numbers were given, they seemed unrealistically low. This supports the conclusion formulated here and in the illustrative text box below.

> *"There is often a mismatch between funding for short term projects (funded through commercial, government, agency, EC means) and expectations of long term archival – for which the only existing mechanism at present is national research council funding. This puts data generated through short term projects at risk."*
>
> **Research Infrastructure Provider, UK**

### Box A: Most funding for RDM is indirect

**Funding of RDM services and activities per type of organisations**




**Survey question to Research Infrastructure Providers, Libraries and Universities:**
'How are your RDM services or activities funded?'

**Possible responses:** directly (e.g. from research council, Government.); indirectly (e.g. overhead, project costs); fees from charging for your services; via research applications or proposals you make yourselves

**Interpretation:** For a very few research infrastructure providers, there is direct funding for RDM available, whereas for libraries, and more so for universities, financing RDM via indirect methods (overheads etc.) and via proposals is the major type of funding source.

### 3.1.2 Challenge 2: Budget allocation to RDM and RDI

As data collection, processing and analysis are usually part and parcel of research projects, these activities are funded on a project basis. However, the funding of stable research data infrastructures and long-term services is usually a separate matter, if they are eligible for funding at all. Although the data created in projects are increasingly required to be sustainable, in practice they are not, as funding for infrastructure and services is often not systematically planned or organised. Many respondents consider the drivers regarding the benefits and value to be unclear, especially in relation to 'who pays' and 'who benefits'.

> *"National requirement can only be supported if it would go hand in hand with a national archive and funding strategy."*
>
> Library, Austria

### 3.1.3 Challenge 3: Which data should be preserved and for how long?

The scope of preserving research data (drivers, objectives, terms, responsibilities), especially in the long term, has been explicitly defined only in a minority of cases (research fields/countries). This affects the funding of such services negatively. In spite of this, it is clear that there are several aims for preserving data, among which reuse in later or comparative research and replication of results are important ones. Variations in scope according to discipline may also occur here, for example caused by differences in intensity of reuse, by possibilities to reproduce data in another experiment, or by the historical and cultural value of the data. In some fields commercial exploitation of research data plays a role.

> *"Very important players in this game are the scientists themselves. Funding efforts may not be successful if they are not supported, accepted and used by the scientists. Therefore, it is absolutely vital to include them into all these processes."*
>
> **RFO**

### 3.1.4 Challenge 4: Roles and responsibilities towards funding

In the current situation the roles and responsibilities with regard to funding RDM and RDI, especially in the international context, are unclear. Many respondents from the surveyed organisations stress the need to have both complementary national and international funding opportunities.

> *"It's truly important to have both national and international funding for RDI."*
>
> **Higher Education Institution, Portugal**

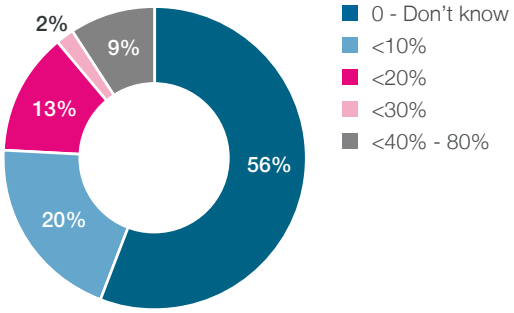
## Box B: Budget allocation to RDM and RDI unknown to many

**Survey question to Research Infrastructure Providers, Libraries and Universities:**
'What percentage of the total budget of your organisation is allocated for RDM and RDI?'

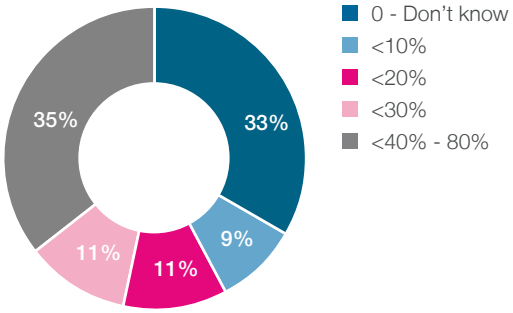
**Possible responses:** 0 - don't know, <10%, <20%, <30%, <40%, <50%, <60%, <70% or <80%

**Interpretation:** The lack of clarity about allocations to RDM and RDI is shown clearly: a substantial share of respondents cannot specify a budget allocation.

### Percentage of the total budget allocated for RDI




### Percentage of the total budget allocated for RDM



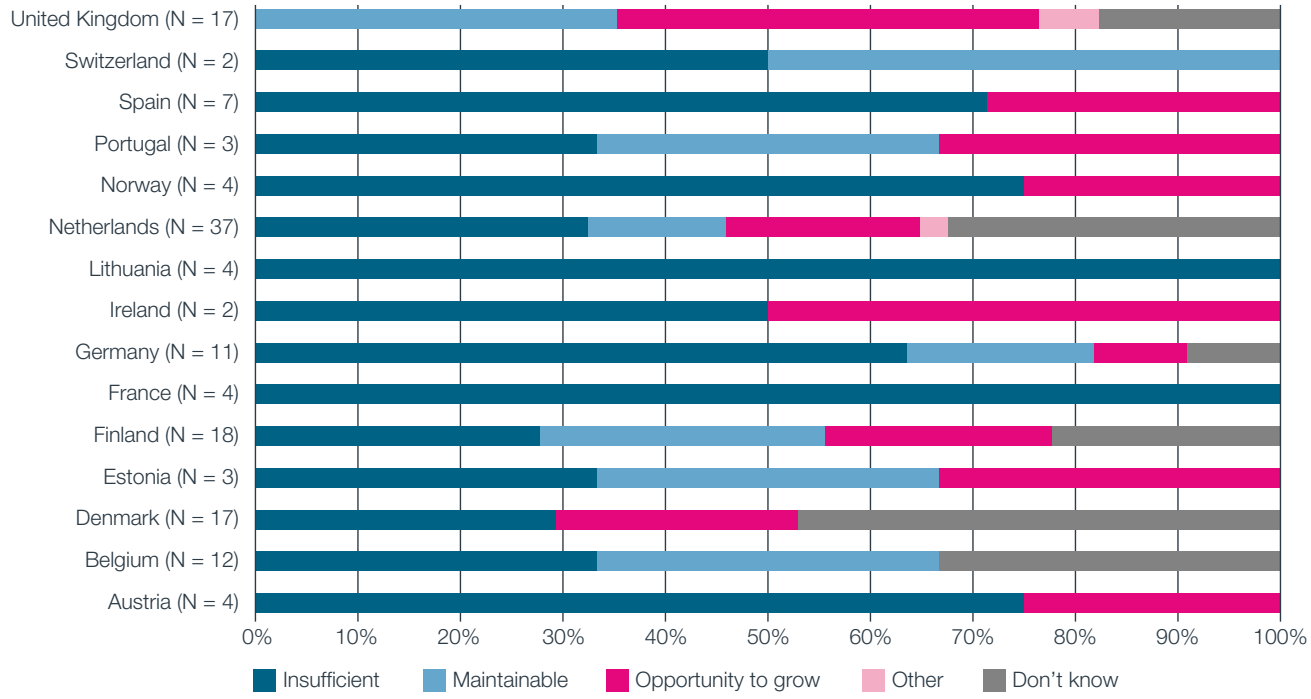
## Box C: Budget allocation to RDI perceived as insufficient by most

**Survey question to Research Infrastructure Providers, Libraries and Universities:** 'What is your organisation's view on the current funding of research data infrastructure at each of the following levels; 'disciplinary', 'local/community', 'national and international?'

**Possible responses:** 'insufficient', 'maintainable', 'opportunity to grow', 'other' and 'don't know'.

**Interpretation:** Although the approach is subjective, these figures seem to indicate that in almost all countries either the majority or a substantial number of respondents perceive that insufficient funds are available for the RDM tasks, or are insecure with respect to the situation.

### Current funding of RDI: views from RI providers, libraries and universities




The answers to factual questions about funding RDM/RDI and services, leads to high scores of 'I don't know' (see text box 'Budget allocation to RDM and RDI unknown to many'). It is clear that there is no systematic or structural allocation of funding, and if there is funding it is a combination of indirect funding from a variety of sources.

At the same time a majority of RPO and RFO representatives answer that the funding is thought to be insufficient for RDI. They may for instance witness a lack of – or limitations in – provision of RDM services.

## 3.2 Funding challenges related to different phases of the research cycle

The research process is often conceived of as a cycle. The collection, processing, analysis, presentation, publication, preservation and reuse of data are obviously closely connected to the different phases in this research cycle, and so are the requirements of RDM and RDI. Since the requirements change throughout the cycle, it makes sense to look at the funding needs and responsibilities in different phases.

The survey findings can be made more meaningful by relating them to these phases, especially because many respondents appeared to experience difficulty in identifying who is responsible for what. By looking at the different phases in the life cycle (= when), we get a clearer view on possible responsibilities, benefits and costs for the various stakeholders. In particular, the distinction between 'during the research process' and 'after a research (project) has been concluded' appears to be important from a funding point of view. When focusing on the funding aspects of the research data cycle, four phases can be distinguished:

**Phase 1.** The **actual research stage:** data creation/ collection (including potential reuse of already existing data), processing and analysis of the data. Responsibility for the data management is clearly with the researcher. The challenge for the researcher here is to calculate the costs to acquire, process and analyse the data accurately and in advance of the project starting. If budgeted well, data-related costs in this phase are usually funded via the research grant.

**Phase 2.** Immediately after the research phase: **deposit of the data in a designated storage solution** seeks to meet the preconditions for preservation, sharing and reuse. The amount of work and associated costs are usually specific to the collected data. Many RFOs now require a Data Management Plan (DMP) at the grant application stage. By working through the DMP sections and considering the specifics of data sharing and curation, researchers can better identify and justify additional budget requests relating to RDM in their applications. Few funding organisations consider the preparation and documentation for preservation to be eligible for funding via research grants. This depends on the extent to which RDM policies are in place and what they specify.

**Phase 3** concerns the first period of keeping the data available **for replication purposes, to check the validity of scientific claims** (including the possibility to investigate fraud). The institution to which the researcher belongs is usually considered responsible, and therefore the costs are paid by that institution, although one might argue that the responsibility should lie with the funder whose policy stipulates that the data is available for replication and /or reuse.

A growing number of codes of conduct and policies by research organisations require the preservation of research data for at least five or ten years, usually without stipulating who is to cover the costs for this. The funding responsibility is often not clear, also because the objectives of preservation change over time between this phase and the next. As long as the preservation is meant to enable replication, to check claims in publications, and possibly to investigate fraud, the institutions (RPOs) display a growing tendency to take responsibility for funding, as they have an interest in protecting their reputation.

**Phase 4. Continued preservation** after this initial period of five to ten years will also come at a cost, but who will gain from this long-term preservation? There is no (market) mechanism to bring costs and benefits together in identifiable responsible stakeholders (one cannot predict which stakeholders will gain from continued preservation and should contribute towards the incurred costs). In Open Science visions, research data is increasingly declared to be a 'public good' which is worthy of preservation, certainly as long as it represents scientific or socio-economic value; this poses new questions on selection and on who should pay for this long-term preservation.[7]

A proliferation of data repositories has sprung into being, at different levels (local, national, international), with a great variety of stakeholders (institutional, national, European research infrastructures and even commercial publishers), with various degrees of sustainability, and often with unknown guarantees for survival in the long run.[8] The main funding models are either structural funding or income earning through value added services, or a mix of these two.

A distinction could be made between situations where a direct demand for data reuse exists and situations where this is not immediately so, based on the idea that the user could pay for the reuse. However, this profitability principle may be in contradiction with open data principles; moreover, it is difficult to predict when and how often research data is going to be reused. For a minority of high-demand data a commercial model based on value-added services is possible, but for the majority of research data it is far from certain that this model will work.

# 4 Results of the joint Knowledge Exchange - Science Europe study in a wider context

There is a growing awareness of important and unsolved matters concerning financial aspects of research data, both from the cost perspective and from the funding side. In recent years, various initiatives have articulated recommendations concerning the costs and funding of services related to RDI and RDM; some of the resulting papers are presented in this chapter. Although not exhaustive, this overview is useful to put the results of the joint Science Europe – Knowledge Exchange in a wider context.

The **RECODE project** (2013–2015) formulated policy recommendations for Open Access to Research Data in Europe.[9] One of these recommendation (n° 2) states that funding bodies should:

> *"adopt a comprehensive approach in funding the implementation of open access to and preservation of research data. Appropriate financing and comprehensive planning is necessary for the following: collaborative and scalable infrastructures and services for access to and long-term preservation of research data; innovative actions that boost data re-use in the research and innovation sector; development of skills among researchers and information specialists, both formal (curriculum development) and informal (training activities). In achieving this comprehensive approach, they are encouraged to mobilize complementary funding instruments."*[10]

The **4C Project** (Collaboration to Clarify the Costs of Curation) aimed to help organisations across Europe to invest more effectively in digital curation and preservation. The project emphasises that the point of this investment is to realise a benefit. It provides an instrument to compare the curation costs across institutions, but also gives a number of valuable funding considerations, such as:

- *Make funding dependent on costing digital assets across the whole life cycle*
- *Design digital curation as a sustainable service*
- *Provide domain-wide shared infrastructures to exploit economies of scale*
- *Design funding constraints to ensure that sustainable digital curation is underpinned by proven cost-effectiveness*

Moreover the 4C Project:

> *"recommends that funds are not awarded to initiatives (e.g. research projects, development projects) that aren't able to give a plausible estimate of how much it will cost to sustain and make available the data they will be funded to create."*[11]

Funding data infrastructure and supporting services is also high on the agenda of the **Research Data Alliance (RDA)**, which aims to promote international cooperation and infrastructure required for scientific data-sharing. In the 2010 report, '**Riding the Wave**', a High Level Expert Group on Scientific Data recommends that additional funds be earmarked for scientific data infrastructure:[12]

> *"Development of e-infrastructure for scientific data will cost money, obviously – and as there is a significant element of public good in this, so there must be a significant degree of public support. [...] We call upon the European Council to expand the funding possibilities."*

The successor report, **'The Data Harvest'**, had similar recommendations in 2014:

> *"Much work is needed to develop the underlying infrastructure, identifiers, meta-data, systems and networks – and for that, again, public funding in Europe and international coordination by RDA will be needed".*[13]

The call in 'Riding the Wave' for a collaborative data infrastructure was responded to by the **Knowledge Exchange** paper '**A Surfboard for Riding the Wave: Towards a Four Country Action Programme on Research Data**'.[14] Based on the situation with regard to research data in Denmark, Germany, the Netherlands and the United Kingdom, it offers outlines for a possible coordinated action programme for the four countries in realising the envisaged collaborative data infrastructure, requiring the involvement of all stakeholders from the scientific community.

In 2013 Knowledge Exchange addressed issues relating to costs and value of research data in two workshops. The workshop report, **'The Price of Keeping Knowledge: Financial Streams for Digital Preservation'**,[15] showed that diverse sources of income are required to run a data centre and called for more value to be assigned to research data. In the Knowledge Exchange workshop report **'Making Data Count – Research Data and Research Assessment'**[16] RFOs are advised to set up policies, to implement well-defined mandates linked to grants encouraging data reuse, to observe development of practice and to provide funding in key areas.

The **Research Data Working Group of the League of European Research Libraries (LERU)** has formulated a Roadmap for Research Data.[17] Chapter 5 of this report is devoted to costs (pp. 24–27). Section 102 states:

> *"The revolution that data-driven science has initiated presents great challenges for a university and its finances [...] and this makes the identification of costs in supporting research data management of significant importance for university planning. Alternative funding sources could be the EU and individual research funders, although not all costs (such as recurrent staffing costs) would be considered as eligible costs by external funders. The extent to which research funders will fund the storage of, and access to, research data after the end of a project is also a factor to be taken into account when costing the construction and sustainability of research data infrastructures."*

In 2015 Research Councils UK published a **'Draft Concordat on Open Research Data'**, building on an earlier set of open data principles.[18] The first principle of the Concordat recognises the value of research data for high quality research as a facilitator of innovation and a safeguard of good research practice, and states that:

> *"Funders of Research will support open research data through the provision of appropriate resources as an acknowledged research cost."*

It is interesting to note that out of the 565 research infrastructures currently registered in **MERIL,** an inventory of openly accessible research infrastructures in Europe, 297 include the word 'data' in their description.[19] Moreover, of the 1476 data repositories registered in 65 countries (including two international categories) worldwide, 132 have 'research infrastructure' in their description.[20]

The **Science Europe** survey report **'Funding and Pan-European Cooperation for Research Infrastructures in Europe'** (January 2016) points in the same direction as the outcomes of this study:

> *"The landscape in Europe is diverse, with a range of approaches to issues such as the strategic priorities and the procedures used to define them, the funding of RIs and the exchange of information".*[21]

And:

> *"Funding instruments and procedures for RIs vary across the surveyed countries of the Science Europe Member Organisations, and sometimes it is difficult to classify these clearly."*

The **European Open Science Cloud** (EOSC) aims to provide an important shared infrastructure for research, including a data stewardship component. The report of the High Level Expert Group (HLEG) for the EOSC is being written in parallel to this report, and therefore it is not possible to quote it. It is, however, anticipated that its recommendations will be relevant for RFOs and RPOs throughout Europe.

The HLEG stresses that current funding mechanisms are biased towards the data-sparse and 'narrative' scientific publishing system of the past, while nowadays support for data publishing and tool sharing are required. Only proposals to develop infrastructure that include a sustainability plan for how it will persist should be eligible for funding. An overall average of 5% of the total project costs is seen as a reasonable estimate for the amount of funding required to sustain and share data and other non-traditional outputs.

The **OECD Global Science Forum (GSF) document 'Sustainable Business Models for Data Repositories'** builds on a study conducted by a Working Group co-sponsored by the Research Data Alliance (RDA) and the ICSU World Data System (WDS), initiated in September 2014. The group identified the most significant income streams of data repositories and developed a typology of the various business models encountered.[22] It was found that:

> *"Although many established national and international data repositories have reliable sources of income from research funders, these sources of income are generally inelastic and may be vulnerable (whether to short-termism, ill-considered re-prioritisation or attempts to pass responsibility to other budgets). Some data repositories are exploring means of diversifying their income streams to increase sustainability [...]"*

These include data deposit charges and selling curation and preservation services to various government, public and private institutions. Many data repositories are also substantially dependent on short-term project funding relating to research, training or infrastructure development activities. Some repositories charge for value-added services, and a number of others are exploring ways in which this can be done while conforming to Open Access principles.

The OECD GSF paper states that the funding models for data infrastructures and data repositories remain uncertain:

> *"As OECD Countries increasingly look to Open Science as a means to advance knowledge more rapidly, to increase the benefits and return of investment in research and to foster innovation, the sustainable funding of the data infrastructure necessary for Open Science needs to be addressed."*

In February 2016, the **ERAC Task Force on Open Access to Research Data** published a report recognising the complexity of the cost structure of open research data, distinguishing (1) overarching costs; (2) infrastructural costs; (3) handling costs; and (4) legal costs.[23]

The question whether research data should be made freely available without cost or if it is legitimate or justified to charge end-users for access to data is answered as such:

> *"We as a Task Force feel that, although it is debatable whether costs for depositing, (long term) preservation, value adding or other types of actions to make the data (better) reusable are justified, costs for access in itself does not fit the principle of open research data at this moment."*

The report also states that:

> *"Researchers must be sure costs will not be an obstacle or impediment to access data."*

> *"It should be assessed whether costs involved in realising open research data could be eligible in different funding schemes/for different funding organisations."*

> *"Supporting research data must not be looked at only from the perspective of costs since significant overall savings – considering the greater research context – could be achieved due to long-term preservation and preparation of data for further reusability."*

In April 2016, the **Amsterdam Call for Action on Open Science**[7] was released as the main result of the Conference on 'Open Science – From Vision to Action' hosted by the Netherlands' EU presidency on 4 and 5 April 2016. The underlying vision is – by 2020 – to reach "full open access for all scientific publications" and to adopt "a fundamentally new approach towards optimal reuse of research data". With regard to the development of research data infrastructures, the introduction of FAIR (Findable, Accessible, Interoperable and Reusable) and secure data principles and the setting up common e-infrastructures are called for.

# 5 Conclusions

Policy formulation among RFOs and RPOs with respect to RDM and access to research data is an ongoing concern, and is taking place at different paces among countries and organisations.

RPOs are taking responsibility more for the data produced in the institutes and centres that they encompass rather than formulating more general policies. RFOs are usually in a better position to formulate such general policies, because the funding instruments can be used as a mechanism to enforce RDM guidelines or Open Data principles. Exchange of practices can further inform the policy formulation processes, as a greater alignment on the matter is desirable.

However, not all funding organisations consider it their responsibility to formulate such policies, let alone feeling responsible for funding the consequences. The funding mechanisms simply do not yet seem adapted to the shifting demands that are being made concerning the management, preservation, and sharing of research data. Most funding mechanisms are geared to funding research on a project basis, whereas the services and infrastructure for data management and access require a good amount of permanence. Many grants and investments in data facilities require that the grantee guarantees continued access to the results of the project after it has been finished, which usually implies that the institution employing the researcher inherits this obligation. Obviously this means that (a) commitments are often made in a 'soft' way, the reality of which is difficult to check after several years have passed; and (b) where the commitment is met, the continued care for the data is paid from the (research) budget of the institution inheriting the obligation. And although the number of data repositories and other facilities and services is rapidly growing, the situation is far from transparent.

## 5.1. EU and national funding

The surveys essentially demonstrate that funding organisations in Europe think (and act) very differently about their responsibilities for the funding of data services and infrastructures. Therefore, RFOs and RPOs in Europe should (re)consider their position on the funding of data facilities. Moreover, there should be better and more effective coordination between national and international (European) funding mechanisms/schemes/responsibilities of such facilities.

The sustainability of research that is only funded on the basis of projects is low if this is not coupled with investment in adequate data infrastructure and services. Also, the majority of European investments in establishing infrastructures (e-infra, research and data infrastructure – EUDAT,[24] PASTEUR4OA,[25] OpenAIRE,[26] and many more) is on a project basis. It is hoped that those infrastructures that serve their purpose and audience well will become sustainable entities with sound business models, but this is by no means certain.

Proper criteria for the selection of infrastructures that are worth maintaining (and worthy of stable funding) do not exist. The implicit 'strategy', if it may be called a strategy, of "*let a hundred flowers bloom*" may result in a situation where a substantial proportion of the flowers will perish after some time. It is evident that not all flowers are perennial, but the garden of research data services and facilities obviously needs both watering and weeding.

It seems that the task of maintaining the data produced during projects (operations around curation, storage, archiving, sharing) will remain primarily a responsibility at the national and/or local level. However, strategies and policies with respect to research and data infrastructure are as yet fragmented at the national level, and there is no clarity about if, and how much of the research budget, national RFOs and RPOs are willing or need to invest in data infrastructure and services. The 5% 'data overhead' mentioned by the High Level Expert Group on the EOSC seems a reasonable starting point. Moreover, there is also a role here for research and data infrastructures at the EU and wider international level, and a certain amount of EU funding will also be necessary to complement the national and local funding.

An additional issue is the different size and requirements per discipline. With data volume and complexity increasing, the amount of expertise and labour required to keep the data (re)usable over time only adds to the problem. A solution, which will have to come primarily from national investments, will have to take such requirements into account. The balance of costs and benefits will probably have a disciplinary component.

Irrespective of the business model and funding channels chosen (such as lump sum, project, value added services, mixed), ultimately the money for data infrastructure originates from two sources: government funding and (depending on the discipline) private sector research funding, where the latter is rarely the core funding. Based on this situation, an optimal balance between the two, with agreed responsibilities and clear incentives, is urgently needed. The main issue at stake is the question of how far it is reasonable to allow privately co-sponsored data not to be openly shared. Various reports maintain it is reasonable for companies sponsoring research to expect that research data will be made openly available if doing so creates no significant commercial disadvantage to them. There is therefore a need to develop protocols on when and how data that may be commercially sensitive should be made openly accessible, taking account of the weight and nature of contributions to the funding of collaborative research projects, and providing an appropriate balance between openness and commercial incentives.[17]

## 5.2. The research cycle

When formulating policies with respect to the funding of data facilities, it makes sense to take into account the full research cycle. There is clearly a challenge in ensuring the sustainability of research results after project funding has ended. The question 'who pays' is a direct result of the responsibilities and the underlying motivations to preserve the data. The answer is different for the reproducibility in the short to medium term (formulated in codes of conduct) and for the availability of data for reuse in the long term.

Host institutions may be inclined to cater for the reproducibility of data, but, as long as data volumes continue to increase faster than storage costs per unit drop, the maintenance costs of data facilities tend to increase over time. The impact of these increasing costs on key functions of the host institutions (such as research and education) is unclear.

It is also unclear what 'guarantees' can be expected for the long-term availability of research data for further reuse. Although research institutions partly provide this essential function of the RDI, the safekeeping of research output is usually not explicitly part of their remit, and sustainable income to support these activities is not sufficiently secured through stable funding or additional revenues. Irrespective of how institutions are funded, for research it is critical that the research funding ecosystem will allow such infrastructures to be put in place.

Even when sustainability is formally required by a funding organisation, it is hard to check whether there is compliance with this requirement a couple years after the end of the project funding. It is not realistic to expect that the 'guarantees' for sustainability of data resources given by

research organisations hosting projects can be met if there is no follow-up funding, or when there is no data infrastructure in place that has the mission and reliable funding to ensure sustainability.

Research Funding Organisations may or may not consider RDM activities during active research to be eligible for funding. The Knowledge Exchange – Science Europe survey exercise indicates that RDM often seems to be funded indirectly, without any clear budget (see chapter 3). This adds to the already apparent problem of ensuring that data remain available after active research. For good RDM during the active research phases, and for ensured reusability of quality data at a later stage, RDM activities and resulting costs should be considered to be part of the costing breakdown in research funding programmes.

## 5.3. Possible ways forward

In the preceding chapters, both the surveys of this study and the findings in other reports on funding provide indications of what to do next. Moreover, it is also wise to look at approaches followed outside Europe, for example in the US, where the NSF has designed a programme of 'Data Infrastructure Building Blocks' to develop a robust and shared data-centric cyber infrastructure.[27] Such a strategy of components that should fit together like Lego bricks seems conceptually interesting. However, it requires a vision of an infrastructure that is expandable and modifiable over time.

It is clear that given the diversity in Europe, such a common vision – and related strategy and funding programmes – is not easy to accomplish. Perhaps the Open Science visions currently being formulated, both at the EU-level and nationally or locally, offer a good starting point for the layout of a layered, component-based infrastructure with complementary functions at various levels: international/national/local, mono/inter/multidisciplinary, type of service infrastructure (computing, storage, network, data, research).

Finally, any future activities for Science Europe, Knowledge Exchange or others on the theme of funding RDM and RDI should consider the joint active engagement of representatives from the various stakeholders involved in funding decisions relating to data infrastructures. This would include representatives of funding organisations and science policy makers, data repositories, research performing organisations, and the academic community.

# 6 Notes and reference

1 Research Data Management (RDM) is defined as the process, services and policies covering how the data used by or generated from research is organised, structures, stored, and cared for to ensure both its preservation and re-use.

2 The Science Europe Roadmap (2013) states: "It will be beneficial to the advancement of research, and ultimately to the European taxpayer, to address common issues in relevant policies and funding structures globally, or at least Europe-wide. Science Europe Member Organisations have already issued a number of general principles, policies and detailed requirements related to research data. They have proposed best practices related to data management, and have identified how the absence of such measures can lead to scientific misconduct. They fund and routinely operate elaborate data infrastructures in an increasing number of fields" (p. 9).
See: **http://scieur.org/roadmap**

3 Throughout this report, the acronyms RFO and RPO to indicate Research Funding and Performing Organisations will also occur regularly.

4 These four case studies were: 1: The Natural Environment Research Council's (NERC) Data Policy; 2: The National Financing Initiative for Research Infrastructure (INFRASTRUKTUR); 3: Nordic PIAAC Database; 4: A federated infrastructure - Research Data Netherlands (RDNL). The interviews were conducted in order to clarify some of the surveys' results.

5 This means that the survey results cannot be interpreted as a representative sample, and that the outcomes cannot be statistically generalised to the whole population of RFOs and RPOs.

6 **http://cessda.net/**

7 **http://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science**

8 This situation makes a quality hallmark for repositories (such as the Data Seal of Approval) desirable. It is ironic that even the sustainability of the hallmark itself is not guaranteed because of a lack of stable funding or a secure business model.

9 **http://recodeproject.eu/**

10 **http://policy.recodeproject.eu/assets/recode-funders.pdf**

11 **http://4cproject.eu/**

12 Riding the Wave: How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data A submission to the European Commission, October 2010, p.5.
**http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204**

13 The Data Harvest: How sharing research data can yield knowledge, jobs and growth. A Special Report by RDA Europe, 2014.
**http://www.e-nformation.ro/wp-content/uploads/2014/12/TheDataHarvestReport_-Final.pdf**

14 **http://www.knowledge-exchange.info/index.php/event/riding-the-wave**

15 **http://repository.jisc.ac.uk/6276/1/KE_Workshop_Report_-_Price_of_Keeping_Knowledge.pdf**

16 **http://repository.jisc.ac.uk/6275/1/KE_**

Workshop_report_-_Making_data_count_-_Research_Data_and_Research_Assessment.pdf

17 LERU Research Data Working Group. LERU Roadmap for Research Data. ADVICE PAPER no.14 - December 2013. http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

18 RCUK Open Data Principles (Version 10, July 2015). http://www.rcuk.ac.uk/research/opendata/

19 http://portal.meril.eu/

20 http://re3data.org/

21 Kas Maessen et al. Science Europe Working Group on Research Infrastructures, Strategic Priorities, **Funding and Pan-European Cooperation for Research Infrastructures in Europe**. January 2016. D/2016/13.324/1 http://scieur.org/rif-survey. Recommendations 7–12 and 15 in this report are directly relevant for the funding of data infrastructures as well.

22 Global Science Forum (prepared by Ingrid Dillo, Simon Hodson and Anita de Waard), '**Sustainable Business Models for Data Repositories**'. OECD Headquarters, 26–27 November 2015. DSTI/STP/MS(2015)13; the original report and more information on the RDA and ICSU/WDS working group can be found here: https://rd-alliance.org/groups/rdawds-publishing-data-cost-recovery-data-centres.html

23 ERAC stands for European Research Area and Innovation Committee (ERAC), a strategic policy advisory committee that advises the European Council, the European Commission and member states on the full spectrum of research and innovation issues in the framework of the governance of the European Research Area. ERAC Secretariat, ERAC Opinion on Open Research Data. Brussels, 3 February 2016. ERAC 1202/16. http://www.earto.eu/fileadmin/content/Website/ERAC_Opinion_on_Open_Research_Data.PDF

24 www.eudat.eu

25 www.pasteur4oa.eu

26 www.openaire.eu

27 More details can be found in the 2012 CIF21 vision, see: http://www.nsf.gov/cise/aci/cif21/CIF21Vision2012current.pdf

Knowledge Exchange is a collaboration between five national organisations, each responsible for supporting the development of ICT infrastructure for higher education and research. More information on its mission and activities is provided at: **knowledge-exchange.info**.

To contact Knowledge Exchange, email **office@knowledge-exchange.info**

Science Europe is a non-profit organisation based in Brussels representing major Research Funding and Performing Organisations across Europe. More information on its mission and activities is provided at: **scienceeurope.org**.

To contact Science Europe, email **office@scienceeurope.org**