International Coordination of Data Infrastructures for Open Science

March 2017 OECD-WDS Workshop

Briefing Paper, Theme 5

Sustainability, effectiveness and funding for international data infrastructures

Lead author: Juan Bicarregui

1. Introduction and Background

The argument for open science has been made elsewhere (e.g. OECD, 2015, Making Open Science a Reality). It has been widely recognised that open science is both more effective and more efficient as well as being a stimulus and basis for greater innovation. Central to open science is open data. That is, researchers must be able to access the data from research teams elsewhere in the world. Thus, the availability of high quality and stable data infrastructures with global scope is an essential component in delivering the vision of open science.

Data infrastructures provide a way for research data to be shared and reused between different research teams. If researchers are to have access to the best data possible, then this sharing should be done at a global level. However, the vast majority of public funding for research is available at national level for nationally-based research teams. Thus there is an inherent tension between how data infrastructures are funded and how they are best delivered. In some domains, these infrastructures also provide data that is also relevant to private sector research. This raises questions about under what conditions publicly funded services should be made available to the private sector.

Many policy statements over several decades have promoted the benefits of open data[1] and some data infrastructures already have a history of decades of service supporting these aims. However, there is currently no consensus as to the best way to fund the provision of such infrastructures. The debate centres on several aspects, some of which are also relevant when considering physical research infrastructures. Firstly, whether research infrastructure funding should be supported

[1] eg. NSF (date - referenced in RS Science as an Open Enterprise), OECD 2006?, EC, RCUK 2011, OSTP 2013 and G8 2013.

directly or whether funding should flow through the users of the infrastructure. Secondly, what mechanisms can be used to support provision of an international service, given that funding comes primarily from national sources. A third aspect, particularly relevant to data infrastructures, is that the benefit may come to people other than those who created the data, and a fourth aspect is how to make the cost-benefit assessment regarding data preservation, given that it may not be known at the time that data is created what the long term value of that data may be. Furthermore, there is an intrinsic tension between the structural character of research infrastructures and the time-limited, project-based funding models available for many research programmes.

It has long been argued that data arising from publicly funded research should be considered as a public good [Arzberger et al 2004, OECD 2006, RCUK 2011] and as such access should be provided at no cost to the consumer of the data. However, this principle does not address the issues of how the cost of provision should be met or of how data creators and controllers should be recognised for their contribution. A great deal of research has gone into addressing these concerns. For example, a recent session at SciDataCon 2016 on Sustainable Business Models for Data Repositories included 9 witness statements and 14 papers[2]. In "*Squaring Circles: Economics and Governance of Scholarly Infrastructures*", Neylon discusses governance for transitioning from club-like to public-like goods and argues that governance objectives need to be understood before sustainability models can be decided.

Other briefing papers in this suite cover some of the barriers that need to be overcome in making the transition to a networked global research data infrastructure. Here we discuss issues related to funding. How should funding be managed to maximise the utility of the infrastructure, ensuring it best meets the needs of researchers whilst ensuring efficiency of the provision. These needs include access to as broad a range of data as possible; ability to integrate data from different sources; availability of value added services on top of raw access to the data; and stability and reliability of the infrastructure and services built upon it.

[There is a separate ongoing GSF project that is focusing on *business models for sustainable research data repositories*. Hence the funding issues related to individual data repositories are not the direct focus of the current briefing, which concentrates more on the issues relating to support for international data infrastructure networks. At the same time it is recognised that the sustainability of these networks depends on the support for their individual components and vice-versa.]

2. Analysis

2.1 Funding: sources and stability

The interviews demonstrated a very large variety in the type of funding received and the security of that funding over the long term. Many different sustainability issues were reported.

Smörgåsbord

[2] http://www.scidatacon.org/2016/sessions/45/

In all cases, the international networks analysed are funded through a variety of sources. As well as receiving funding explicitly for support of the infrastructure, most reported that funding also comes for various other purposes including research projects, policy research, provision of training, consultancy and international cooperation. Often the networks rely on a small number of funding sources for the bulk of their centralised components but these are topped up with a larger number of smaller, time-limited grants for particular projects. Most networks reported that their lines of funding are not secured and that they need to continuously apply to many different calls to get sufficient funding. The success or otherwise of these applications is often beyond the control of the applicants, with significant implications. For example, in one instance, support that was initially for inter-university collaboration was cut short when a particular initiative ended due to changes in funding policy.

Cyclic funding

Longevity of funding was often reported to be a major concern. Funding specifically to support the infrastructure normally comes in cycles of 3 or 5 years. Often, additional funding is acquired through a number of research grants which are often relatively small and are provided for research projects rather than for infrastructure. This project based funding also comes in fixed length projects normally of between 3 and 5 years. Some networks reported that initial funding had been for a fixed period after which more permanent funding was very difficult to acquire. Several networks reported that a change in approach from funders was required so that infrastructure can be supported in its own right, not just as a bi-product of funding for research projects.

National and International

Many of the interviewed networks reported a mixture of nationally and internationally based funding. This was particularly the case in Europe, where several infrastructures are primarily European initiatives supported by the EU to bring together and integrate national provision. The ERIC (European Research Infrastructure Consortium) is an important model in this context. Sometimes the national components themselves were established from units in separate organisations brought together specifically in order to provide a national focus for the European activity. Some networks were trying to move towards a funding model where strong national programmes are federated and coordinated to integrate national provision through relatively light international frameworks. The challenges faced by a central hub of an international network are different from those of the infrastructure nodes in the member countries. It is sometimes difficult to use national funding for activities outside the country where it originated. The failure to obtain funding in each node can mean both unsustainability for itself and a membership problem for an international network. Many international data infrastructure networks are also supported by member countries, and so membership problems have important implications for their overall sustainability.

Models of provision

Sometimes the very construction of the data infrastructure network was designed in order to match the available funding streams. A Hub and Node model is fairly common although there is no consistent model or vocabulary for this. In one large scale infrastructure, the single central hub is

funded internationally whilst the largest part of the provision is in the operational nodes which are funded nationally. Nodes can be of different sizes. Core nodes may have bigger budget allocation than other nodes as well as broader scope, for example in providing training. Funding is sometimes allocated transiently and nodes may not get continued funding if they are not successful.

Evolution from research services to infrastructure services

The model of funding often changes as the network evolves. Initial project based funding to establish something new may move toward more stable funding as the service matures and becomes embedded in the research culture of the domain it supports. In some cases, permanent grants to the institutions participating are a very important component of the provision.

Some data infrastructures networks reported a transition in the nature of the services they provide with data storage increasingly becoming a more important aspect of their remit with this leading to a shift in emphasis from research to infrastructure provision. With this shift comes a need for new lines of funding dedicated to supporting these new primarily infrastructural services.

Public or Private

For the networks interviewed, it was reported that funding came mostly from public sources even when significant use was from the private sector. Only a few networks reported receiving any funding from private sector sources, this being through subscriptions. Some were considering membership fees as a scalable way to reduce reliance on short term, unstable funding from public funds, assuming that if services are adopted and members are happy with the services provided, then it might be possible to sustain the infrastructure on this basis. However, in other cases membership based funding was reported as potentially leading to problems when members themselves are financially constrained.

Strategic national support

There was a common view that sustainability at the national level was critical where it is necessary for each participating country to have plans for a sustainable digital infrastructure that are brought together to provide the international dimension. However, the level to which provision of the infrastructure was considered strategic varied. On the one hand, in one country, where there is a governmental desire for science to become more international in order to ensure the comprehensiveness and sustainability, there is a clear intention to provide continued support. Whilst, at the other extreme, one national node of an infrastructure had been told with a relatively long lead time that the current cycle of funding will be the last and that they needed to find other ways to fund themselves. In response to this, they have set up a sustainability working group to lobby politicians but are uncertain about their chance of success.

Costs of connecting

One international data infrastructure network reported that they had underestimated the support required to build the connecting “glue” between infrastructures provided by its members. The current funding for that network is really available only for the coordination effort and there is now an ongoing effort to expand the funding specifically to establish the distributed infrastructure and sustain it in the long-term. On the other hand, another reported that the funding they receive is just to maintain the infrastructure hub, as well as supporting the governing board and some standing committees dealing with the nodes and organisational aspects. This one is multilateral initiative

established by intergovernmental agreement and based on a non-binding Memorandum of Understanding (MoU) so funding is not really an issue thanks to the MoU.

One already very mature data infrastructure network reported that over the timescale of decades, it may move to a situation where international funding for the integration of national infrastructures may become irrelevant as national funding becomes embedded and provides the sustainability of the services.

2.2 Commercial and industrial involvement

Disciplinary characteristics

As would be expected the level of involvement of the private sector varies largely across disciplines. In some purely academic fields there is no involvement with the private sector at all and the infrastructures are run entirely on public funding. However, in other fields industrial participation is critical. For example, in the chemical sectors there are publically collected data sets and associated services that are provided to pharmaceutical companies and universities on a commercial basis. Access can sometimes be quite expensive and some argued that, where use is for publically funded research, it might seem that the public purse was paying twice.

National Characteristics

The interviews revealed that views about the involvement the private sector varied significantly between countries. Whilst in some countries it is viewed as completely natural, even desirable, that the private sector pays for access to these infrastructures, other countries provide support entirely from public funds. Some concerns were expressed about private provisioning, arguing that it is more economical in the long run for the activity to be supported from public funds than it is to use private provision that is later recompensed through commercially provided services. In one case, an infrastructure that was entirely publicly funded in most countries, was delivered in a specific country through a national project delivered by a university with funding from the private sector.

Private use of publicly funded infrastructure

In general, the publicly funded data infrastructures make their data available at no cost to the consumer irrespective of whether that consumer is from the public or private sector. In some cases, industry is the largest user of the data. For example, in the bio sectors, the open data provided by the large infrastructures is critical for example for the development of new drugs or enzymes. Many companies are relying on it and, although there is no clear way to track this or to measure the benefit of this to the economy, it is clear that the accessibility of biological data to industry does enhance competitiveness and innovation.

Several infrastructures were trying to strengthen their industrial user base and were actively trying to make their research more relevant to industry needs. The possibility of charging for services provided to the private sector was being considered by some of them.

Public Private Partnerships and Open Innovation

Most of the data infrastructure networks interviewed were not-for-profit organizations primarily focused on public funded research. However, some networks have members or nodes that do have close connections or collaborations with the private sector and some have associate members that

are themselves private sector, for-profit entities. These private sector organisations can be data providers, data users, or technology companies.

Some data infrastructure networks reported increasing collaboration with private organisations through projects that were that were being delivered through public private partnerships (PPPs). Some projects are aiming for open innovation including sharing of data openly across the scientific community, including public and private research teams. In some domains, large projects funded by private companies are being delivered through mixed public-private research teams. For example, pharma companies sometimes collaborate in this way in early stage research combining public and private funding for specific problems, accepting that resulting project outputs become public.

Public-Private Partnerships are also relevant for provision of data services. Whilst some data infrastructure networks are just talking with the private sector data holders about how they might open up their data, others reported that services like the ones they are provisioning might in future be commodified by the very large software service providers. Some were working with these large commercial organisations who were providing technical expertise for example in database design. In some cases these organisations were participating in a very active way in the community meetings and developing consumer products exploiting the availability of the scientific data.

Fostering Private-Private collaboration

One network reported that companies sometimes provide their data, or at least their data catalogues, through the global catalogue so potential customers can see what data they hold as a way to develop collaborations and build their business to business offering.

3. Conclusions

Data Infrastructures and associated services in international networks are generally provided by data specialists from within the research communities they serve. Typically these data infrastructure networks are vertically integrated, dedicated to serving related fields within one research domain with little sharing of resources across domains. Funding normally comes from a variety of sources and often the network itself is structured to match the funding streams. For example, there may be a number of national nodes, each with its own national funding lines, perhaps with a central hub funded by through an international agreement. Such distributed models of provision require significant coordination and the amount of work required for effective integration that delivers seamless access to the combined resources is sometimes underestimated.

Not surprisingly, there are many different funding models, and sources of funding vary significantly across disciplines. Funding is often piecemeal with different sources providing funding for different purposes and this fragmentation of funding militates against efficient delivery of services and the need for infrastructure to be stable over the long term. In some cases, it was reported that it was difficult to obtain funding specifically for infrastructure provision, particularly after a start up phase, and that infrastructure was only being supported as a by-product of research funding. Some data infrastructure networks were being encouraged towards sustainability by providing paid-for services. However, there were worries about stability of such funding particularly when customers

themselves were squeezed and concerns about the efficiency of recycling of funds from one public budget to another.

Where long term sustainability had been achieved it had sometimes come about through a gradual evolution over several decades from ad-hoc collaborations between projects to securing coordinated national provision formalised through an international agreement.

The level of industrial involvement in data networks depends on the research domain. Most of the networks that provide services to both public and private research made no distinction between public and private use, arguing that stimulating private sector competitiveness and innovation was within their mission. However, some were considering mechanisms for private contributions towards costs.

Some successful Public Private Partnerships were in place for technology provision but others expressed worries about drivers in commercial provision being contrary to the aims of open research and open innovation. Some observed moves by very large software providers to productise the data and expressed worries about building in dependence on commercially provided services.

Recommendations

1. Data infrastructure networks need to be recognised for the vital role they play in modern research methods and dedicated structural funding needs to be provided that is sufficient to operate the network without the need to augment it with piecemeal project based funds. Long term commitments to data infrastructures need to be made at national level as these form the basis of international infrastructure provision.

2. Ad hoc mechanisms for funding international data infrastructures are suboptimal as they fail to provide the stability of funding required to deliver services as effectively and efficiently as possible. Greater coordination between funders, formalised through international agreements, is required to guarantee the stability of funding necessary to successfully deliver large scale data infrastructures.

3. Short term funding cycles are not compatible with stable provision of high quality infrastructure and research assessment methods based on novelty and innovation are not appropriate for infrastructure that needs to be reliable and secure. Funding pipelines for data infrastructure should therefore be distinct from research budgets in order to avoid tensioning of funding across activities with incompatible measures of success.