Received 24 January 2015
First revision approved
18 May 2015

Funding models for Open Access digital data repositories

Rob Kitchin
Maynooth University Social Sciences Institute, National University of Ireland Maynooth, Maynooth, Ireland
Sandra Collins
Digital Repository of Ireland, Royal Irish Academy, Dublin, Ireland, and
Dermot Frost
Trinity Centre for High Performance Computing, Trinity College Dublin, Dublin, Ireland

Abstract

Purpose – The purpose of this paper is to examine funding models for Open Access (OA) digital data repositories whose costs are not wholly core funded. Whilst such repositories are free to access, they are not without significant cost to build and maintain and the lack of both full core costs and a direct funding stream through payment-for-use poses a considerable financial challenge, placing their future and the digital collections they hold at risk.

Design/methodology/approach – The authors document 14 different potential funding streams for OA digital data repositories, grouped into six classes (institutional, philanthropy, research, audience, service, volunteer), drawing on the ongoing experiences of seeking a sustainable funding for the Digital Repository of Ireland (DRI).

Findings – There is no straight forward solution to funding OA digital data repositories that are not wholly core funded, with a number of general and specific challenges facing each repository, and each funding model having strengths and weaknesses. The proposed DRI solution is the adoption of a blended approach that seeks to ameliorate cyclical effects across funding streams by generating income from a number of sources rather than overly relying on a single one, though it is still reliant on significant state core funding to be viable.

Practical implications – The detailing of potential funding streams offers practical financial solutions to other OA digital data repositories which are seeking a means to become financially sustainable in the absence of full core funding.

Originality/value – The review assesses and provides concrete advice with respect to potential funding streams in order to help repository owners address the financing conundrum they face.

Keywords Repositories, Archives, Open Access, Open data, Funding

Paper type General review

Introduction

Societies have collected, stored and analysed data for several millennia as a means to record and manage their activities. For example the ancient Egyptians collected administrative records of land deeds, field sizes and livestock for taxation purposes, the 1086 *Domesday Book* captured demographic data and the first national registry was undertaken in Sweden in the seventeenth century (Bard and Shubert, 1999; Poovey, 1998; Porter, 1986). However, most of the data generated throughout history has been lost or destroyed because they were stored informally rather than in an institutional

Online Information Review
Vol. 39 No. 5, 2015
pp. 664-681

1468-4527
DOI 10.1108/OIR-01-2015-0031

The research for this paper was funded through the Programme for Research in Third-Level Institutions Cycle 5 administered by the Irish Higher Education Authority for the Digital Repository of Ireland.

archive, or it was decided to keep the information derived from the data (such as papers and books) which were considered more valuable, storing them in libraries. In general only the most valuable data sets were retained, such as those associated with key scientific and cultural endeavours, government records, economic transactions and legal contracts. Indeed research funders have traditionally not required projects to retain and store data, or if they did it was only for a short time. The same is true of much born digital and digitised data generated over the past 50 years which has been lost due to storage media and equipment obsolescence, bit-rot and the lack of preservation strategies and infrastructures.

Over the past two decades there has been an attempt to change this situation with the research agencies of national governments and supra-national bodies such as the European Union, along with philanthropic organisations, investing extensively in funding a wide variety of digital data infrastructures aimed at preserving and sharing scientific and cultural data. For example in Europe there are large-scale programmes such as the European Strategy Forum on Research Infrastructures and e-Infrastructures Reflection Group, and thematic large-scale European Research Infrastructure Consortia relating to supporting access to research data in the humanities and social sciences, such as Digital Research Infrastructure for the Arts and Humanities, Common Language Resources and Technology Infrastructure and the Council of European Social Science Data Archives, as well as many others related to the sciences.

Increasingly there is a commitment to making such repositories Open Access (OA) in nature. OA in its purest form is "digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber, 2013). In other words it seeks to remove both "price barriers (subscriptions, licensing fees, pay-per-view fees) and permission barriers (most copyright and licensing restrictions)" (Suber, 2013) so that material is freely available "on the public internet" and can be used for "any lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" (Budapest Open Access Initiative, 2002). Here digital data and products paid for by the public purse are seen as public goods and their sharing represents a public good. However, rather than a standard form of OA being adopted, a range of different positions have emerged that take varying stances on issues such as permission barriers, timing and who pays and how for production (given that OA is not cost free, involving significant labour, service and technology costs). Different models include gratis OA (free of charge, but not free of copyright or licensing restrictions), libre OA (free of charge and expressly permits uses beyond fair use), delayed OA (paid access initially, becoming open after a set time period), green and gold OA (pay-for-production followed by delayed publication in an OA repository or gratis OA) and so on (Suber, 2013).

There are a host of good reasons to establish and maintain OA digital data repositories (see Table I). From a scientific perspective they: facilitate the re-use of data and enable data sets to be conjoined, increasing the likelihood of new discoveries and innovations; promote research integrity through the promotion of transparency about the research process and facilitate the replication of results; enable data to be exposed to the power of computational analytics, meaning that procedures and calculations that would be difficult to undertake by hand or using analogue technologies become possible in just a few microseconds; and ensure the best opportunity for reaching as large an audience as possible (Borgman, 2007; Lauriault *et al.*, 2007). Data sharing also makes available key data for teaching, thus improving pedagogical resources. The financial benefits of data infrastructures centre on: the scales of economy created

Direct benefits	*Indirect benefits (costs avoided)*
New research opportunities	No re-creation/duplication of data
Scholarly communication/access to data	No loss of future research opportunities
Re-purposing and re-use of data	Lower future preservation costs
Increasing research productivity	Re-purposing data for new audiences
Stimulating new networks/collaborations	Re-purposing methodologies
Data available for teaching and student projects	Use by new audiences
Knowledge transfer to industry	Protecting return on earlier investment
Improves skills base	Tools and standards have potential to increase data quality
Increasing productivity/economic growth	Reduces ad hoc queries concerning data
Verification of research/research integrity	
Fulfilling mandate(s)	
Short-term benefits	*Long-term benefits*
Value to current researcher and students	Secures value to future researchers and students
No data lost from researcher turnover	Adds value over time as collection grows and develops critical mass
Widens access where costs prohibitive for researchers/institutions	Increases speed of research and time to realise impacts
Short-term re-use of well-curated data	Stimulates new research questions, especially relating to linked and derived data
Secure storage for data-intensive research	
Availability of data underpinning publications	
Private benefits	*Public benefits*
Benefits to sponsors/funder of research/archive	Input for future research
Benefits to researchers and institutions	Motivating new research
Fulfils grant obligations	Catalysing new companies and high-skills employment
Increased visibility/citation	Transparency in research funding
Commercialising research	

Source: Compiled from Beagrie *et al.* (2010) and Fry *et al.* (2008)

Table I.
Benefits of data repositories/ infrastructures

by sharing resources, avoiding replication and reducing wastage; the leveraging effects of re-using costly data where entry costs to a field might normally be prohibitive; and the generation of wealth through new discoveries (Fry *et al.*, 2008).

Aiming to leverage these benefits, in 2009 the European Union stated that: "The vision underlying the Commission's strategy on open data and knowledge circulation is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full. This means making publicly-funded scientific information available online, at no extra cost, to European researchers and citizens via sustainable e-infrastructures, also ensuring long-term access to avoid losing scientific information of unique value".

Three years later, the European Commission (2012a) re-iterated that: "Policy action on access to data is […] urgent and should be recommended to Member States". Subsequently, Horizon 2020 clearly states that the EU intends to build on OA pilot projects funded under FP7, with clear recommendations that Member States "reinforce the preservation of scientific information" (Spichtinger, 2012) and a commitment to continue to fund "relevant Open Access projects (research, coordination and support) and infrastructure" (European Commission, 2012b). Moreover, in July 2014 the European Commission (2014) launched a major public consultation on "Science 2.0", in order to develop a more open, data-driven and people-focused way of doing research

and innovation. Science 2.0 includes OA, open code, open lab-books and open data. Further, the EU Commission is also currently developing a Charter for Access to Research Infrastructures – a voluntary code of practice for transparent access to publicly funded repositories (European Association of Research and Technology Organisations, 2014). Similarly, in a major policy decision in the USA, an executive memorandum issued by the White House requires all federal agencies with research expenditures greater than $100 million per year to demonstrate how they will make taxpayer-funded research freely available to the public (Maron, 2014). Other initiatives which enable open data sharing and preservation include the global Research Data Alliance and the Digital Preservation Coalition. In other words there is a concerted drive towards ensuring that research data infrastructures are OA in nature to ensure that the data they hold are freely available for re-use.

However, whilst the rhetoric of state and supra-national agencies such as the EU suggests that OA digital data repositories will be financed entirely by the public purse through core grants, the reality for many repositories is somewhat different. For example the Netherlands' Data Archiving and Networked Services, the Netherlands Institute for Sound and Vision, UK Data Archive and Swedish National Data Service typically receive approximately 70 per cent of their funding directly from the state, having to make up the difference through other funding mechanisms. Here it is important to note that the monies involved are not trivial: OA digital data repositories are not simply data stores or back-up systems, but are actively planned, curated and managed, staffed by dedicated and specialist personnel who are dealing with multiple depositors, diverse interlinked data sets, and varying formats, standards, protocols and technologies, and seek to add value and ensure continuity (Borgman, 2007; Lauriault *et al.*, 2007; Kitchin, 2014). They tend to be much more complex to build and sustain than OA publication repositories.

Repositories that are not national in status, such as specific digital collections held by local galleries, libraries, archives, museums and universities are less likely to secure sufficient ongoing state subventions and are therefore under more pressure to identify and source other funding streams. Consequently whilst there is significant pressure being exerted on institutions to make their digital data collections open for wider use at no cost, many OA digital data repositories lack a sustainable financial model and find it difficult to fund their activities, placing their collections at risk. Identifying and assessing potential funding models to address this financial shortcoming is the central concern of this paper.

Motivation and method

Our interest in funding models for OA repositories stems from our work building the Digital Repository of Ireland (DRI; www.dri.ie). The DRI is a national research infrastructure for the humanities and social sciences that also serves as a trusted digital repository for Irish galleries, libraries, archives and museums (GLAM). Its partners include several universities and 30 stakeholder organisations, including all the major national GLAM. It is an open digital repository using open source software and open metadata CC-BY licence, and advocates for OA; however, the content owners can set the rights and access conditions, with some data being under copyright and access to sensitive social science data restricted for legal reasons. Access to the DRI collections is free to use. The DRI was initially funded for four years by the Irish Higher Education Authority through its Programme for Research in Third-Level Institutions, Cycle 5. It presently has a staff of 35 (not all of whom are full-time and a number of whom are

funded by additional research grants, or by their host institution, rather than the initial core funding). The core funding for DRI terminates in September 2015 with an expectation that it will transfer to a new financial model at this point with either no state funds or a reduced amount of state funds.

The search for a new financial model for DRI is taking place within a very challenging fiscal situation given the severe economic recession in Ireland, ongoing austerity measures, severe cutbacks in public finances (including all the major stakeholders in the repository), cutbacks to research funding and a prioritisation of remaining research funds towards industry-focused research and job creation. Moreover, the competition to secure funding has increased dramatically as agencies seek to replace lost core funding with soft monies. Establishing a new financial model that prohibits charging for use in such a context is a major challenge and requires scoping out all potential viable options.

Our prime motivation in examining OA funding models, therefore, is pragmatic given the need to source sufficient funds to continue operation of the DRI. The discussion presented in this paper presents our findings from desk-based research to identify various potential funding solutions and our evaluation of the pros and the cons of using these solutions to fund DRI. Our analysis is not based on a detailed economic cost-benefit analysis of each funding source, but rather a more general evaluation of its relative merits and the likelihood that it could be successfully deployed. What is key here is not simply whether the funding source could work in theory, but whether it would work in practice in terms of being able to convince relevant parties to back implementation and would provide sufficient funds to enable ongoing operation.

From our perspective then, the exercise was not an academic research project focused on assessing funding models for OA in the abstract, building and testing economic cost-benefit models and evaluating the soundness of financial models of other repositories. Nor was it an exercise in building or testing a theory concerned with the funding of OA models. Rather it was an exercise in producing a business case to convince government ministries, funders and various stakeholders as to a potential funding solution for DRI. As such, our approach was to search the literature for various OA funding sources, consult with other repositories, to brainstorm additional potential solutions and then to critically assess if and how they might work in practice.

In the remainder of the paper we provide a synoptic overview of the 14 potential funding sources we identified, organising them into a basic typology of six classes (institutional, philanthropy, research, audience, service, volunteer), and assess their relative merits. We next discuss the challenges that delimit what models might be pursued and the potential risks arising from failing to find sustainable funding models. Given the relative paucity of discussion concerning funding models for OA digital data repositories in the wider literature our assessment should be of interest to other repository managers seeking to identify new potential sources of funding.

Funding models for OA repositories

"For digital projects to remain vital, current, and discoverable, and be used by the people who want to use them, takes hard work from the project leaders and teams that create them. Creating a model that balances the desire to keep a resource openly available, with the need to cover the costs associated with continuing to actively develop it, is no simple task" (Maron, 2014, p. 5).

The key challenge for OA repositories that are not wholly funded by the state is to generate a sustainable funding model that ensures that the repository is maintained

and can continue to develop, providing new tools and storing new data sets, at the same time as ensuring that the repository is free to access and retains the trust of its users. In other words the challenge is to find a way to deliver core services with no or limited for-fee income. Our research into how various repositories have sought to fund their endeavours has identified 14 archetype funding sources, which can be divided into six classes (Table II), each of which varies in its potential to provide a viable funding stream.

Assembled from Ferro and Osella (2012, 2013); Maron (2014); consultation with stakeholders and team discussion.

Source income through institutional arrangements

A. Core funded. Traditionally data produced and released by the various sectors of the state has been funded by the state. In some cases the costs of producing and distributing such data have been recuperated in full or in part through cost-recovery charging. For example mapping agencies often operate as trading (cost recovery) funds, charging users to access and employ the data. Similarly libraries, national archives and statistical agencies often provide free access to resources, but charge for some specialist services or for commercial re-use of their data. Nascent research data infrastructures have followed a similar model, being core funded by research agencies and being free to access for researchers with the exception of some services. However, access for the wider public or commercial entities is often restricted, often for good reason (e.g. social science archives that house sensitive personal information).

These models of core funding are under threat in two main ways. First, the open data/OA movement has made a concerted attack on trading funds and payment for data or services. The argument advanced is that the citizens and companies have already paid for the data produced by public bodies (e.g. government departments/agencies, universities) through tax payments, and moreover opening data will produce public sector savings (by reducing transaction costs, such as staffing required for marketing, sales, communicating with customers and monitoring compliance with licence arrangements), increase taxation revenues through new innovative products that will create new markets, and leverage diverse consumer surplus value by providing significant public goods (Pollock, 2006, 2009; de Vries *et al.*, 2011; Houghton, 2011). In other words zero or marginal cost approaches are seen as being more advantageous over the long term than cost-recovery strategies (de Vries *et al.*, 2011).

Second, whilst this argument holds in theory there is little concrete evidence that open data does pay for itself in real terms, and even if it does that the corresponding savings/taxation are spent on such initiatives. In reality the massive growth in digital data and the pressure to store and retain ever more of them and to make them OA means huge pressure is being exerted on existing resources at the same time that the means to raise funds to support the development and maintenance of repositories is being restricted. Moreover, hugely increasing the number and size of OA repositories requires a commensurate increase in core funding at a time when public sector finances are under pressure to downsize. What this means is that core state funding, if it is secured, often needs to be supplemented by other sources of income.

B. Consortia (membership) model (shared service). In a consortia (membership) model, rather than a large subvention from a single state agency, many stakeholders provide subscription fees of a smaller amount. The benefit for the stakeholders is gaining access to a sophisticated shared resource and its tools that delivers more

Model	Description
Institutional	
A Core funded	The state provides the core operational costs through a subvention as with other state data services such as libraries, national archives, statistical agencies, etc
B Consortia (membership) model	A consortium collectively owns the data, pools labour, resources and tools and facilitates capacity building, but charges a membership fee to consortium members to cover shared value-added services
C Built-in costs at source	When research grants are awarded by funders, applicants must build in the costs for archiving the data and associated outputs in a repository at the end of the project. This funding is transferred to the repository for any services rendered
D Public/private partnership	The public sector provides the data and private companies provide finance and value-added services for access and re-use rights
Philanthropy	
E Philanthropy/ corporate sponsorship	Funding is sourced from philanthropic organisations as grants, donations, endowments and/or corporate sponsorship. If an endowment is sizable then core services can be funded from the interest. The donations can also be used to leverage other funding, for example, matched money from the state. This can also be reversed, so that state funding is used to try and leverage philanthropic funding/corporate sponsorship
Research	
F Research funded	The majority of funding is generated through the sourcing of research grants from national and international sources, with overheads being used to subvent core services
Audience	
G Premium product/ service	This option offers end-users a high-end product or a service that adds value to data (e.g. derived data, tools or analysis) for payment, either as fixed payment, recurrent fees or pay-per-use, without using monopoly rights. This enables the data producer to gain first-mover advantages in the marketing and the sale of complementary goods
H Freemium product/ service	This model offers end-users a graded set of options, including a free of charge option that includes basic elements (e.g. limited features or sampled data set), with more advanced, value-adding options (e.g. special formats, additional functionality, tools) being charged a fee. It opens up the product/service to a wider, low-end market and more casual use, whilst retaining a paid, high-end product/service for more specialised users
I Content licensing	This makes the data free for non-commercial re-use, but charges for-profit re-users
J Infrastructural razor and blades	An initial inexpensive or free trial is offered for products/services (razor) that encourages take-up and continued paid use (blades). It might be that access is free through APIs, but that computational usage is charged on a pay-as-you-go model, with the latter cross-subsidising the former
Service	
K Pay per purpose	There is a charge for services beyond data use, such as ingest, archiving, consulting and training services
L Free with advertising	Products/services are provided for free, but users receive advertising when using the product/service (revenue generating) or the products/services are provided by different companies and branded as such to encourage use of their other products/services (cross-subsidisation)
M White-label development/platform licensing	A customised product/service is created for a client and branded for their use, with that client paying a one-off fee or subscription that includes maintenance and update costs
Volunteer	
N Open source	This offers end-users data products/services for free, with the infrastructure maintained on a voluntary basis, including crowdsourcing

Table II.
Models of funding open repositories

collective value than any single contribution. This shared services model has been successfully employed within the public sector in many jurisdictions, across many domains and is a key part of the funding model for organisations such as the Digital Preservation Coalition. For relatively new repositories establishing a membership/shared service model can be a challenge because institutions are being asked to invest in a resource under development, rather than at maturity, at a time when their budgets are being squeezed. At the same time a shared service should help ameliorate budget cuts through the sharing of costs for a key service. If the model can be established, it provides a relatively robust, non-cyclical source of income.

C. Built-in costs at source. Many funding agencies now expect the data from the projects they fund to be deposited in an OA repository to ensure potential future re-use and to ensure research validation and integrity. The built-in costs at source model, used by UK research grant agencies and elsewhere, requires that archiving costs are factored into the original grant application. These costs are either used by the research team to prepare the data for archiving or are transferred from the grant to an OA data repository for ingest, storage and other services. This model is attractive with respect to providing a sustainable funding base for ingesting research data and for increasing the data available, but the funds typically pay for those services rather than the core costs (unless an overhead is factored in). The establishment of such a funding model is beyond the control of any single repository and is reliant on a central government mandate. Moreover, it has to be phased in over time meaning funds in its initial years will be small, though they should grow to a sustainable level. However, it should be noted that the Archaeology Data Service in the UK has found such funding to be non-linear, making it difficult to plan around.

D. Public/private partnership. Public private partnerships (PPPs) have been used extensively by governments over the past couple of decades to co-fund the development of public infrastructure such as housing, roads and service provision. Such partnerships only work where there is a clear benefit to both parties, delivering a profit to the private partner. While PPPs might have a role in repository projects, with the private partner making money from advertising revenue, ingest services, white-label development or by producing commercial products from the archived data, the success of such a venture will, in large part, be dependent on the type of data being archived. Data sets such as transport, weather, health and map data all have potentially high commercial value. However, cultural heritage and data from relatively esoteric research projects have much weaker direct commercial value. It is therefore likely that PPPs will only be an attractive option where the private partner can envisage some means to leverage the data, or attract traffic to the site or are getting involved on a philanthropic basis.

Source income through philanthropy

E. Philanthropy. Philanthropy is an important source of funding for research in many nations. Philanthropy might therefore be a key source of funds for archiving the data resulting from these projects. It might also be the case that philanthropic donations can be used to leverage matched state funds, or alternatively state funding is used to try and leverage philanthropic funding or corporate sponsorship. There are two issues with philanthropic funding. First, it is usually best sourced with respect to specific sub-projects rather than core activities. Second, it is cyclical in nature, meaning it is difficult to plan multi-annual budgets given the uncertainties over funds raised.

Source income through research

F. Research funded. Many aspects of research data infrastructures are funded through research funding, including the building of an infrastructure itself and projects that add to and utilise it. However, research funding typically does not cover core maintenance costs, but funds new developments. Contractual obligations with respect to these grants mean that funds cannot be diverted for non-project purposes. And while research grants typically have associated overheads, it would take a continuous supply of very large volumes of research income to provide sufficient overhead to fund core costs in addition to the costs of running the new projects including such overhead items as office space, facilities, etc. Research funding is also highly competitive (and getting more so) and cyclical, meaning that it cannot be relied on to provide a sustained income stream. That said, research funding can form an important part of a blended model of repository income.

Source income from audience

G. Premium product/service. In the absence of core or subscription funding, funds can be raised through the selling of services. A premium product/service approach involves selling end-users a high-end product or a service that adds value to data and they cannot gain elsewhere. Such a premium approach works best with data that has a high commercial utility and will add value to the work being undertaken by the purchaser. However, it also runs against the ethos of OA to date and is therefore of limited utility to OA repositories.

H. Content licensing. Depending on the content, a potential source of funding is content licensing. Here content such as art images, manuscript screenshots and audio-visual files is made available for commercial re-use in publishing, media and advertising/marketing. Such content licensing can be highly profitable if well organised, with some select digital archives in the UK and France generating revenue in the hundreds of thousands of euros (Maron, 2014). To be able to licence content the repository must either own the content, or have struck a deal with those that do. There are also associated costs, with the ability to realise fees requiring cost recovery, licensing and marketing expertise and resources. Again, the funding stream is likely to be cyclical and difficult to predict, and also at odds with OA.

I. Infrastructural razor and blades. This is a commercial funding model for encouraging initial usage that might translate into a paid service. Users are given an initial trial run. When this expires they are offered continued service for a fee. This might be combined with a freemium model, though it clearly works against the wider ethos of OA.

Source income through services

J. Freemium product/service. Some new data infrastructures, such as Dublinked (www.dublinked.ie/), have been experimenting with freemium product/services. All users are offered a free of charge set of options that include basic functionality and key data sets. However, for a fee additional services are available. Maron (2014) identifies six types of such value-added services: charging for a higher-quality version; charging for additional formats; charging for additional features; offering more storage for a fee; charging for an advertising-free environment; and charging for different end uses (free for education and non-for-profit use, but charging for commercial use). A freemium model is a more attractive option than the premium model, but still means that some of the infrastructure is not OA. That said, it might provide some sustainable

lines of funding whilst providing a workable free service for non-specialist users. To generate sizable income it would require a large number of users to opt for the paid services, which will depend on the value of the data sets to users, with many research data sets having intrinsic rather monetary than value.

K. Pay per purpose. This is a form of cost recovery for specific services such ingestion, archiving, consulting and training, with data access being free. As with research funding, the monies are to be used to provide the services paid for and cannot be simply diverted to cover core costs, though any overhead on such payments could be used in this way. Moreover, it is a cyclical source of potential income. The extent to which such service provision can provide a viable funding stream is dependent on potential demand, which will vary between repositories in line with expertise levels across depositors and their ability to pay.

L. Free with advertising. Many internet services such as Google, Twitter, Flickr and Facebook offer their services to users for free, funding their services through advertising revenue (and also selling data about users to data brokers). However, such a model requires a high volume of site visits to provide a sustainable source of income. For example Maron (2014) reports that to generate US$50,000 a year in advertising revenue, a web site would need around two million page views annually. Given that most research repositories are serving quite small constituencies of academics and interested commercial and lay readers, site traffic is likely to be quite modest and advertising revenue therefore small. There is also a wider question as to whether public sites should be delivering commercial advertising content.

M. White-label development. This is another internet funding model where versions of a web service are tailored and branded for a specific entity for a fee or subscription. Here the repository and its underlying architecture is used as the "engine" for other initiatives. For example in our case the DRI content and back-end architecture was used for an Irish government web site "Inspiring Ireland", where the front page and the look and feel of that site is independent of the DRI site. In this sense "Inspiring Ireland" (www.inspiring-ireland.ie/) is powered by DRI hardware, software and expertise, but this is not immediately obvious to users. Ongoing maintenance of the site is either taken on in-house by those who commissioned the white-label development or paid for through an ongoing service contract. Again, such initiatives pay for a specific service, with only overhead contributing to core costs, and IP ownership needs to be treated carefully.

Volunteered resourcing

N. Open source. Enterprises such as Open Street Map and Wikipedia use the power of crowdsourcing and voluntary labour to create comprehensive mapping and encyclopaedia data that are free to use. Whilst crowdsourcing has its benefits, bringing many minds to focus on a task, it is notoriously difficult to mobilise and manage a crowd and to keep it motivated, and to assure data quality, integrity and standards (Carr, 2007; Dodge and Kitchin, 2013). Whilst an open source approach to OA data repositories might include the running of hackathons to develop new tools and APIs, or to source specific data, it is unlikely that it can be relied upon to provide core services for a long-term repository that requires specialist knowledge, trust and continuity, except in a few specific cases where there might be significant buy-in by potential users and where the service is cross-subsidised by other projects (providing necessary infrastructure and staffing, for example, through research projects).

Challenges in funding OA data repositories

Identifying and rolling out potential funding streams is no easy task and it is made more fraught by a set of challenges that provide context and frame the options open to those operating repositories. These challenges take two forms, general and specific, and also create a set of risks that potentially jeopardise the realisation of a sustainable funding model.

General challenges

A key general challenge that is beyond the control of a repository is the financial and political climate in which it operates. There needs to be political will not just for the notion of OA, but to fund it in practice, and the state and funding agencies have to be in a position to supply such funding, and to coordinate their approach, policies and even legislation. In the context of DRI, as noted, the Irish state is presently severely constrained financially, and there are multiple demands for what funding is available. This places an inherent constraint on sourcing core additional funding.

Another challenge facing many repositories is persuading data holders to share a valuable commodity. An underlying principle of academic research is that all aspects of knowledge production should be freely available for others to inspect and test through replication. In practice this principle has never worked optimally as researchers are often reluctant to share data which has been time consuming and costly to produce and provides a competitive advantage in advancing knowledge production. As Borgman (2007) notes, sharing is only common in a handful of disciplines such as astronomy, genomics and geomatics which rely on large, distributed teams and large and expensive equipment and infrastructure where research funding agencies have demanded collaboration in return for the massive investments required. In other disciplines it is shared occasionally or not at all. She concludes that "[t]he 'dirty little secret' behind the promotion of data sharing is that not much sharing may be taking place" (Borgman, 2012, p. 1059), noting a number of disincentives to the sharing of data:

- a lack of rewards to do so;
- the effort required to prepare and archive the data;
- a lack of expertise, resources and tools to archive data;
- concerns over being able to extract value prior to others in terms of papers and patents given the effort invested in generating the data;
- concerns over how the data will be used, especially if they relate to people, or how they might be mishandled or misinterpreted;
- worries over the data generating queries and requests that will create additional work;
- concerns over issues with the data being exposed and research findings being undermined through alternative interpretations of the same data;
- intellectual property issues; and
- a fear that the data will not be used, thus archiving constituting a wasted effort (Borgman, 2007, 2012; Strasser, 2013).

As such, ensuring data are archived for future re-use requires more than creating OA data repositories; it is going to require a cultural change in research practices. This change is starting to be driven using a carrot and stick strategy. On the one hand,

incentives are starting to be used to encourage researchers to deposit data, such as promoting data citation and attribution (Borgman, 2012), and building adequate funding for archiving into grant awards. Standardised data citation is however still in its infancy and needs to be adopted by the major publishers. On the other hand, research agencies are starting to compel researchers to deposit data, taking into account ethical and IPR issues, as a condition of research funding. Importantly the funding mechanisms for supporting OA can be a vital part of strategies designed to compel researchers to deposit data. Without such strategies it is likely that the move to OA data repositories will be stymied by resistance from researchers.

Specific challenges

Specific challenges relate to particular conditions of individual repositories, with the adoption of any funding model having to align to its ethos and position in its life cycle, operating policies and licensing requirements of software adopted. It must also consider who will use the data and how that data will be used. If charging does occur it will need to have a clear, justified and transparent cost model. By way of illustration we discuss these issues with respect to DRI.

At the time of formulating its future plan with respect to financing its activities the DRI was in year three of a four year programme of development, testing and roll-out. It was therefore at an immature phase with only a pre-launch demonstration version that lacked full functionality to show potential funders and stakeholders. Typically repositories require core funding until a project is not just complete but has reached maturity, with its value to stakeholders firmly established and able to be proven using metrics. Trying to transfer from core funding to other sources, or even to significantly reduced core funding, is therefore difficult as it requires investors to have faith and trust in a largely unproven endeavour, and exposes it to major risks with respect to sustainability. Moreover, the kinds of data that DRI stores have weak direct commercial value, restricting the viability of some potential funding streams.

Moreover, choices made with respect to the technology used and software licences placed limits on the ability to charge for use of the software and also obligated the project to adopt an open source ethos and contribute back to the wider development of such software. In its design and requirements phase DRI made the decision to use a number of open source software components such as Hydra (interface framework), Fedora Commons (core data repository), Apache SOLR (search) and CEPH (preservation), the first three of which are used under an Apache 2 licence, the latter a LGPL licence. The Apache 2 licence allows DRI to use, modify and re-distribute the code for any purpose with no royalty issues. The LGPL requires any modifications made to the code to be released under an LGPL (or compatible) licence. The terms of these various open source licences make it difficult, if not impossible, to charge for the software itself. Instead, most business models using such software are built around support services (consultancy, hosting, documentation and training) and development on demand.

DRI is committed to open, free access to data wherever possible, but makes a distinction between access to data and provision of services such as ingestion and preservation services, recognising that it will need to charge for them because they involve significant time, expertise, labour and resources beyond maintaining core functions. In charging for these services, however, consideration needs to be given to the nature of this charging and whether the model being pursued seeks profit maximisation or cost recovery/partial cost recovery (Pollock, 2009). Given DRI's

mandate to be a public service and serve the public good, profit maximisation is not an option. Without subvention through core funding, partial cost recovery is also not a viable option. It is therefore trammelled into using a full cost-recovery model for services, but to do so requires establishing a charging model. Cost models assess the costs of services, factoring in key figures relating to operational areas such as administration, ingestion and validation, format migration, upgrading hardware, retrieval and dissemination of content, and preservation planning. Established cost models for preservation generally align with best practice preservation processes (e.g. OAIS) and quantify the value of services to stakeholders, funders and end-users; justify their costs in providing these services; and provide transparency and accountability in charging. A number of EU and international projects have developed and published cost models and cost modelling tools aimed at repositories undertaking digital preservation and/or curation. These tools provide a framework by which costs can be estimated or assessed, and determine either broad projected costs or specific figures, depending on the tool used and the data entered. Some available cost modelling tools and projects include:

- cost model for digital preservation, developed by the Royal Danish Library and the Danish National Archives (www.costmodelfordigitalpreservation.dk/);
- cost modelling for sustainable services, by California Digital Library/Technology at Berkeley (https://wiki.ucop.edu/download/attachments/163610649/TCP-total-cost-of-preservation.pdf);
- digital preservation for libraries, developed by the Deutsche Nationalbibliothek (http://dp4lib.langzeitarchivierung.de);
- keeping research data safe project, led by Charles Beagrie Ltd with funding from JISC (www.beagrie.com/krds.php);
- 4C: collaboration to clarify the costs of curation, EU funded project launched February 2013 (www.4cproject.eu/); and
- life cycle information for E-Literature (LIFE), a collaborative project undertaken by University College London and the British Library (www.life.ac.uk/tool/).

Although published cost modelling tools appear to provide generic cost modelling services to repositories, they nearly always require adjustments to cater to specific projects and use-cases. The APARSEN (2013) project report on cost models for digital repositories maps how the cost parameters generally used in these projects can be assessed against the activities defined by the OAIS model and the International Standard for Trusted Repositories (ISO 16363).

Risks associated with failing to secure a sustainable funding model

Failing to secure a funding model or to create a robust and transparent cost-recovery model puts an OA repository at risk and creates potential impediments to future investment in at least four ways. First, and the most significant risk, is that the repository closes because it cannot cover its core costs. Clearly the raison d'être of repositories is that they preserve and make data collections available for re-use. However, in the case of closure, unless a repository's collections can be transferred elsewhere the danger is that important and valuable data sets will be potentially lost, denying access to researchers, students, citizens and companies and foreclosing the insights and knowledge that might be drawn from such collections. Moreover, if

existing and potential depositors start to become worried that a repository is going to vanish it will undermine trust and faith in the integrity of the repository and affect their willingness to deposit data.

Second, there is the risk of major reputational damage to those associated with the repository and its original funders. Significant financial resources, as well as time, energy and political and social capital, will have been invested in establishing a repository in the expectation that it will operate as a long-term endeavour. A repository closing tarnishes all those involved, but also seriously impedes the ability to remobilise those resources and capital in the future to re-establish, or create a new, repository. In our own case we have built an extensive network of key stakeholders, persuading them to invest and back the repository and to collaborate with each other. If the repository was to close, failing to fulfil its promise, it would take a significant effort to persuade these stakeholders to re-invest in a new, replacement repository, to undertake necessary related work and redirect their own resources, and to assure them that the project would not fail again for the same reasons, namely, the lack of a sustainable funding model.

Third, repositories do not just consume financial resources, but also enable funds to be leveraged on their technologies and collections. For example the repository partners and stakeholders can apply for research and other funds to extend and develop the repository, or to analyse the data held within them. In some cases funding is dependent on a repository existing. For example some projects are now only funded if there is a guarantee that the data and publications from a project are deposited for re-use. A repository closing would foreclose any such leveraging.

Fourth, the closure of a repository leads to a significant loss of consolidated human resource expertise, stakeholder networks, technical infrastructures, and the legal and policy frameworks developed. Again it takes time, energy and resources to assemble necessary staffing and communities, and to put in place technology, software, databases and frameworks. Re-establishing a repository will not simply be a case of assembling a new team of actors, but will require all the other elements to be rebuilt from the ground up in line with latest best practices and technical specifications. In other words the benefits of the investment in a repository will largely disappear, with any attempt at re-establishing it being just as costly.

Rather than closing altogether it might be the case that a repository can continue operation but on a reduced basis. For example enough funding might be secured to run the repository using a skeleton staff, limiting the work it can perform and forgoing additional development work or the addition of new data sets. While this might be a short-term, plug gap solution it will create progressively more harm the longer the arrangement persists. Over time a funding model cannot simply maintain present resources but needs to enable investment in new technologies and platforms to allow data to be migrated as machines come to the end of their operational life and to take advantage of new software and techniques. Indeed digital data are highly vulnerable to loss due to obsolescence in software and hardware. As O'Carroll *et al.* (2013) note, "While it is possible for anyone to pick up, look at and read a page from a book written 100 years ago, the same would not be true of a floppy disk containing Word Perfect files from 20 years ago". Without such costs and financial stability, the risk is of "digital decay" and the repository failing to evolve to meet user expectations (Maron, 2014). At the same time raising necessary leveraged finance needs to be balanced against the core mission of the repository to avoid drift through "following the money". Digital preservation is a long-term core commitment.

Conclusion

Significant investment is directed at funding research. Such research produces much data and outputs and there is now significant political pressure to make these openly accessible through digital repositories for no cost. While such an aim makes sense in terms of transparency, accountability and scientific endeavour, there are significant legacy issues to be dealt with regarding existing dissemination models, the funding of those models, and researcher practices. As a result a number of different OA models have been developed with respect to publications. However, the development of finance models for OA data repositories is lacking. Such repositories are much more demanding to build and maintain than publication repositories given the diversity of the data to be stored and associated standards, protocols, legal obligations, and the need for active curation and management. They are therefore not without significant cost to build and maintain.

In this paper we have sought to document and critically examine 14 different potential funding streams, grouped into six classes (institutional, philanthropy, research, audience, service, volunteer), for OA research data repositories. With the exception of core funding from a state agency, each of these funding streams have associated issues, such as being cyclical, creating new services rather than supporting the core functions, and they undermine the notion of an open, free resource. Moreover, a repository seeking to create sources of income faces a number of challenges, some relatively generic such as austerity and competition with respect to public finances and a reluctance on behalf of researchers to deposit data, and some more specific relating to the choices and decisions made with respect to the ethos of the repository and its technology and software.

The critical issue is that regardless of the various constraints and difficulties OA repositories do need to find ways to fund their activities or they place the collections they hold at significant risk, as well as risking loss of expertise, trust, stakeholder networks, technical infrastructures, and the legal and policy frameworks developed that have been created at some expense. In formulating a funding model for DRI our strategy has been to create a blended model that seeks to ameliorate cyclical effects across funding streams by seeking income from a number of sources rather than relying on a single one. It is clear from our analysis, however, that a large proportion of the budget will need to continue to be core funding, with other prioritised sources of funding (stakeholder membership fees, built-in costs at source, leveraged research income, philanthropy, paid for specialist services and white-label development/ platform licensing) providing a smaller proportion of income. In our business plan we have this set upon a sliding scale with core funding reducing over time to a ceiling and other funding streams making up the difference.

Whether this business plan is achieved is at present still an open question. Moreover, even if it is accepted, the other funding streams still have to be realised: stakeholders persuaded to pay membership fees, grants to be secured, philanthropists persuaded to donate and services to be sold. In other words in the absence of sufficient core funding the struggle to source income will be an ongoing endeavour. Given that other existing national data repositories are funded in such a fashion suggests that this precarious situation will become the norm for many OA repositories, and the degree of insecurity will increase for more localised repositories. This clearly has to be a source of concern as it places OA repositories at risk. As such, whilst the arguments advocating OA are important, just as salient are further debates and models as to how such repositories should be funded. To date there has been little concerted attention paid to this

conundrum and our intention has been to fill in part this lacuna. A very useful extension to our initial mapping would be a set of economic cost-benefit analysis studies that would evaluate the 14 potential sources we have identified and the relative merits of different blended approaches, as well as assess the financial security and sustainability of OA digital data repositories that have adopted particular funding models.

References

APARSEN (2013), "Report on cost parameters for digital repositories", Didcot, Oxfordshire, 28 February, available at: www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D32_1-01-1_0.pdf (accessed 13 January 2015).

Bard, K.A. and Shubert, S.B. (1999), *Encyclopaedia of the Archaeology of Ancient Egypt*, Routledge, London.

Beagrie, N., Lavoie, B. and Wollard, M. (2010), *Keeping Research Data Safe 2*, JISC, London, available at: www.beagrie.com/jisc.php (accessed 22 May 2015).

Borgman, C.L. (2007), *Scholarship in the Digital Age*, MIT Press, Cambridge, MA.

Borgman, C.L. (2012), "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078.

Budapest Open Access Initiative (2002), "Read the Budapest open access initiative", available at: www.budapestopenaccessinitiative.org/read (accessed 24 October 2014).

Carr, N.G. (2007), "The ignorance of crowds", *Strategy+Business Magazine*, Vol. 47, Summer, pp. 1-5.

de Vries, M., Kapff, L., Negreiro Achiaga, M., Wauters, P., Osimo, D., Foley, P., Szkuta, K., O'Connor, J. and Whitehouse, D. (2011), "Pricing of public sector information study (POPSIS)", available at: http://epsiplatform.eu/sites/default/files/models.pdf (accessed 11 August 2013).

Dodge, M. and Kitchin, R. (2013), "Crowdsourced cartography: mapping experience and knowledge", *Environment and Planning A*, Vol. 45 No. 1, pp. 19-36.

European Association of Research and Technology Organisations (2014), "European charter for access to research infrastructures", Brussels, available at: www.earto.eu/fileadmin/content/04_Newsletter/Newsletter_3_2014/13_may__Draft_European_Charter_for_Access_to_Research_Infrastructures.pdf (accessed 13 January 2015).

European Commission (2012a), "C(2012) 4890 final, 17.7.2012: commission recommendation on access to and preservation of scientific information", Brussels, available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf (accessed 24 October 2014).

European Commission (2012b), "European commission background note on open access to publications and data in horizon 2020", available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/background-paper-open-access-october-2012_en.pdf (accessed 24 October 2014).

European Commission (2014), "Have your say on the future of science: public consultation on science 2.0", available at: http://europa.eu/rapid/press-release_IP-14-761_en.htm (accessed 24 October 2014).

Ferro, E. and Osella, M. (2012), "Business models for PSI re-use: a multidimensional framework", paper presented at Using Open Data: Policy Modeling, Citizen Empowerment, Data Journalism, European Commission Headquarters, Brussels, 19-20 June, available at: www.w3.org/2012/06/pmod/pmod2012_submission_16.pdf (accessed 22 May 2015).

Ferro, E. and Osella, M. (2013), "Eight business model archetypes for PSI re-use", Open Data on the Web, Workshop, London, 23-24 April, available at: www.w3.org/2013/04/odw/odw13_submission_27.pdf (accessed 13 August 2013).

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J.W. and Rasmussen, B. (2008), "Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes", available at: http://repository.jisc.ac.uk/279/ (accessed 21 October 2013).

Houghton, J. (2011), "Costs and benefits of data provision", report to the Australian National Data Service, Centre for Strategic Economic Studies, Victoria University, Melbourne, September, available at: http://ands.org.au/resource/houghton-cost-benefit-study.pdf (accessed 14 August 2013).

Kitchin, R. (2014), *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, Sage, London.

Lauriault, T.P., Craig, B.L., Taylor, D.R.F. and Pulsifier, P.L. (2007), "Today's data are part of tomorrow's research: archival issues in the sciences", *Archivaria*, Vol. 64, Fall, pp. 123-179.

Maron, N. (2014), *A Guide to the Best Revenue Models and Funding Sources for Your Digital Resources*, Ithaka S+C and JISC, New York, NY.

O'Carroll, A., Collins, S., Gallagher, D., Tang, J. and Webb, S. (2013), *Caring for Digital Content, Mapping International Approaches*, NUI Maynooth, Trinity College Dublin, Royal Irish Academy and Digital Repository of Ireland, Dublin.

Pollock, R. (2006), "The value of the public domain", available at: www.ippr.org/publication/55/1526/the-value-of-the-public-domain (accessed 13 August 2013).

Pollock, R. (2009), "The economics of public sector information: profit-maximisation, cost-recovery, marginal costs and zero costs", Working Paper Economics No. 0920, Cambridge, available at: www.econ.cam.ac.uk/research/repec/cam/pdf/cwpe0920.pdf (accessed 13 August 2013).

Poovey, M. (1998), *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*, University Chicago Press, Chicago, IL.

Porter, T.M. (1986), *The Rise of Statistical Thinking*, Princeton University Press, Princeton, NJ.

Spichtinger, D. (2012), "Open access in horizon 2020 and the European research area", available at: www.scienceeurope.org/uploads/GRC/Open%20Access/2_DanielSpichtinger.pdf (accessed 24 October 2014).

Strasser, C. (2013), "Closed data … excuses, excuses", *Data Pub: California Digital Library*, 24 April, available at: http://datapub.cdlib.org/2013/04/24/closed-data-excuses-excuses (accessed 18 September 2013).

Suber, P. (2013), "Open access overview", available at: http://legacy.earlham.edu/~peters/fos/overview.htm (accessed 24 October 2014).

About the authors

Rob Kitchin is a Professor and an ERC Advanced Investigator at the National University of Ireland Maynooth. He has published widely across the social sciences, including 23 books and over 140 papers and book chapters. His latest book is *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (Sage, 2014). He is the Editor of the journal *Dialogues in Human Geography*, and was the Editor-in-Chief of the 12 volume *International Encyclopedia of Human Geography*. He is currently a Principal Investigator for the Digital Repository of Ireland, Programmable City, the All-Island Research Observatory and the Dublin Dashboard. His book *Code/Space* (with Martin Dodge) won the Association of American Geographers' Meridian Book Award for the outstanding book in the discipline in 2011, and he was the 2013 Recipient of the Royal Irish Academy's Gold Medal for the social sciences. Professor Rob Kitchin is the corresponding author and can be contacted at: Rob.Kitchin@nuim.ie