Income Streams for Data Repositories

RDA-WDS Cost Recovery Interest Group: Ingrid Dillo, Simon Hodson, Anita de Waard

V. 1.00, 10 February 2016 - Final Report, published prior to RDA Plenary 7, Tokyo, Japan

Summary. Basic funding of data infrastructure may not keep pace with increasing costs. There is a need, therefore, to consider alternative cost recovery options and a diversification of revenue streams. In short: who will pay for public access to research data? The RDA/WDS Interest Group Publishing Data Cost Recovery for Data Centres aims to contribute to strategic thinking on cost recovery by conducting research to understand current and possible cost recovery strategies for data centres.

1. Introduction

1.1 Organisational background and history of the Interest Group

The RDA/WDS Interest Group Publishing Data Cost Recovery for Data Centres was recognised and endorsed by the RDA Council and TAB in 2014: https://rd-alliance.org/groups/rdawds-publishing-data-cost-recovery-data-centres.html
The Interest Group functions together with a number of other groups under the umbrella of the RDA/WDS Interest Group Publishing Data. The group currently has a membership of over 40 persons and is co-chaired by Simon Hodson, Executive Director of CODATA, Anita de Waard, VP Research Data Collaborations at Elsevier, and Ingrid Dillo, Deputy Director of DANS.

The original value proposition, definitions of scope, plan of work and deliverables for the group are laid out in a case statement [1]. During the course of activities focus of the group shifted a little. It was deemed most important to build up a view of the current landscape of funding or business models for data repositories. To do this we undertook a series of in-depth interviews with over twenty data repositories around the globe in order to develop a view of the funding landscape and the various income streams on which resourced their activities. We also enquired into any perceived or imminent changes in funding streams and what innovative funding models may be emerging.

The group held several sessions during the subsequent plenaries of the RDA that were all well attended. At the RDA Plenary 5 in San Diego in March 2015 the group presented the first preliminary results of its interviews and discussed these with representatives from data centres as well as funders. At Plenary 6 in Paris in September 2015 the group presented and discussed a first draft of this survey report. We also identified four common business models that had emerged from the survey and tested these with stakeholders using a SWOT analysis. This final report, containing the results of that analysis as well as conclusions and recommendations will be presented at the RDA Plenary 7 in Japan in March 2016.

1.2 The challenge: understanding and recovering costs

National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data. Data centres are faced with larger volumes of data, that also become more complex over time.[4]

With the volume and variety of data increasing, and budgets to manage these data unable to keep pace, investments in digital curation must be strategic and targeted to ensure the best value for money. Transparency of digital curation costs will help data centres identify greater efficiencies and pinpoint potential optimisations. Insight into how and why peers target their investments can lead to better use of resources, help identify weaknesses and drivers in current practices, and inspire innovations. [5, 6]

Although many established national and international data centres have reliable sources of income from research funders, these sources of income are generally inelastic and may be vulnerable to political and other changes in the landscape. There is concern that basic funding of data infrastructure may not keep pace with increasing costs. Some very large data services are finding that it is possible to manage and preserve exponentially increasing data volumes with a relatively stable budget due to a combination of strategically layered storage, well-timed support renewal and automation of certain processes. However, where the (human) curation costs per volume ratio is high such savings cannot be obtained.

There is clearly a need to improve our understanding of costs and where and how costs may be restrained. However, for sustainability, it is also important to explore alternative cost recovery options and a diversification of revenue streams. [6,7]

The Interest Group contributes to strategic thinking in the latter area by the research we are presenting here, which aims to help develop an understanding of current and possible future cost recovery strategies for data centres.

As the main body of work, the Interest Group has conducted a survey of over twenty data centres around the globe, and in different domains. The survey consisted of structured in-depth interviews with high-level representatives of these organisations and was focused on identifying various existing approaches to cost recovery, the range of income streams available and current and possible business models.

The central questions in the survey were:

- Which cost recovery models are currently being employed by data repositories?
- What trends are perceived by data repositories with regard to the vulnerability of funding?
- What are the possible responses to diversify income streams? What new options are funders and data repositories considering to diversify data repositories' income in the future?
- Have any of these alternative funding models been tried out?
- How have they been justified and 'sold' to stakeholders?
- To what extent are available models compatible with a commitment to Open Access to research data?

The principal beneficiaries of this work will be data centre managers on the one hand and research infrastructures on the other. Each of these stakeholders will gain a valuable insight into alternative options for cost recovery, substantiated by the survey results. Other stakeholders, including the users of data repositories and research performing organisations are also likely to value a transparent consideration of income streams and business models that can help maintain data infrastructure on a sustainable footing.

Some answers to the question: **"What - if anything - would you like to know about other repositories' business models and thoughts about future revenue streams?"**

"I would be very interested to examine other costing models. Where repositories are publicly funded this information could help us present a case for funding from the public purse."

"How do repositories satisfy long-term archival commitments, particularly when funding commitments can be short-term?"

"What kind of fees [are other repositories charging] for different services and what level are researchers willing to pay? Interested in the results of the survey for inspiration."

"What are other repositories dependent on federal funding agencies planning? Are there developments at policy level to shift costs of data curation to RPO/investigators/projects?"

"Would like to know how other repositories deal with larger datasets - when they are collecting larger amounts of data, how do they maintain that? How do you avoid contrasting principles of research transparency, avoid a lot of barriers with that?"

1.3 Overview of Our Approach

The purpose of our study is to understand the structure of data repositories' funding and to identify trends and options for their sustainable financing. The study provides a snapshot of the funding landscape, identifying the components of data repositories' income streams. It is intended that this overview will help data repositories, funders and other stakeholders consider the options available to put data repository funding on a sustainable footing. As an important caveat, our goal is not to quantify the costs of data stewardship. This is an important task but others are doing this, notably the EC funded 4C project [2].

Advocates of the importance and value of research data maintain that it is an essential part of the scholarly record and a part with considerable reuse value. They argue, additionally, that the costs of making data available and maintaining an excellent infrastructure for data management and stewardship are integral to the costs of doing research. Yet when confronted with the choice: 'more data infrastructure and less research or their contraries?' it is not clear how funders and researchers will respond. Funders are wary of underwriting an ongoing and growing commitment from a central budget.

Discussions of the sustainable funding of data repositories can be greatly advanced by unpacking the mechanisms by which data repositories are resourced. Much of the infrastructure that stewards research data that is produced as a result of public funding itself relies ultimately on government resources. However, the way in which a given data repository receives this money may affect its sustainability.

This study was motivated by the concerned hypothesis that many data repositories essentially rely on a block grant (structural funding) from a single research or research infrastructure funder and that such structural funding would not keep pace with the increasing proportional cost of data stewardship and may even be squeezed as research budgets tighten. If data repositories are under pressure to increase and diversify their income, what options exist to do this? Are there any funding mechanisms which provide scalable additional income? Conversely, might it be that for funders and data repositories alike, structural funding is actually the most straightforward and responsive means of maintaining a data infrastructure?

Our purpose is primarily to cast light on the current breakdown of funding streams. Secondarily, through discussion with stakeholders, we aim to understand which out the current models might be most suited to ensuring the sustainability of an adequate data infrastructure.

Some answers[1] to the question: "**Do you expect the current funding streams to be adequate for the tasks your repository will need to perform in the future?**":

""If they [government funding agency] continue to fund the infrastructure, sure. But I don't expect that will be happening."

"[We] need to step up from outsourced capacity to in-house capacity and this requires a drastic change in funding scheme. Stakeholder and data volumes are growing rapidly and funding [is] not following. The current model is not scalable to support a doubling of stakeholders and data…"

"The cost of long term preservation is now only covered for the first five years of preservation. More data and thus higher costs are expected in the near future. Demands and requirements will grow."

[1] Throughout this report, we will refer to some verbatim answers to the questions posed during the interviews, which are indicated in boxes and in this green font.

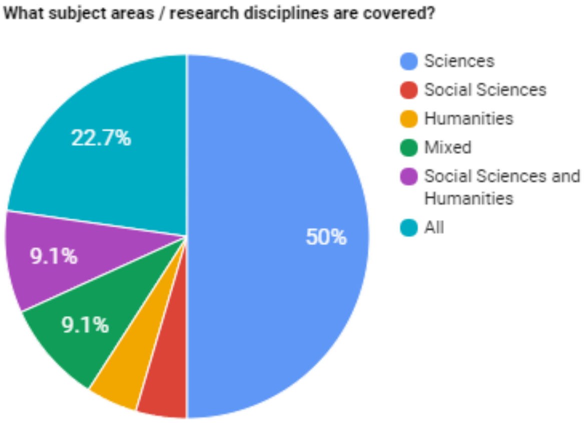
2. Cost recovery for data centres: topics and trends

This chapter contains the results of the in-depth interviews with 22 data repositories around the globe, which were analysed with the help of staff of the Center for Quantitative Social Sciences at Harvard. The interviews were held in person or over the phone, and were guided by a questionnaire (available at https://docs.google.com/forms/d/14rWiyPq8vxma2sQpXNReilAtkC-O-lHr2Yl6o5U3jpl/viewform?edit_requested=true).

2.1 Typology of Data Centres Studied

2.1.1 What subject areas / research disciplines are covered?

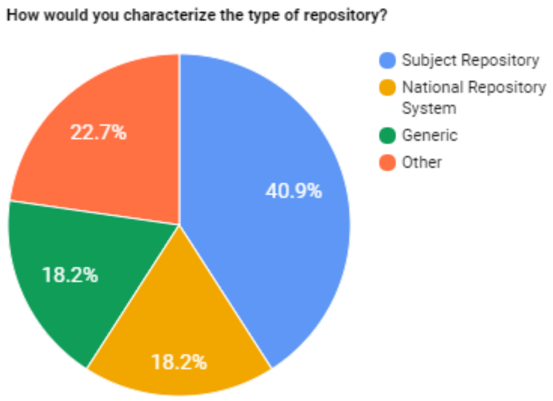
Sciences	11
Social Sciences	1
Humanities	1
Mixed	2
Social Sciences and Humanities	2
All	5



As shown, we surveyed repositories in many different domains, thought there was a focus on science.

2.1.2: How would you characterise the type of repository?

Subject Repository	9
Institutional	0
National Repository System	4
Generic	4
Libraries or Museum	0
National/Governmental Archives	0
Other	5



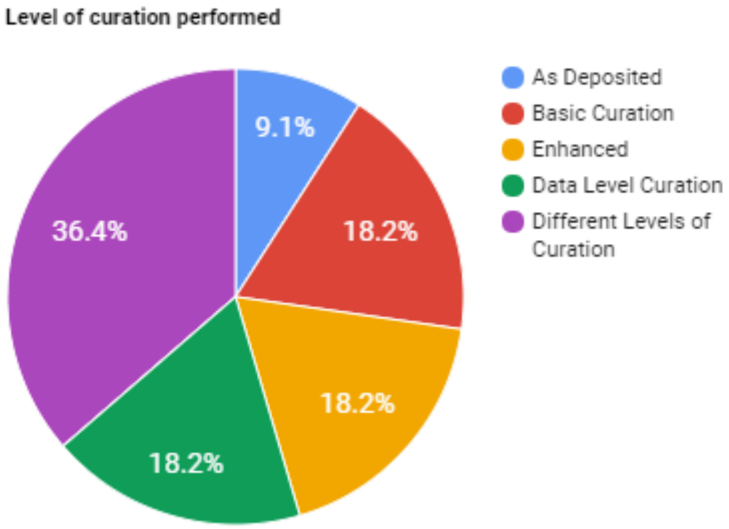
In terms of repository types, 40% classified themselves as Subject repositories, and 18% as ‘Generic’ or ‘National’ repositories (see Table and Figure 3.2). There were no government archives or Institutional repositories part of the survey, which can be seen as an omission.

2.1.3. Level of curation provided?

For this question, we were seeking to find whether the repository employs different approaches (and potentially income streams) for different data collections:

1. As deposited: the repository makes the data available as it is deposited without intervention.
2. Basic curation: the repository undertakes e.g. brief checking, addition of basic metadata or documentation.
3. Enhanced curation: the repository undertakes e.g. enhancement of documentation, basic data editing, format migration.
4. Data-Level Curation: as c. above but with additional editing of deposited data for accuracy as well as harmonization of metadata and data according to community standards.
5. Different levels of curation in different circumstances

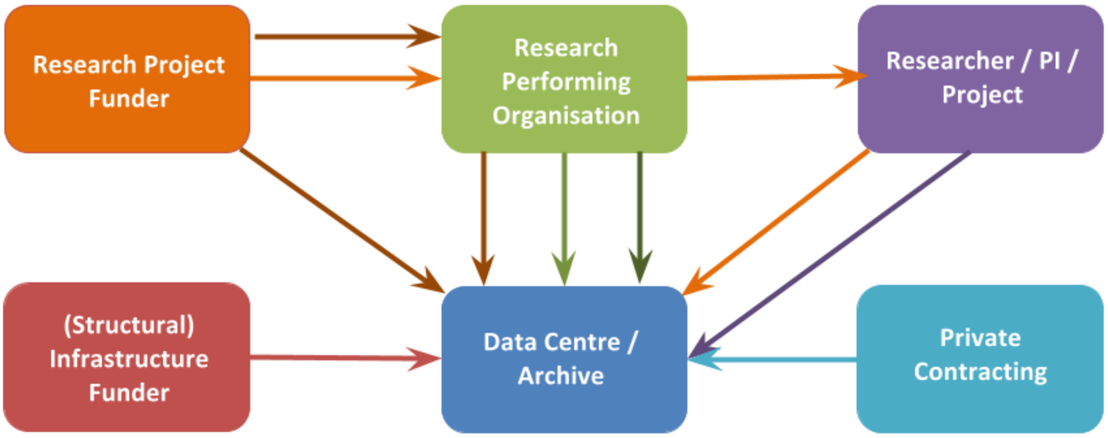
As Deposited	2
Basic Curation	4
Enhanced	4
Data Level Curation	4
Different Levels of Curation	8



We received a varied response to this question; this reflects the different cost models under which the repositories operate, as well as the differing mission statements. This will tie in questions about where money comes from (i.e. depositors or other stakeholders paying for different kinds of curation) and questions about the term of preservation (which links back to the repositories' mission statements). In future work it will be important to consider the stakeholders' willingness to pay or to fund against the level of curation (what levels of curation are necessary for reuse, what are required by users, what data 'merits' high or low levels of curation)? However, such stakeholder analysis is beyond the scope of the current study.

2.2. Typology of Income Streams

Through discussion and desk analysis, the Group developed the following typology of data repositories' income streams:


Research Project Funder
Research Performing Organisation
Researcher / PI / Project
(Structural) Infrastructure Funder
Data Centre / Archive
Private Contracting

1. Structural (central contract)
2. Hosting Support (indirect or direct support through institutional hosting)
3. Annual Contract (from depositing institution)
4. Data Deposit Fee (may be paid by researcher, RPO or publisher; may originate with funder)
5. Access Charge (for the data or for value-adding services)
6. R&D Projects (to develop infrastructure or value-adding services)
7. Private Contracting (services to parties other than core funder)

1. **Structural funding:** a major central contract or grant from a funding agency or an institution to perform a given role as a data repository. Many data repositories rely on a major central contract with a given research or infrastructure funding agency.
2. **Hosting support:** the direct or indirect financial support of being hosted at a given institution. Many data repositories are hosted at universities or other research institutions and this may result in significant direct financial or indirect in-kind support.
3. **Contract from depositing institution:** repositories may provide a data stewardship service to a research institution. As universities and research performing organisations are increasingly required to ensure data is looked after, outsourcing this activity to an expert repository organisation may become appealing and may provide a source of income for the repository.
4. **Data Deposit Fee:** a single, one-off fee paid for the long-term stewardship of data. With analogies to the Open Access Article Processing Charge, some data repositories are now generating income through Data Processing Charges calculated to cover the costs of curation and long term preservation of the data. Data deposit charges might be covered by a block grant to Research Performing Institutions, such as the Gold OA grants in the UK or they may be funded by dissemination budgets in research grants, as is also the case with Gold APCs.
5. **Data Access Charge:** some repositories charge users to access data. On the model of subscription-based publishing, it can be argued that the data user has the greatest interest in the data and so should fund the data repository by means of subscription, membership or other access charges. The challenge is that this model may now find itself at odds with Open Access principles when the data is produced as a result of public funding.

6. **Research and Development Projects:** data repositories look to project funding for research and development, building infrastructure or refining technology, policies or materials. Such funding can be financially important for the data repository and can also be an important source of additional expertise, partnerships and fundamental research and development.
7. **Private Contracting:** this covers contracts other than structural funding, contracts with specific depositing institutions and grant or contract based R&D projects. It may include, for example, consulting and expertise provided to the public or private research sectors, to other curation or memory institutions etc.

In the detailed structured interviews, data repository representatives were asked to estimate the proportion of income obtained from each of these channels and to identify any significant sources of income not covered by this typology.

2.3 Responses of the Interviews on Cost Recovery Types

*Some answers to the question: "**What are the current income streams for the repository?**"*

There are two main income streams […]:
1) financial contributions received for value-added services, for example: providing systematic search and data analysis functionality in front of the archive
2) software license fees via scientific software applications developed [here]

Our main income streams are:
1) Donor project funding
2) Training [for various organisations]
3) [Hosting institution] provides funding for IT infrastructure
4) Service Level Agreements with university projects

Roughly 25% comes from member dues; 10% from fees for [a training] program […]; the remainder comes from contracts and grants [where we] offer data archiving to agencies (funders).

Almost entirely funded by research infrastructure (government) funder.
Data flows in from a variety of sources, but is mainly funded as research infrastructure.

Three sources:
1. funding based on the annual contracts
2. funding based on data deposits by the consortium partners
3. funding coming from additional services (training, consultancy)

Development of the software: mostly Research Grants […]
Support of the repository itself: mostly [hosting institution], hard money
Support from [various commercial companies] for projects

1) Research infrastructure funder: core infrastructure is maintained by a grant.
2) Research Project Funders: have generated money to [develop specific software]
3) [Public] foundations: don't want to foot the bill for full infrastructure, pay to make their data available.
4) Researcher/PI/Project: We [co-submitted] several grants; one got funded, where we helped host data.

The findings of the interviews reveal the following broad funding models for data centres.

1. Largely structurally funded.
2. Reliant on data access charges.
3. Exploring data deposit fees.
4. Supported by host institution.

5. Substantial diversification.
6. Supported by project funding.

Additionally, although some data centres were experimenting with data deposit fees and with contracted services of various sorts, we can observe that the most important means of supplementing core funding or diversifying income streams was through project funding of various kinds. If not strictly speaking a funding model, this phenomenon deserves comment below.

2.3.1 Largely Structurally Funded

Just over half of the data repositories surveyed rely on structural funding for the majority of their income. The proportion of funding ranges from 67%-100% and comes from a research funder with a direct interest in the stewardship of a particular type of data. Half of these again (i.e. a quarter of the overall sample) receive all their income from structural funding in relation to a contract or grant to provide a relatively closely defined data service. For the remaining half (a quarter of the overall sample) a fraction between 33% and 10% is provided by various other sources. For the most part this additional income comes in the form of project funding for additional R&D or relating to specific expertise / consultancy, but there are indications of other forms of diversification for small proportions of total income (including deposit fees and contracts with particular institutions).

This is the predominant funding model, even if there are signs of diversification in the questionnaire results. For a number of those data repositories surveyed, this model was viewed as the most appropriate and sustainable. The contracts are reviewed at intervals varying between three and five years. The general impression from the interviews is that data repositories with this funding model feel that core services are valued by the funder and relatively secure, the settlement is reasonable, but leaves little room for innovation, R&D or additional services unless these are specifically contracted.

2.3.2 Reliant on data access charges

Only one of the data centres surveyed relied for the majority of their income on data access charges. What is more, that repository provides subscription-based access to enhanced data with value added features, rather than to relatively raw experimental outcome data.

Only one other repository relies for a significant proportion (c.25%) on access fees. The repository in question has a considerably diversified business model.

There is little doubt that the principles of Open Access to the outcomes of publicly funded research militate against business models which rely heavily on access charges. That said, a number of repositories saw potential income streams in charging for value added services - enhanced data, visualisation, analysis or interpretative services - built on top of data which are held in the archive in their deposited state and which were created by publicly funded research. However, it remains to be proven that such statements were not largely aspirational and without significant R&D investment the development of such services will be hard to realise.

It should be observed that some data repositories that rely on structural funding do so with an explicit remit of providing ‘data products’ and value-added, researcher assisting data services.

It seems unlikely that any repository caring for data produced by publicly funded research will start charging for access to those data, without adding very considerable value (beyond peer-review and quality assurance). If value-added services were widely referenced as a possible source of additional income, there remains a gap between aspiration and realisation - and one that will require investment to bridge.

2.3.3 Exploring data deposit fees

For two of the repositories surveyed data deposit fees form the core of their business model. For one of these repositories, data deposit fees / data processing charges was the business model at the outset. The repository in question is transitioning from project funding to an increasing reliance on data processing charges. For the other, a shift to data deposit fees, largely from research grant holder, was necessitated by a change in approach from the major national research funder in that field. Over a short period of time, the data repository in question has successfully transformed its business model to be over 50% reliant on data deposit fees.

Two other data repositories currently rely on data deposit charges for a small amount of their income. A small number (2) expressed an interest in exploring data deposit fees as an additional source of income.

The Open Access movement with regard to papers has suggested that the dissemination of knowledge created by publicly funded research should be provided for by the research project or by the host institution. In the ‘gold’ OA model, Article Processing Charges pay for the dissemination of the research and other key infrastructure activities, like the creation of permanent identifiers and long-term preservation. Should / can a similar model be applied to data? Perhaps surprisingly to the investigators, there was relatively little exploration of this model. Institutional data repositories discussing the financing of additional infrastructure have speculated whether internal data despot fees, taken from research grants, might provide a flexible, scalable, on demand / PAYG means of funding institutional data infrastructure, one which fits with the attraction of Infrastructure as a Service (IaaS).

A next step for this research, might be to analyse further - with stakeholders drawn from research funders, RPOs, data repositories and research grant holders - whether IaaS can be effectively funded by data deposit fees.

2.3.4 Supported by Host Institution

Contrary to expectations, based on initial discussions in the working group, we found that most data repositories did not receive substantial funding, directly or in kind, from their host institutions. Indeed, a number complained earnestly, even bitterly, that far from subsidising the repositories activities, the host institution sliced off contributions from the central grant / structural funding or research projects.

There were exceptions to this trend: one repository was substantially supported by the host, two received 15-20% support and two more received support in single figures of % income. The majority of data repositories paid indirect hosting costs from structural funding or research grants into the coffers of the host institution.

2.3.5 Substantial Diversification

Nearly half of the data repositories surveyed demonstrate significant diversification of funding. For some of the repositories, this involves contracts with depositing institutions. However, this remains relatively small. Similarly, as noted above, only a small number have started exploring deposit fees; a larger number are interested in this approach.

It should be clearly stated that the most significant source of alternative funding is found in time-limited projects from established research funders. For the repositories whose income from structural funding ranges between 70 and 90%, it is for the most part R&D projects that makes up the difference. One of the repositories surveyed was nearly 50% reliant on project funding. Another, as a new initiative, was still almost entirely project funded.

In this survey, a number of options were encountered for revenue diversification.

These include:

a) contracts with research performing institutions;
b) data deposit fees;
c) charged value added services;
d) R&D projects.

Of these, the first three are relatively under exploited.

2.3.6 Supported by Project Funding

Our data infrastructure, the organisations on which we rely for the long-term stewardship of research data remains considerably - it might be argued, disproportionately - reliant on short term project funding. Let us be clear: project funding is highly valued by many data repositories. Some data repositories with substantial structural funding, nevertheless, regard short-term R&D projects as highly valuable. The funding helps with particular activities that might not be covered by structural funding (including, but not limited to travel to conferences and other valued research activities). However, for this class of institution it is organisational development through the intellectual, research value of the project work that is most highly valued.

That may not be the case with all data repositories surveyed. Some demonstrate a considerable relative reliance on research projects. Among such institutions, it was acknowledged that while certain sources of funding (host institution, central contract or deposit fees) covered the core function of data curation, project-based R&D funding was absolutely essential for the repository to progress, to develop its business processes and to enhance its services.

Some answers to the question: **“Which additional income streams are you exploring?”**

"A potential increase in the future contract work revenue stream exploiting in-house expertise, while at the same time recognizing that this increase is limited and comes with an overhead of additional staff during contract work.”

“We're experimenting with a different funding model: question is whether that is adequate for something that is inside the university, might be better handled in a commercial setting. Commercial sector is designed for services and that's who should handle these things.”

“Restricted use data: desire to recover some costs from this, [this model is] becoming more common as a percentage of data received. Providing the service is expensive and needs to be recovered. Need to move costs to the users benefiting from those services. Also risks involved in distributing confidential data - the services required have costs."

“Possibility of private contracting of curation service, but this is more good citizenship rather than genuinely income generating - i.e. will not charge much more than real cost.”

"Hoping to attract revenue directly from funders for region-wide provision of services. This would offer economies of scale and reduce the possibility of disparity between large and small universities.”

“Considering the possibility of offering terabyte levels of storage to end users. [Exploring the] possibility of richer end-user services for publishers."

“We would love to open up the ability to cover some of our infrastructure cost through sponsorships; would love to migrate some of this to an investigator-initiated model. User pays and sponsorship, are two of the things we are working on; [however, finding the right] costing mechanism is a challenge!”

"1) Hosting of restricted use data, with user access charges.
2) Providing branded repository services for other institutions, journals etc”

"To be able to keep up with the technology developments and enhance services and increase remit, [we are] considering more commercial contracts."

> *“We plan to charge all depositors who deposit more than 2GB of data. This will be done within the framework of an agreement with the institutions for which the researchers work (universities, of the research institutes).*
> *Secondly, [we will] consider storing datasets off line after 20 years if they have a low re-use. These datasets will be kept, but access to them will be charged."*

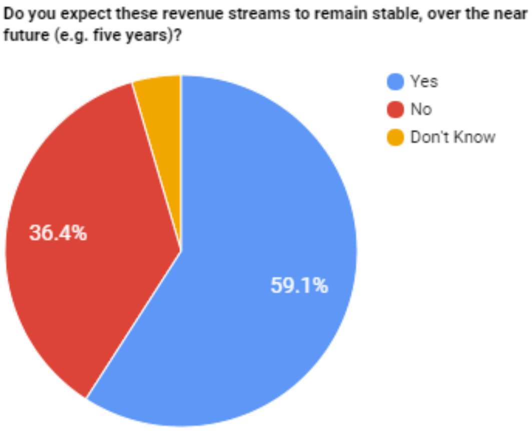
3. Views on the Future

3.1 Views on the Future of Funding

Repositories were asked for their views on the future of funding: whether the repository expected the income streams to remain stable, and whether it was expected that the existing funding streams would remain sufficient for the tasks the repository would need to perform in the future. We will briefly discuss the responses we received.

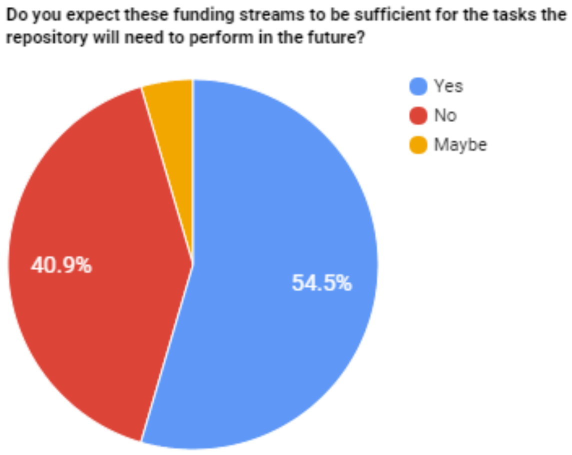
Do you expect these revenue streams to remain stable, over the near future (e.g. five years)?

Yes	13
No	8
Don't Know	1



Do you expect these funding streams to be sufficient for the tasks the repository will need to perform in the future?

Yes	12
No	9
Maybe	1

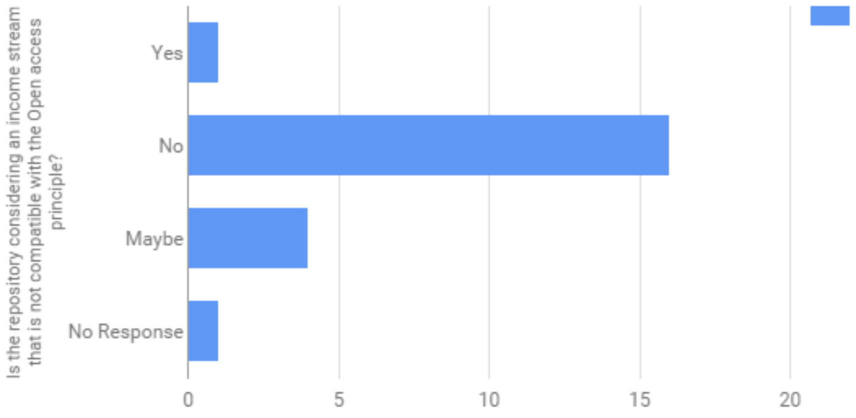


Just over half those surveyed felt that total income would be sufficient for the repository to perform the tasks needed in the future. That begs the question of whether the metaphorical glass of future funding is half-full or half-empty. We might be reassured that just over half of the repositories felt that funding would keep track with mission requirements. Equally, we might hope that with so much written about the ‘data revolution’, that more data repositories felt optimistic that funding for the important services provided would keep pace with growing data quantities and costs.

With a small sample, it is not possible to draw significant correlations between the type of income stream and the responses to this question. Six of the nine repositories that answered that they did *not* expect to funding to be sufficient for all mission tasks relied on structural funding or a major contract. This does not, however, mean that other means of funding were perceived as more secure. Concern about whether funding would keep pace with the requirements of mission tasks was spread reasonably evenly across the funding models listed above.

Is the repository considering an income stream that is not compatible with the Open access principle?

Yes	1
No	16
No Response	1
Maybe	4



Yes: 4.5%, No: 72.7%, Maybe: 18.2%, No Response: 4.5%

Are there activities that a repository would like to be doing but can't because current revenue streams don't provide sufficient revenue?

Yes	18
No	4

The answers to this question might be taken to indicate a greater degree of concern about the availability of funding than those in response to the question of whether funding streams would be sufficient for the tasks the repository will need to perform in the future'. We surmise from this, and from some of the answers reproduced below, that even when a data repository is relatively confident that funding will keep pace with mission tasks, there are always additional developmental, R&D and enhancement activities which are required.

Some answers to the question: **"Are there activities that a repository would like to be doing but can't because current revenue streams don't provide sufficient revenue?"**

"Yes, in particular feature development and functionality. Increasing efficiency and functionality requires investment. This largely depends on R&D grants. Also need investment in outreach activity."

"We are not allowed to do the analysis of the real business model cost, this is not a contracted deliverable! Don't have the funding to be able to investigate feasible business models."

"Yes:
1. [We] would like to offer better services for dynamic data.
2. [We] would like to provide software sustainability solutions. Presently, the intelligence in the application software is lost after a certain number of years."

"Yes, we would like to increase accessibility to larger datasets. Currently technical issues prevent this."

"More effort in the areas of awareness raising and data stewardship with [the institutions we work with]. More [money to develop an] automatic ingest functionality."

"We are not allowed to do the analysis of the real business model cost, since this is not a contracted deliverable! Don't have the funding to be able to investigate feasible business models. Trying to figure

out the right cost for a Terabyte of data, or 7 hours of curation?”

“Yes, we are keen to undertake data rescue projects […], as much data is lost through poor curation. We also have data quality input we cannot do currently because of staff shortages.”

"Revamp platform (5 years old) but no funding. Improve internet connectivity, distributed system with many stakeholders and connectivity is a challenge.”

“[We would like to improve our] curation and increase services such as DOI minting and discovery services.”

"

“[We would like to] improve the usability of our interfaces [and develop] tools for data submission and the [I]mplementation of bibliometrics"

‘[We] need to move towards really open data. Open data demands a large investment in man power, because open data need more work in order to make them interpretable for a larger audience.”

“This is an issue of scale. The ingest of collections is limited due to funding constraints.”

3.2 Alternative Revenue Streams

A substantial number of the data repositories surveyed are exploring alternative revenue streams. Two of those that were not exploring alternative revenue streams reported that this was because their focus was on expanding *existing* revenue streams determined by their business model. For the remainder (four) a tightly circumscribed mission or stable relationship with the major funder meant that exploring funding streams was either not in scope or not necessary.

Are you exploring alternative revenue streams?

Yes	15
No	6
Maybe	1

Yes: 68.2%, No: 27.3%, Maybe: 4.5%

The need to expand revenue and the willingness - or need - to do this through a diversification of income streams is evident. Also notable is the very diverse list of alternative income streams mentioned. These include: contract work of various types (including particular mission related services for funders or other institutions, consultancy and sale of expertise etc), research projects and other R&D activities, developing greater contributions from host institutes, access charging (generally in relation to some value added service), data deposit charging and sponsorship. Cost *saving* was also mentioned.

The alternative revenue streams will now be discussed in more detail.

3.2.1. Contract Work

A substantial number of repositories are developing income streams through various forms of additional contracted work. In some cases this is an expansion of direct contracts with the existing major funder for new services, in others it is seeking new markets. In the latter case, it may be providing data curation and repository services for research institutions, for the private sector, NGOs and government. The possibility of providing data stewardship for research institutions at a regional scale was also mentioned. Along with research project activities this was the diversification being most widely explored. At least one respondent commented that the repository did not currently have the resources necessary to pursue this type of income diversification effectively.

3.2.2. Research Grants and other R&D Funding

Similarly, many repositories are already supported to a substantial degree by grant funding of various sorts (this includes research projects, infrastructure development, grants for training and outreach activities etc). A number of repositories regarded such activities as important both for income diversification and to enhance the repository's 'offer' beyond the core service.

3.2.3. Funding from Host Institution

Some repositories are seeking to enhance income from the host institution: 'We are currently exploring options such as more sustainable funding from the [host university]'.

3.2.4. Access Charges

A number of repositories are considering the possibility of charging for access. Given the policy environment this strategy was always associated with charging for some value-added service. The variety of value-added service that might be considered is of great interest.

Charging for access to restricted data (e.g. sensitive data which has to be restricted). Some repositories already charge for this on a cost recovery basis for the extra cost of running secure services.

Charging for access to data that has been deep archived. Mentioned by a number of repositories, the idea of charging for data that has been deep-archived would allow some cost recovery for long term preservation of data that the collection policy and practices had determined should not longer be on top-level storage but which should nevertheless be retained and in which there may still be occasional research interest.

Charging for enriched, value-added services. Some data repositories already have charging for value added data services as part of their business model and a number mentioned it as an option they were considering. The Open Access principle for the outputs of publicly-funded research means that there is a presumption that data as produced by data collection exercises or research projects should be openly available as deposited. However, there is interest among some data repositories to charge for access to data where there has been significant enhancement, enrichment and value-adding work. This may range from enhanced metadata, data linking services, visualisation and analysis tools, synthesised/integrated data products and so on. There may be opportunities to develop enhanced services for particular stakeholders: for example, data visualisation widgets that enhance journal articles with views onto underlying data. Moreover, the economic benefits gained through developing an innovative market of enhanced data product or services built on top of Open data is one of the broader advantages of Open data policies.

3.2.5. Data Deposit Fees

A number of data repositories are exploring data deposit charges: for two of the repositories surveyed, this source of income is the major component of their business model. As noted above, the deposit fee model may provide a transparent way of ensuring that infrastructure for data from research projects is funded in a sustainable way through hypothecated components of project grants. The model also has the virtue of aligning with the author-pays Open Access approach, increasingly being adopted, in which dissemination of research outputs and results is regarded as an integral part of research activity and the cost of research. Deposit fees have the *potential* to scale effectively with the growth in data being produced; conversely, a risk is that without effective policy and monitoring support, data deposit charges will be squeezed by other research priorities.

Some repositories were considering reducing overheads by only charging for deposits which incur additional costs: for example, charging for data deposits over a certain volume or for data deposits requiring additional or special curation.

3.2.6. National Infrastructure Funding

Some data repositories argue that data and the institutions that steward data are an essential component of the infrastructure necessary for research. Like the familiar shared infrastructure for transport, this is one of the things that government rightly provides for the public good. In the case of data infrastructure, it might be argued, the structural grant is the most appropriate. A number of data repositories mentioned negotiations, broader partnerships, infrastructure reviews or roadmaps as well as the offer of new or enhanced services as means of increasing structural funding.

3.2.7. Sponsorship

Sponsorship to cover infrastructure cost, or to pay for access to otherwise access-charged data, was mentioned as a potential source of additional income.

3.2.8. Cost-Saving

All data repositories are concerned to manage and reduce costs, where possible. The declining cost in storage certainly assists proportional cost savings. The use of tiered storage, saving costs on infrequently accessed data was mentioned specifically as a means of accelerating this. However, it should be acknowledged that uncovering the varied means of achieving cost-savings was not an explicitly targeted objective of this study: perhaps it should have been; and it should certainly be part of a future study.

3.3. Preliminary Findings: alternative revenue streams

There is a recognised need to expand revenue and a large proportion of the data repositories surveyed are exploring alternative income streams. The drivers are increase in costs (due to increasing volumes, deposits and demand for data), the pressure to provide new services, the need for research and development activities, and the concern that existing sources of income will not keep pace with these pressures.

4. SWOT Analysis of Funding Streams and Business Models

Section 3.2 describes six broad funding models for data centres. For the purpose of a SWOT-analysis of these different models, we have grouped them into four principal categories:

1. Largely structurally funded (including support by the host institution)
2. Reliant on data access charges (including membership fees)
3. Exploring data deposit fees
4. Substantial diversification (including project funding)

A SWOT-analysis is a useful technique for identifying the strengths and weaknesses, as well as the opportunities and threats of a strategy. The Group used the RDA Plenary 6 in Paris to ask stakeholders to evaluate the four funding models on the basis of such an analysis. The outcomes are presented below.

1. Largely structurally funded data centres

The main strength of this model is the long-term sustainability it provides, which allows data centres to plan in advance and build an effective organisation. Fundraising does not consume too much time. It was also argued that there currently exists a real window of opportunity since data, their use and accessibility, are a 'hot' topic and funders have increasing infrastructure budgets.

On the other hand there is a danger that structural funding will not keep pace with growing data volumes, that the data hype curve may soon reach its peak. The primary weakness and threat posed by this form of funding is its inflexibility and that it can constitute a single point of failure.

STRENGTHS • Longer-term stability: easier planning and achieve efficiency • Stronger commitments and communication with stakeholders • Larger chunk of investments can cover operational costs • Up front funding can help plan budget and build effective organisation • Immune to marketing and collateral effects • No need to spend too much time fundraising	**WEAKNESSES** • If only renumeration for capital, this is a risk • Fixed funding is a weakness wrt the context of (immensely) growing volumes of data • Can reduce the efficiency; no incentive to improve; long evaluation cycles make you lazy! • Inflexibility of funding, can't adapt easily
OPPORTUNITIES • Data is hot and funders are more amenable to provide structural funding • Riding the hype and gaining structural funding can help raise the profile of institutions (win-win) • Funders have increasing budget for infrastructure • Data is/can be recognized as infrastructure • Institutions (universities, RPO, etc.) recognize their responsibility over funding the data infrastructure	**THREATS** • "Today it's hot, tomorrow it's not!" • Not receiving structural funding because of big national initiatives with which you are not aligned • Increase demand cannot be handled easily • Not in control of your funding – dependent on small nr of sources • Funder itself may be descoped (e.g. US)

2. Data centres reliant on data access charges and membership fees

Data access charges are scalable and the charges can reflect the value of the data products. The access charges model is the most market-oriented approach and also offers the possibility of creating a basis for free access to basic data products. In the case of the membership model there is a relatively stable and predictable income and the members have an influence on the organization that allows for creating a loyal community.

The major weakness of the access charges model is that it is conflict with the Open Access principle. It can lead to unequal access to what should be a public good and for data that have already been generated through public funds. It might also create unhelpful competition between data centres for high impact data. Finally the access charges could be vulnerable to economic downturns.

STRENGTHS	WEAKNESSES
• Consumer pays for what consumer wants • Cost scales with access if cloud-based storage • Income reflects value of data product • Stable and predictable income from membership / subscription fees; income continues in the long-term • Loyalty of community of members • Members have influence over priorities • Model applicable to similar/related services such as DataCite • Model can accommodate licensing flexibility	• Data worth serving in the long term might not be available • Causes competition amongst repositories for high impact data; data poaching • Not affordable to people at underprivileged organizations; unequal access to what should be a public good • Could lead to data purging • Must think carefully about licensing terms and conditions
OPPORTUNITIES	**THREATS**
• Monitor demand, change/improve services • Create free access for basic products, build consumer base and create demand for value-added services • Most market-oriented approach • Valuable for private sector • Creates captive market	• Vulnerable to economic downturns – although there is actual increased use of archival data when primary funding is down • Expectation is for free access, other providers might undermine business

3. Data centres exploring data deposit fees

In contrast with data charges, data deposit fees are compliant with the Open Access principle. The model puts charge on the data depositor and, in principle, this works well with grant funding. Furthermore, the model is neutral to the value of data to end-users and it is scalable.

On the other hand there is a danger of large administrative overheads and depositors might rush to the cheapest option. The model requires a very clear policy framework to be effective and to avoid such risks.

STRENGTHS **Puts charge on data producer (works well with grant funding)** **OA compatible** **Scalable** **Closely linked to the research community – responsive to science need** **Competition** **Neutral to value of data to end users (no a priori value judgment)** **Potentially fair/proportional distribution of funding**	**WEAKNESSES** **Defining the cost (POSF)** **Does it meet the challenge of diverse data types** **Market weakness vs structurally funded repositories** **Administrative overheads** **Neutral to value of data to end users (data centre has to accept all paid data)**
OPPORTUNITIES **Autonomous generation of revenue** **Scaled deposit fee model** **Compatible with subscription as part of business model**	**THREATS** **PI pushback (vs top-slicing research grant)** **Rush to cheapest option?** **Needs very clear policy framework** **High cost will put off depositors** **Hostage to future storage and preservation costs** **Infrastructure costs are estimated too low**

4. Data centres with substantial funding diversification

Diversification of funding is attractive, as it has no single source of failure; financial risks are spread over different income sources. It offers flexibility to experiment with new services and markets and can stimulate innovation.

The flipside of the coin is likely to be a relatively high administrative overhead. With different funders with possibly diverse and/or shifting interests attention could be drawn away from the core mission of the data centre.

STRENGTHS	WEAKNESSES
• No single source of failure • Flexibility to experiment with new services and markets • Stimulates innovation • Focuses attention on value to users	• Access fees exclude users/limit uses • Funding is short term; obligations long term • Sponsor priorities change • High administrative overhead • Requires highly skilled staff • Host universities are not stakeholders of national repositories • Sustainability of funded projects • Draws attention away from core mission
OPPORTUNITIES	**THREATS**
• Research funding is project based • Data management requirements are creating demand from researchers for services during the project funding • Sponsor priorities change	• Competition • Commercial companies • Institutional repositories • Variability of funding

4. Conclusions and Next Steps

The report above provides a snapshot on the basis of the results of a questionnaire conducted with 20 or so data repositories. It provides an overview of current funding streams and business models for data repositories. It is hoped that, as such, it will give those repositories, other stakeholders in data infrastructure and funders, some insight into the current opportunities to diversify income streams and develop sustainable business models that are elastic enough to meet new and growing data needs.

The workshop held at the 6th Plenary of the RDA in Paris, September 2016, was particularly helpful in allowing the Group to conduct a SWOT analysis that elucidated some stakeholder reaction to the principal income streams and business models identified.

The work of this Interest Group has produced a useful landscape study of the income streams and business models of 22 data repositories worldwide. It has also demonstrated the utility of stakeholder consultation – here in the form of a SWOT analysis.

We present here a summary of the preliminary findings of this study:

Funding budgets, instruments and procedures vary widely across the repositories surveyed: there is no single formula for how data repositories should be funded. There is as much variation within as between national funding. systems.

Overall, data repositories' funding appears to be determined and allocated more on a per-initiative basis with considerable variety in approach. Funding horizons are relatively short (if compared to other memory institutions), and budgets are substantially augmented by short term, project based funding. The sustainability and adequacy of funding is a matter of concern for many repositories and there are many unsolved issues in this area. According to a recent survey report of Science Europe it is important to realise that sustainable funding is supported by regular funding opportunities.[8]

Structural funding in the form of a major renewable grant suits a lot of data repositories, so long as the repository feels that the funder has a major long-term commitment to the service and there are responsive mechanisms to ensure that the funded service can evolve and advance as required.

The possibility of charging for value added services is seen by many data repositories but is under exploited. Data access charges are at odds with current policy requirements, unless they are specifically for value added services.

Data deposit fees are being explored, but only in a limited way and there are concerns about administrative overheads and negative market effects. However, they are consistent with the Open Access policy principle and potentially scalable. Further stakeholder and cost-analysis should be conducted.

Many data repositories value participation in research projects for organisational development and in a number of cases project funding is an essential additional source of income. Are data repositories overly reliant on short-term project funding for activities that relate to long term core activities or at the very least to 'business intelligence' and development?[9] The overheads incurred by project funding should also be analysed. Stakeholders should consider carefully what proportion of the data infrastructure should be supported by short-term grants: a maximum of one third was mentioned a couple of times as representing an appropriate upper limit. This would need to be tested with stakeholders in the context of further work.

It is customary in a conclusion to observe that further research is required – and the present occasion is no exception. Two areas in particular need more attention in future work in this area: innovative income streams and cost savings.

Relatively few innovative income streams were identified by the data repositories interviewed. To do so in future work, we will need to cast our net wider and consult with business experts and not just data repository managers. Identifying the means by which data repositories may reduce or at least restrain increasing costs is another area requiring more focussed attention.

Happily the opportunity to perform further research has been confirmed. Running for 15-18 months from March 2016, an OECD Global Science Forum project will build directly on the work described above. That project will: 1) identify further emerging and innovative income streams; 2) identify existing and possible means of reducing or restraining costs; 3) test more rigorously possible business models with various stakeholders and against budgets and ability to pay; and 4) on the basis of these findings make policy recommendations to promote sustainable business models for data infrastructure.

Specific task to be undertaken by the project include the following:

1. Extend the number of data repositories analysed to include a broader number of OECD member countries and to increase coverage non-European data repositories.
2. Conduct a focus group on innovative income streams.
3. Conduct a focus group on means of reducing and restraining costs.
4. Extend the number of potential business models.
5. Perform a thorough economic analysis of the business models, assessing their viability and the value of the services offered and exploring the alignment of these with stakeholder and beneficiaries' willingness to pay and budgetary authority. This task will be conducted in two stages. First, an economic analysis and presentation of business models will be conducted by appropriate experts. Second a focus group meeting will test the business models with stakeholders and refine the analysis.
6. On the basis of these findings, policy recommendations to promote sustainable business models for data infrastructure will be formulated with input from the GSF.

The OECD GSF project will consider the alignment of business models with such statements as ‘The Principles for Open Scholarly Infrastructure’ [9], current policy developments and the international interest in Open Science and the infrastructures to support this. In particular, we will be concerned to examine stakeholder acceptance, and therefore likely sustainability of a variety of income streams and business models.

Building on the work presented above, the OECD GSF Project promises to advance considerably our efforts to develop sustainable business models for data repositories.

5. Acknowledgements

Most of the data processing work in this report was done by Lucy Taylor of the Harvard Center for Quantitative Social Science - we are deeply grateful for her efforts.

We would like to acknowledge the staff from 22 data repositories worldwide that spent considerable time (1-2 hours each) on detailed interviews. Interviews were conducted by the following members of the RDA-WDS Interest Group: Ian Bruno, Matthew Viljoen, Anita de Waard, Mustapha Mokrane, Simon Hodson, Ingrid Dillo, Frank Toussaint. Infrastructure for virtual meetings and collaboration was provided by both WDS and CODATA. RDA Plenary meetings provided valuable occasions to consult with Interest Group members and the wider community. We are grateful to those who participated in the series of sessions and who shared their insights and expertise on the basis of presentations and previous versions of this report.

6. References

1. Cost Recovery for Data Centres Working Group Case Statement: https://rd-alliance.org/sites/default/files/case_statement/RDA_WDS_IG_Publishing_Costs.pdf
2. 4C: Collaboration to Clarify the Costs of Curation: http://4cproject.eu/
3. Curation Costs Exchange: http://www.curationexchange.org/
4. Ember, C. and Robert Hanisch, Sustaining Domain Repositories for Digital Data: A White Paper, December 11, 2013, http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf
5. Downs, R.R., and Robert S. Chen, Towards Sustainable Stewardship of Digital Collections of Scientific Data, 2013http://www.gsdi.org/gsdiconf/gsdi13/papers/130.pdf
6. Baker, M. Databases fight funding cuts, Online tools are becoming ever more important to biology, but financial support is unstable. 05 September 2012, http://www.nature.com/news/databases-fight-funding-cuts-1.11347
7. Berman, F. and V. Cerf, Who will pay for public access to research data? Science Magazine 09 August 2013 http://www.sciencemag.org/content/341/6146/616.full.pdf?keytype=ref&siteid=sci&ijkey=.e2Ezowko%2FxF2
8. Strategic priorities, funding and pan-European co-operation for research infrastructures in Europe, Survey Report Science Europe, January 2016 http://www.scienceeurope.org/uploads/PublicDocumentsAndSpeeches/SE_Infrastructures_SurveyReport_web_FIN.pdf
9. Bilder G, Lin J, Neylon C (2015) Principles for Open Scholarly Infrastructure-v1, http://dx.doi.org/10.6084/m9.figshare.1314859

7. Appendices

- List of interviewees
- Questionnaire