Sustaining the scholarly record and enabling further research, learning and impact

Briefing paper for the Jisc and CNI meeting 2016

Introduction

The nature and scope of the many challenges facing the scholarly community with respect to research outputs are considerable but steps must continue to be taken towards the goal of finding and achieving effective and sustainable solutions. The Jisc and CNI meeting is therefore a timely opportunity to think about the future, providing a forum in which leading experts from the UK and USA can constructively identify key issues and challenges, and look forward with a confident direction of travel.

In preparation for the meeting, the organisers have sought the views of a number of the speakers on four key issues, two on the topic of sustainability of the scholarly record and the supporting infrastructure and two focused on enabling the scholarly record for future research and impact. This short briefing paper encapsulates many of those opinions and is designed to provide “food for thought” in the run up to the event in Oxford. In addition to detailed consideration of the four issues selected for discussion by the meeting organisers, the event is also designed to identify opportunities for greater international collaboration and co-operation.

Sustainability of the scholarly record and the supporting infrastructure

We are witnessing the growth of research outputs that are more ephemeral, inter-connected and dynamic than traditional channels such as regular books and journals and the number of new services in the scholarly space continues to grow apace. The sustainability of scholarly outputs and the infrastructure that supports their creation, distribution and curation is a challenging issue that has received much attention recently, particularly in relation to the Open Access environment. Knowledge Exchange\(^1\), for example, has done a lot of work looking into the sustainability of Open Access infrastructure and content services, resulting in the development of its Sustainability Index\(^2\). More recently, Bilder, Lin and Neylon\(^3\) have proposed a set of Principles for Open Scholarly Infrastructures to help galvanise the discussion about sustainability.

---

\(^1\) http://www.knowledge-exchange.info
\(^2\) http://www.knowledge-exchange.info/event/sustainability-oa-services
Retrieved: 11 26, Jun 14, 2016 (GMT)
Issues and developments

A number of critical issues relevant to sustainability have been proposed and are listed below. Although these will not be in the spotlight at the meeting, they are important and may be useful in terms of your preparation for the event.

- Is there a common understanding of what “infrastructure” means in the context of supporting the creation, distribution and curation of scholarly outputs?
- To what extent do we trust that key elements of the infrastructure can be sustained, individually and collectively?
- How could Bilder et al’s principles apply in different contexts, such as for different types of services or for underpinning different types of research output or activity?
- Are there particular elements of the infrastructure that should be more widely used to help ensure persistence and sustainability and, if so, how might this be achieved?
- What opportunities exist to improve the effectiveness and efficiency of infrastructure?
- Is repository software adequate to ensure the persistence of the scholarly record and if not, how might shortcomings be addressed?
- What are the respective roles of market mechanisms, regulation and public interest activity in ensuring the persistence of the scholarly record?

Many readers will be familiar with some or all of the following sustainability-related developments. In the USA Apereo serves to support the development and maintenance of educational software4 while work on the Digital Preservation Network5 and the Data Conservancy6 project addresses preservation challenges. Relevant developments in the UK and Europe include the European Open Science Cloud7 and Open Science Policy Platform and the Knowledge Exchange initiative on the sustainability of global open access infrastructure services such as SHERPA8, DOAJ9, arXiv10, OpenAIRE11 and CORE12. Specific services such as the Jisc Research Data Shared Service13 are also in scope, as are global initiatives such as the RDA (Research Data Alliance)14, Open Scholarship Initiative15, LOCKSS16, DataCite17 and ORCID18.

4 https://www.apereo.org/content/about
5 http://dpn.org
6 https://dataconservancy.org
7 http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
8 http://www.sherpa.ac.uk
9 https://doaj.org
10 https://arxiv.org
11 https://www.openaire.eu
12 https://core.ac.uk
13 https://www.jisc.ac.uk/rd/projects/research-data-shared-service
14 https://rd-alliance.org
15 http://osinitiative.org/
16 https://www.lockss.org/
17 https://www.datacite.org
18 http://orcid.org
**Key issue one: The persistence of digital monographs**

For monographs and new types of long form publication, how far are we from having infrastructure that reduces the risks to the persistence of the scholarly record? What are the major challenges and opportunities?

It is important to differentiate, at the outset, between regular ebooks and other long form publications and the more complex dynamic monographs that typically contain links to or embedded external content, data, software models, visualisation tools and so on. In the case of ebooks, standards such as PDF/A and EPUB have been widely adopted and the infrastructure used by libraries for electronic or online journals can also be used for electronic books. Although ebooks may be relatively straightforward to manage and preserve, some think that basic pdf-based monographs have limited value since it is felt that scholars ought to be able to do the things they can do with paper books such as make annotations. There exist opportunities for the growing number of new or revived university presses to enrich the ebook publishing universe though some have expressed concern about their business models and sustainability.

There are a number of challenges associated with dynamic long form publications. For a start there is little consensus on what they should look like and even less on underlying standards. Monograph publishers often do not adopt DOIs and metadata can be insufficiently granular to enable the identification of individual components such as images. There are, as may be expected, issues with licensing and sometimes digital rights management; digital monographs may link to source material that is not open, a problem particularly associated with audio-visual material. There is, therefore, a need for systems that accommodate licensed as well as openly available material. As well as challenges there are opportunities to add value and these are seen to be particularly evident when the process of producing a dynamic multimedia book pays close attention to understanding the user experience and involving users in their development.

When it comes to preservation, because each dynamic monograph is somehow unique, so the preservation problem is also unique and cannot easily be scaled to regional, national or international levels. In short, some people take the view that we don't yet have an answer to the question of how to preserve this type of publication at scale.

Other people take that view that the fundamental problem with preserving the scholarly record, whether in monograph or other form, is a social and economic one. It is argued that technical problems are largely tractable but that the lack of collaboration and trust (whether due to concerns over SLAs, governance or security of content) precludes the development of scalable preservation solutions. There are notable exceptions such as the HathiTrust Digital Library but in general institutions tend to look to their own resources to develop or procure their own local solution. Scaling to regional, national or global levels then requires one or more layers of technology to make sense of heterogeneous metadata and content; this is argued to be a suboptimal solution.

**Key issue two: How best to achieve sustainable infrastructure**

Research needs sustainable infrastructure to ensure the evidence on which it is based remains accessible. Sustainability might be achieved through means such as public commitment, market mechanisms or cooperative action. How do the local, regional, national and global actors responsible for providing the infrastructure use these approaches to sustain it? Do these approaches work and what risks are associated with them?
Achieving sustainability for infrastructure is one of the most challenging subjects to be addressed by the meeting and one for which there are no clear answers. The three main approaches set out in the question each have their advantages and disadvantages. Although helpful particularly in the early stages of the development of a project or service, funding from public sources can be short term in nature and can also be subject to the risk of disruption due to political whim or expediency. It has been suggested that governments should put money into endowments to shelter recipients of public funding from such risks. There is also a role for philanthropic pump priming, a source of funding that is especially important in the humanities. Although public funding is seen by some to be problematic in the scholarly communications arena, infrastructure in other sectors – such as transport and power – either is, or has been, successfully run for long periods in the public sector often with strong regulatory involvement even where such infrastructure is in private hands.

While people acknowledge the power of the market to drive innovation and customer service there is also fear and suspicion that the involvement of commercial entities leads to a loss of transparency, trust and community control. In terms of storage infrastructure some say it makes sense to use community expertise to build value-adding services on top of commercial services such as Amazon cloud. It has been suggested that if funders were to make clearer their expectations and red lines that would “de-risk” the investment decision for smaller entrepreneurial companies; at present, with an uncertain investment landscape, larger companies (often content providers) with deeper pockets are more likely to be active in the scholarly information market. It is also suggested that unbundling content and services would enable commercial providers to focus on providing competitive services while the underlying data could be openly available and transparent.

The notion of cooperation is attractive to many but in reality it can be difficult given the propensity of individual institutions to propagate their own solutions. Services such as SHARE and Jisc Monitor therefore need to take an open interoperable approach, trying to connect disparate pieces of a jigsaw. There is a growing desire for collaborative ventures that rise “above the campus” and a recognition that these would work best at the national or supra-national levels.

In reality the answer lies in a mix of these three approaches, though some would prefer the market approach not to be such a dominant force. ORCID is held up as an example of an initiative that did a lot of things right but in such cases early development is often “faith based”. That is, sustainable levels of funding only begin to materialise when a critical mass of players become interested.

It is a fact of life that some databases will go out of business. This is a significant risk for those in the scholarly community who rely on such services. There is a need to identify and manage this risk as databases are migrated to new players and responsibility for their management is re-apportioned. The closure of the Arts and Humanities Data Service in the UK is a relatively recent example of what can happen even to parts of the infrastructure previously believed to be stable and resilient.

19 http://www.share-research.org
20 https://www.jisc.ac.uk/rd/projects/monitoring-open-access-activity
Enabling the scholarly record for future research and impact

Exciting opportunities exist to build on the scholarly record, for example, making new discoveries and creating new knowledge through the use of text and data mining technologies. The scholarly record is being used by a broader group of stakeholders, notably those newly involved in clinical and drug development, new uses like business analytics for science and real advances are being made in analytics-based understanding of research, teaching and learning. Against this background a number of important questions can be posed.

- How might infrastructure be built to enable re-use and new forms of analysis through, for example, data aggregations of different types?
- What types of openness and “walled gardens“ are developing in the scholarly landscape and which of them are especially important?
- How prepared are universities to support researchers curate the outputs from new types of research?
- Is the scholarly record incomplete and, if so, how should this be remedied?
- To what extent are the legitimate ambitions of researchers limited by factors such as intellectual property rights and how might such limitations be mitigated while recognising the rights of creators?
- What roles do or should local, regional, national and global actors play in enabling the use of the scholarly record?
- What opportunities exist for enhancing learning and teaching by building on an open scholarly record?

The meeting organisers have selected two key issues for particular attention but before considering those it is worth reminding ourselves of developments that fall into the category of enabling people to build on the scholarly record. In the USA, the OSTP memo, the launch of the Genomic Data Commons and the policy stance taken by the Gates Foundation demonstrate senior-level support for Open Access. Examples of enabling infrastructure and services include SHARE, CHORUS, the Hathi Trust and the Digital Public Library of America.

In the UK the Concordat on Open Research Data serves to establish core principles of research data sharing and management and Jisc has been advancing specific infrastructure initiatives such as the Research Data Discovery Service, SHERPA Services and CORE. At the global level there are a number of well-known ventures providing identifiers, standards, services and infrastructure including ORCID, CrossRef, DataCite, the NISO altmetrics group, GitHub, OpenAIRE and, more controversially, Sci-Hub. Significant commercial players in this area include Elsevier, Thomson Reuters, Digital Science, ResearchGate and Academia.edu.

**Key issue one: Analytics for research and learning and the challenges of “walled gardens”**

Extensive and reliable data are needed to underpin the responsible use of metrics for research and learning. These data arise as a result of a variety of interactions with the scholarly record by researchers, learners and others. Under what conditions are these data and associated algorithms available to the academic community? If the data and algorithms are held in “walled gardens“ what challenges does that pose and how might they be addressed?

The production, use and even manipulation of research metrics have been the subject of discussion for some time. It has long been fashionable to count things and then add qualitative factors to the mix to provide metrics.
that seem to be irresistible to some but many argue that there is very little evidence that such metrics correlate positively with the quality of research, that the “impact factor” is grossly misapplied and that similar metrics are easy to “game”. Furthermore, such measures are virtually impossible to independently reproduce, they are not transparent and they are increasingly moving into the realm of commercial organisations. The lack of transparency and the potential for bias is what troubles people most. There is a commonly held view that the data from which research and learning metrics are derived should be open, shared and transparent since, it is believed, they are a public good. There is a need for community agreement on what data about the research process must be open.

Efforts to develop alternative metrics have been under way for a number of years and HEFCE’s Metric Tide report together with its proposal for more responsible metrics has led to new initiatives. Jisc and the Open University run CORE, which is a service to aggregate Open Access material, and perform analyses. Recently CORE has been working on the development of “semantometrics” based on analysis of full text as well as citations. There are now a number of “altmetric” tools and services and NISO has recently published a draft version of its Recommended Practice for data metrics. The opportunity to develop research and learning metrics that are transparent, reproducible and trusted by the scholarly community is thought to be compelling.

There are different factors at play when it comes to learning metrics. There is often debate about what outcomes are to be measured and what base they should be evaluated against in the absence of standardised tests at the university level. In the USA, since most funding for Higher Education is at the state level, learning metrics will need to be at the state rather than, or in addition to, the federal level. There is far greater interest in measures of retention and completions of degree courses in the USA leading to the detection of problems, mainly with a social component, and the development of interventions. In the UK Jisc is developing a learning analytics service due to start next year and has been doing a lot of work in this area in recent years. Increasingly data on learners’ activities in higher education are being captured through systems such as learning management systems and e-textbooks leading to better insights into how teaching and learning may be enhanced. There are some issues common to both learning and research metrics including, for instance, data literacy to crate and use the information responsibly, standards to ensure comparable metrics and the challenge of “walled gardens”.

The algorithms and data behind many of the key learning and research metrics exist in “walled gardens”. These are closed environments wherein parts of the scholarly record are created, curated and used, but where access to those parts of the record is restricted by non-academic criteria such as the ability to pay, adoption of proprietary standards or a willingness to give up certain privacy rights. Walled gardens are a cause for huge concern; there are many different types and several examples were cited in the discussions underpinning this document including Research Fish, Research Gate, Figshare and SlideShare in addition to the main commercial citation metrics providers. People also cite institutional repository, CRIS systems, student records and e-learning

21 http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html
22 http://www.researchfish.com
23 https://www.researchgate.net/about
24 https://figshare.com
25 http://www.slideshare.net
systems as being walled gardens bemoaning poor data interchange facilities, poor data schemas, proprietary software and often-poor implementation. The onus is on institutions to check the terms and conditions of commercial services and ensure that it is at least possible to get back when required data given to service providers by institutions. It was noted that even when services are apparently open parts of it may be closed. In response to these concerns and the challenges the community faces, there is an opportunity to develop a shared roadmap. What would a set of principles regarding information an institution should be collecting about the research and learning processes and their outputs, and what third parties should be allowed to do with those data? What common problems exist in the research and learning environments?

**Key issue two: Developing infrastructure to enable compliance with policy mandates**

*What are the major challenges in developing an integrated infrastructure to enable those subject to Open Access and research data policies to comply and demonstrate compliance with those policies? What steps might be taken next in addressing these challenges?*

Institutions in the UK are already familiar with the need to comply with funders’ Open Access policies. In the USA often individual institutions or university systems will have their own Open Access policy or mandate and in recent months the OSTP has released guidelines about Open Access compliance in relation to federal funding. Although it may be too early to get a good sense of the forthcoming compliance challenges in the USA, fears have been expressed that different federal funders have interpreted the requirements differently and so institutions are expecting to have to comply with a range of expectations. For many, the biggest problem is the parochial nature of funders’ requirements that makes it difficult to develop tools that will scale. Ideally there should be a global infrastructure that supports compliance; distributed systems or frameworks are deemed not to work very well for this purpose. The consensus appears to support the rationalisation or standardisation of funder policies to facilitate a smoother path to compliance.

Jisc has been at the forefront of developing tools and standards to help institutions to comply with funders’ policies and demonstrate compliance. The RIOXX Application Profile provides a means for capturing metadata that enables funders to track the research outputs they have funded and identify the license that governs the use of each of those outputs. Jisc Monitor, SHERPA Services and SHERPA REF provide important compliance-related information and the forthcoming Research Data Management Shared Services will provide a means by which UK institutions can comply with funder policies on curating data sets.

In terms of the information institutions need to demonstrate compliance, very often the information they seek is generated by their own institution but they sometimes lack the means to capture and process those data. It has been reported that institutions therefore sometimes ask other parties – publishers or funders – for data for which the authoritative source is the institution itself. Capturing data about grants, the nature and scope of research outputs and so forth is a challenge for many institutions but it may also provide an opportunity to collaboratively create tools or services to meet that challenge.

With special thanks to Geoffrey Bilder, Rachel Bruce, Neil Jacobs, Cliff Lynch, Joan Lippincott, Tim Lance, Cameron Neylon, MacKenzie Smith, Don Waters and John Wilkin.