Funding models for Open Access Repositories

First published in 2015 by the Royal Irish Academy.

When citing this report, please use the following wording: Kitchin, R., Collins, S. and Frost, D. (2015), Funding models for Open Access Repositories. Maynooth: Maynooth University. Dublin: the Royal Irish Academy and Trinity College Dublin.

DOI: http://dx.doi.org/10.3318/DRI.2015.4

Design and layout by Fidelma Slattery

Summary

Across jurisdictions and domains (academia, government, business) there has been much recent attention paid to open forms of knowledge production (e.g., open-source software, open data/metadata, open infrastructures) and the creation of open digital repositories for the unrestricted sharing of data, publications and other resources. This report focuses on the latter, documenting and critically examining 14 different funding streams, grouped into six classes (institutional, philanthropic, research, audience, service, volunteer), being pursued by open digital repositories to support their endeavours, with a particular focus on academic research data repositories. While open digital repositories are free to access, they are not without significant cost to build and maintain, and unstable and cyclical funding poses considerable risks to their future and the digital collections they hold. While the political and ethical debate concerning the merits of open access and open data is important, we argue that just as salient are concerns with respect to long-term, sustainable funding for the operation and maintenance of open access digital repositories.

Introduction

The founding of the internet was a significant disruptive innovation with respect to the publishing and sharing of data, information and knowledge. Progressively, online publication and databases have undermined traditional barriers to distributing and accessing the fruits of academic labour (e.g., papers, books, data), and created new forms of scholarly communication (e.g., social media), by enabling thoughts and files to be easily disseminated and accessed through ICT networks. Until recently, however, traditional forms of publishing and academic practices have remained remarkably robust, with academics largely preferring to publish in well-established, for-profit, peer-review journals and with print presses, and to hoard rather than share data. In part this is inertia, but it is also due to perceptions about quality, standards, the ways in which academic labour is assessed with regard to worth, and ingrained academic practices including career progression models built on traditional academic outputs. Current debates concerning open access publishing and the opening and sharing of data, and changes in the terms and conditions of research funding, are set to transform which academic outputs are disseminated and how.

> “ open access in its purest form is “digital, online, free of charge, and free of most copyright and licensing restrictions”

Put simply, open access 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, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles [or databases], crawl them for indexing, pass them as data to software, or use them for any other 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, academic outputs are seen as public goods, having largely been paid for by public monies (through core state funding to universities and research funding through state agencies), and their sharing represents a public good. In the ensuing debate a range of different open access positions have emerged that take varying positions on issues such as permission barriers, timing, and who pays for production and how (given that open access is not cost free, involving significant labour, service and technology costs), including gratis OA (free of charge, but not free of copyright of 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 delayed publication in an open access repository or gratis OA), and so on (Suber 2013).

Internationally there has been significant adoption of open access policies to research publications. For example, by October 2014, the ROARMAP project had documented over 90 policies, drawn from over 45 countries, in which funding agencies mandated open access to research publications. The European Commission expresses its vision on open access as follows:

> "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." (European Union 2009)

In Horizon 2020 all funded projects will be mandated to provide open access to peer-reviewed publications.

The natural progression from opening publications to wider access, to opening up other academic outputs such as research data and research infrastructures is also underway. Over the past two decades the research agencies of national governments and supra-national bodies such as the European Union, along with philanthropic organisations, have invested extensively in funding a wide variety of data infrastructures. For example in Europe there are large-scale programmes such as European Strategy Forum on Research Infrastructures (ESFRI) and e-Infrastructures Reflection Group (e-IRG), and thematic large-scale European Research Infrastructure Consortiums (ERICs) relating to supporting access to research data in the humanities and social sciences, such as DARIAH (Digital Research Infrastructure for the Arts and Humanities), CLARIN (Common Language Resources and Technology Infrastructure), and CESSDA (Council of European Social Science Data Archives), as well as many others related to the sciences. 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[1]. Other initiatives which enable open data sharing and preservation include the global Research Data Alliance (RDA) and the Digital Preservation Coalition (DPC). In 2012 the EU Commission re-iterated their commitment to open access, broadening its focus to research data, noting that:

> "Open access to scientific research data enhances data quality, reduces the need for duplication of research, speeds up scientific progress and helps to combat scientific fraud. ... [T]he High Level Expert Group on Scientific Data emphasised the critical importance of sharing and preserving reliable data produced during the scientific process. Policy action on access to data is therefore urgent and should be recommended to Member States" (European Commission 2012a)

Subsequently, Horizon 2020 has clearly stated that they intend to build on open access 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 open access, open code, open lab-books and open data. Similarly, in a major policy decision in the United States, 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). In other words, there is a concerted drive towards ensuring that research data infrastructures are open access in nature to ensure that the data they hold are freely available for re-use.

This move towards open access research data has been accompanied by a more broadly focused open data movement that has developed in tandem with the right to information (RTI) movement (freedom of information) and open government[2]. The movement is built on three principles: openness, participation and collaboration (White House 2009); that through transparency, sharing and working together the value of data for society can be realised. In particular, attention has been focused on opening data that has been produced by state agencies (often termed public sector information/PSI) for re-use. Since the late 2000s the movement has gained traction with dozens of countries and international organisations (e.g., EU, UNDP), making thousands of previously

[1] http://www.earto.eu/fileadmin/content/04_Newsletter/Newsletter_3_2014/13_may__Draft_European_Charter_for_Access_to_Research_Infrastructures.pdf, last accessed 15 April 2015.
[2] http://www.opengovpartnership.org/, last accessed 15 April 2015.

restricted datasets open in nature for non-commercial and commercial use (see DataRemixed 2013). Such a shift in position has been facilitated by influential international and national lobby groups such as the Open Knowledge Foundation and the Sunlight Foundation, accompanied by the lobbying of knowledge economy industry groups and companies, and local citizen groups seeking to leverage municipal data.

In this report, we focus our attention on open access research data repositories and in particular how they are funded. We start by outlining the logic, work and benefits of digital data repositories. We note that while the arguments in favour of open access data repositories are compelling, most initiatives are funded precariously. This is followed by a critical examination of 14 different funding models, grouped into six classes (institutional, philanthropy, research, audience, service, volunteer), that might be used to provide revenue streams to support their work. We next discuss the challenges that delimit what models might be pursued and the risks of failing to find sustainable funding models, drawing on our own experience of seeking continuation funding for the Digital Repository of Ireland (DRI[3]; www.dri.ie), an initiative funded for four years by the Irish Higher Education Authority through its Programme for Research in Third Level Institutions, Cycle 5. We conclude that while much critical attention has focused on the relative merits of open access initiatives, much less consideration has been paid to how such initiatives are to be sustained in the absence of payment to access. Whilst open digital repositories are free to access, they are not without significant cost to build and maintain, and unstable and cyclical funding poses considerable risks to their futures and the digital collections they hold. It is therefore vital to develop sustainable funding models to support their long term future and ensure their benefits are realised.

> **“ while the arguments in favour of open access data repositories are compelling, most initiatives are funded precariously”**

[3] The Digital Repository of Ireland (DRI) is a national research infrastructure for the humanities and social sciences that also serves as a trusted digital repository for the Irish GLAM sector (Galleries, Libraries, Archives, and Museums). The DRI is an open digital repository using open source software and open metadata CC-BY licence, and advocates for open access, however the content owners can set the rights and access conditions, with some data under copyright and access to sensitive social science data restricted for legal reasons.

Digital data repositories

Societies have collected, stored and analysed data for a couple of 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 Doomsday 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, not in a formalised archive, or it was decided to keep the information derived from the data (such as articles and books) which were considered more valuable, storing them in libraries. In general, only the most valuable datasets were retained, such as those associated with key scientific and cultural endeavours, government records, economic transactions, and legal contracts. The data of most scientists have been informally stored in files and boxes or on various hard-drives in their offices or at home. When they retire or die most of their effects are destroyed, and along with them any data they generated. The vast bulk of data generated for doctoral theses are lost after completion. 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 development of digital storage solutions, which reduce the cost and space of retaining data, makes the wide-scale, long term storage of routine and lower-value data seem obtainable. However, unless such storage is formalised into archives and repositories (collections of archives), it is likely that they will ultimately go the same way as informal paper stores. Indeed, it is already clear that, despite significant investment in their creation, much of our recent born digital and digitised data has been lost, along with its cultural and economic value, due to storage media and equipment obsolescence, bit-rot, and the lack of preservation strategies and infrastructures.

Archives and repositories marry curation practices with institutional structures to ensure that data are preserved for future generations, whilst complying with legislation relating to access, privacy, ethics, copyright and intellectual property rights that delimits who can access data and what they can do with them. They are not simply data stores or back-up systems, but are actively planned, curated and managed, staffed by dedicated and specialist personnel who add value and ensure continuity (Borgman 2007; Lauriault *et al*. 2007; Kitchin 2014). Moreover, an archive seeks to preserve the full record set, not simply the data; that is, all supporting documentation, metadata, and other related material that details provenance and context with respect to how the data were generated and should be treated, analysed and interpreted. The approach to preservation is mindful that technologies, protocols and best practice guidelines are subject to change and obsolescence, and that data will need to be migrated across platforms and technologies as new innovations come on stream, and that without active curation data may become corrupted, lost or shorn of its contextual metadata and supporting documents (Borgman 2007; Dasish 2012). Further, in many cases the archive tries to ensure interoperability between datasets by seeking common technical specifications relating to formats, standards, and protocols. By maintaining the integrity of the data over time, the archive becomes a resource that is trusted as a safe and reliable place to store, access and share data.

There are a host of good reasons to establish and maintain digital data archives and repositories (see Table 1). From a scientific perspective they: facilitate the re-use of data and enable datasets to be conjoined, increasing the likelihood of new discoveries and innovations; promote research integrity through the promotion of transparency about the research process and facilitating 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, improving pedagogical resources. The financial benefits of data infrastructures centre on the scales of economy created 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).

As more and more research data and information are born digital or are digitised it is vital then to put in place and sustain digital repositories that will maintain the records of the past and present for future generations and for re-use. The resulting open access repositories will constitute critical research infrastructure that have significant spill-over benefits. And yet, most digital archives and repositories, with the exception of some national initiatives, are funded precariously, perhaps receiving initial core funding through research agencies and then seeking to survive by raising soft monies generated through a variety of sources. Consequently, they face significant challenges in ensuring their continued operation, which in turn creates large risks vis-a-vis their collections. This is a different situation from national libraries and national archives charged with preserving a nation's paper records; while their funds may be decreasing due to austerity, there is an expectation these institutions will be funded in perpetuity and not on a project basis.

Table 1: Benefits of data repositories/infrastructures

Direct benefits

- New research opportunities.
- Scholarly communication/access to data.
- Re-purposing and re-use of data.
- Increasing research productivity.
- Stimulating new network/collaborations.
- Data available for teaching and student projects.
- Knowledge transfer to industry.
- Improves skills base.
- Increasing productivity/economic growth.
- Verification of research/research integrity.
- Fulfilling mandate(s).

Indirect benefits (costs avoided)

- No re-creation/duplication of data.
- No loss of future research opportunities.
- Lower future preservation costs.
- Re-purposing data for new audiences.
- Re-purposing methodologies.
- Use by new audiences.
- Protecting return on earlier investment.
- Tools and standards have potential to increase data quality.
- Reduces ad-hoc queries concerning data.

Near term benefits

- Value to current researcher and students.
- No data lost from researcher turnover.
- Widens access where costs prohibitive for researchers/institutions.
- Short term re-use of well curated data.
- Secure storage for data intensive research.
- Availability of data underpinning publications.

Long term benefits

- Secures value to future researchers and students.
- Adds value over time as collection grows and develops critical mass.
- Increases speed of research and time to realise impacts.
- Stimulates new research questions, especially relating to linked and derived data.

Private benefits

- Benefits to sponsors/funder of research/archive.
- Benefits to researchers and institutions.
- Fulfil grant obligations.
- Increased visibility/citation.
- Commercialising research.

Public benefits

- Input for future research.
- Motivating new research.
- Catalysing new companies and high skills employment.
- Transparency in research funding.

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

Funding models for open access 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: 5).

The key challenge for open access repositories 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 datasets, while ensuring that the repository is free to access and maintains the trust of its users. In other words, to find a way to deliver core services with no or limited for-fee income. This has been the challenge presented to the Digital Repository of Ireland: after an initial four year period of core funding to identify, put in place, and transfer to a new funding model; to find a way to continue the work presently undertaken by 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). To that end, we have actively been researching how other repositories have sought to fund their endeavours. Our research has identified 14 archetype funding sources, which can be divided into six classes (Table 2). We have evaluated the relative merits of each source in order to construct a new blended funded model, taking into account certain constraints, and produced a business plan and started the work of lobbying relevant organisations to realise this model. The latter is important, because even if a workable model is identified it does not logically follow that there will be relevant buy-in or a flow of resources/income. The model has to be realised in practice, which involves politics and business acumen. In the rest of this section we discuss each of our potential 14 funding streams in turn. This is followed by a discussion of the challenges and risks associated with implementation.

“The key challenge for open access repositories 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 datasets, while ensuring that the repository is free to access and maintains the trust of its users.”

Table 2: Models of funding open repositories

		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	Build a consortium that 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	Public/private partnerships, with the public sector providing the data and private companies providing 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.

Table 2: Models of funding open repositories contd.

		Model	Description
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	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	Offers end-users a graded set of options, including a free-of-charge option that includes basic elements (e.g., limited features or sampled dataset), with more advanced, valuing adding options (e.g., special formats, additional functionality, tools) being charged a fee. Opens up the product/ service to a wider, low-end market and more causal use, whilst retaining paid, high-end product/service for more specialised users.
	I	Content licensing	Make the data free for non-commercial re-use, but charge for-profit re-users.
	J	Infrastructural razor & 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-subsidizing the former.
Service	K	Pay per purpose	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-subsidization).

Table 2: Models of funding open repositories contd.

Service	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	Offers end-users data products/services for free, with the infrastructure maintained on a voluntary basis, including crowdsourcing.Assembled from Ferro and Osella (2013, 2014); Maron (2014); consultation with stakeholders and team discussion

Assembled from Ferro and Osella (2013, 2014); 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 recouped in full or in part through cost-recovery charging. For example, mapping agencies often operate as trading 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 the data. Nascent research 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 are 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/open access 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 through public entities (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 leveraging diverse consumer surplus value 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 (European Commission 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 evermore of them and to make them open access 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 open access 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. Indeed, the national open access data repositories we contacted[4] typically receive approximately 70% of their funding directly from the state, making up the difference through other funding mechanisms. Repositories that are not national in status are less likely to secure significant state subvention and are therefore more likely to be under pressure to identify and source other funding streams.

(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 to gain access to a sophisticated shared resource and its tools that deliver more 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 (DPC). 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 to ameliorate budget cuts through the sharing of costs for a key service.

(C) Built-in costs at source

Many funding agencies now expect the data from the projects they fund to be deposited in an open access 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 open access 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[5].

[4] Netherlands' Data Archiving and Networked Services, Netherlands Institute for Sound and Vision, UK Data Archive, Swedish National Data Service.

[5] http://archaeologydataservice.ac.uk/, last accessed 15 April 2015.

(D) Public/private partnership

Public private partnerships (PPPs) have been used extensively by the 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. Datasets 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 research 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. Firstly, it is usually best sourced with respect to specific sub-projects rather than core activities. Secondly, 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 becoming 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 then 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 that 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 open access to data and is therefore of limited utility to open access repositories.

(H) Freemium product/service

Some new data infrastructures, such as Dublinked[6], have been experimenting with freemium product/services. All users are offered a free-of-charge set of options that include basic functionality and key datasets. 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 open access. 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 datasets to users, with many research datasets having intrinsic rather monetary than value.

[6] http://www.dublinked.ie/, last accessed 15 April 2015.

(I) Content licensing

Depending on the content, a potential source of funding is content licensing. Here, content such as art images, manuscript screen shots, 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 Euro (Maron 2014). To be able to content license 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 open access.

(J) 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 open access.

Source income through services

(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 website 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 website Inspiring Ireland[7], where the front page and the look and feel of that site is independent of the DRI site. In this sense Inspiring Ireland 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 bear 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 open access 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).

[7] http://www.inspiring-ireland.ie/, last accessed 15 April 2015.

Challenges in funding open access 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 towards the notion of open access, 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, Ireland has just suffered a severe economic recession and an ongoing period of austerity that has led to major cutbacks in public finances (including all the major stakeholders of the repository), cutbacks to research funding, and a prioritisation of remaining funds towards industry-focused research and job creation. Moreover, the competition to secure such funds has increased dramatically as agencies seek to replace lost core funding with soft monies. Raising funding in such a context is a major challenge.

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: 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;
- 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 open access 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 open access can be a vital part of strategies designed to compel researchers to deposit data. Without such strategies it is likely that the move open access 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, 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. As 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 can 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 has 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 took 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 license, the latter a LGPL license. 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 have 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 given that 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 the repository's 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*[8] developed by the Royal Danish Library and the Danish National Archives.
- *Cost Modelling for Sustainable Services*[9] by California Digital Library/Technology at Berkeley.
- *Digital Preservation for Libraries* (DP4Lib)[10] developed by the Deutsche Nationalbibliothek.
- *Keeping Research Data Safe*[11] project led by Charles Beagrie Ltd with funding from JISC.
- *4C: Collaboration to Clarify the Costs of Curation*[12] EU funded project launched February 2013.
- *Lifecycle Information For E-Literature* (LIFE)[13] collaborative project undertaken by University College London and the British Library.

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 project report on Cost Models for Digital Repositories[14] 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 open access repository at risk. The most significant risk is that the repository closes because it cannot cover its core costs. Unless its collections can be transferred elsewhere the danger is that significant datasets will be potentially lost, denying access to researchers, students, citizens and companies. Moreover, there will be a loss of human

[8] http://www.costmodelfordigitalpreservation.dk/, last accessed 15 April 2015.

[9] https://wiki.ucop.edu/download/attachments/163610649/TCP-total-cost-of-preservation.pdf, last accessed 15 April 2015.

[10] http://dp4lib.langzeitarchivierung.de, last accessed 15 April 2015.

[11] http://www.beagrie.com/krds.php, last accessed 15 April 2015.

[12] http://www.4cproject.eu/, last accessed 15 April 2015.

[13] http://www.life.ac.uk/tool/, last accessed 15 April 2015.

[14] http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D32_1-

resource expertise, stakeholder networks, technical infrastructures, and the legal and policy frameworks developed, and a network of trust will be seriously impaired. Further, it will foreclose any ability of repositories to leverage existing investment through additional research funding or other funding streams. In other words, all of the benefits of the investment to that date will disappear and also cause major reputational damage to those associated with the repository and its original funders.

Rather than closing altogether it might be the case that 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 foregoing additional development work or the addition of new datasets. 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 and Webb (2012) 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). 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. 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 open access models have been developed with respect to publications. However, the development of finance models for open access 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 open access 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 open access repositories do need to find ways to fund their activities or they place the collections they hold at significant risk, as well as 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 mitigate against cyclical effects across funding streams by seeking income from a number of sources rather than rely 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, pay for specialist services, and white-label development/platform licensing) providing a smaller proportion of income. In our business plan, we have this set up on 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 open access repositories, and the degree of insecurity will increase for more localised repositories. This clearly has to be a source of concern as it places open access repositories at risk. As such, whilst the arguments advocating open access 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.

References

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

Beagrie, N. Lavoie, B. and Wollard, M. (2010) *Keeping Research Data Safe 2*, JISC, London and Bristol. http://www.beagrie.com/jisc.php

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* 63(6): 1059-1078.

Budapest Open Access Initiative (2002) Read the Budapest Open Access Initiative http://www.budapestopenaccessinitiative.org/read (last accessed 24 Oct 2014)

Carr, N.G. (2007) The ignorance of crowds. *Strategy + Business Magazine* 47: 1–5.

Dasish (2012) Roadmap for Preservation and Curation in the Social Sciences and Humanities. http://dasish.eu/publications/projectreports/D4.1_-_Roadmap_for_Preservation_and_Curation_in_the_SSH.pdf/ (last accessed 15th October 2013)

DataRemixed (2013) *Worldwide Open Data Sites*. 8th August. http://dataremixed.com/2013/08/worldwide-open-data-sites (last accessed 4th November 2013)

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)*. http://epsiplatform.eu/sites/default/files/models.pdf (last accessed 11th August 2013)

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

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

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

European Union (2009) *Communication from the European Commission: ICT Infrastructures for E-science: Brussels, 5.3.2009: COM(2009) 108 Final*. Directorate-General for the Information Society and Media, EUR-OP, 2009, Brussels.

EU Directive (2003) Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2003L0098:20130717:EN:HTML

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, 19–20 June 2012, European Commission Headquarters, Brussels. http://www.w3.org/2012/06/pmod/pmod2012_submission_16.pdf

Ferro, E. and Osella, M. (2013) Eight Business Model Archetypes for PSI Re-Use. *Open Data on the Web*, Workshop, 23rd – 24th April 2013, Google Campus, Shoreditch, London. http://www.w3.org/2013/04/odw/odw13_submission_27.pdf (last accessed 13th 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*, JISC, London and Bristol. http://repository.jisc.ac.uk/279/ (last 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. http:*//ands.org.au/resource/houghton-cost-benefit-study.pdf* (last accessed 14th 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* 64: 123–179.

Maron, N. (2014) *A guide to the best revenue models and funding sources for your digital resources*. Ithaka S+C and JISC.

National Principles on Open Access Policy Statement (October 2012). (last accessed 23rd September 2013) http://dri.ie/sites/default/files/files/National%20Principles%20on%20Open%20Access%20Policy%20Statement%20%28FINAL%2023%20Oct%202012%20%29.pdf

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. *IPPR*. http://www.ippr.org/publication/55/1526/the-value-of-the-public-domain *(last accessed 13th August 2013)*

Pollack, R. (2009) The Economics of Public Sector Information: Profit-maximisation, Cost-recovery, Marginal costs and zero costs. *Cambridge Working Papers in Economics* 0920. ttp://www.econ.cam.ac.uk/research/repec/cam/pdf/cwpe0920.pdf *(last accessed 13th 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.

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

RLG and OCLC. (2002) *Trusted Digital Repositories: Attributes and Responsibilities.* Attributes of Trusted Digital Repositories, http://www.oclc.org/research/activities/trustedrep.html (last accessed 18th February 2013)

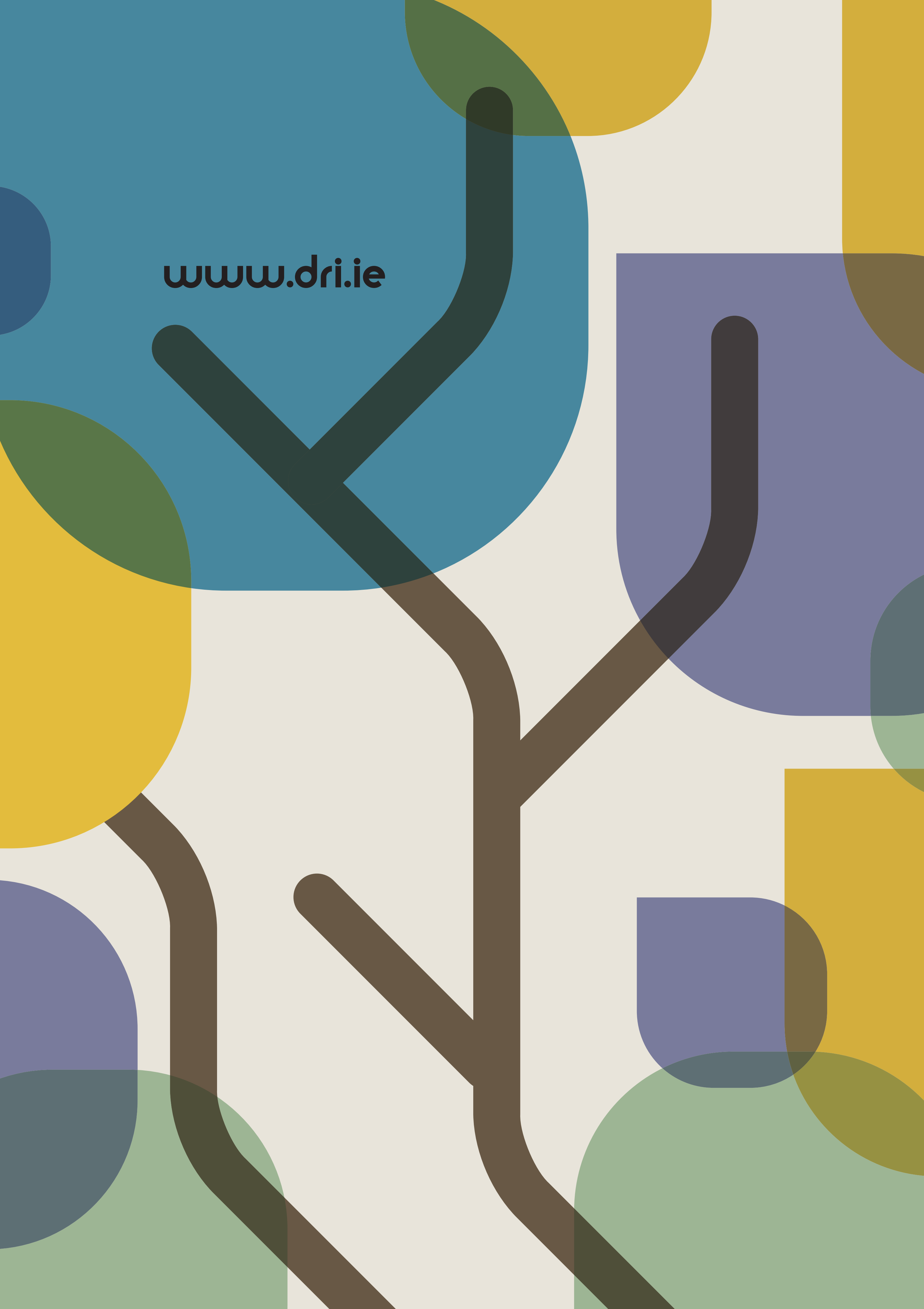
ROARMAP (2014) Registry of Open Access Repositories Mandatory Archiving Policies http://roarmap.eprints.org/ (last accessed 24 October 2014)

Spichtinger, D. (2012) Open Access in Horizon 2020 and the European Research Area http://www.sci-enceeurope.org/uploads/GRC/Open%20Access/2_DanielSpichtinger.pdf (last accessed 24 October 2014)

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

Suber, P. (2013) *Open Access Overview*. http://legacy.earlham.edu/~peters/fos/overview.htm (last accessed 24 October 2014)

White House (2009) Open Government Directive. *Executive Office of the President*. http://www.white-house.gov/sites/default/files/microsites/ogi-directive.pdf (last accessed 19th September 2013)



www.dri.ie