Introduction
Research data has always been a key factor in scientific research (by ‘scientific’ in this paper, I mean any branch of scholarly activity including humanities and social sciences). Different kinds of data include – non-exhaustively – documents (text, Word), spreadsheets, laboratory notebooks, field notebooks, diaries, questionnaires, transcripts, codebooks, audiotapes, videotapes, photographs, films, protein or genetic sequences, spectra, slides, specimens, samples, software, and models. However, historically, publishers only paid attention to the small fragments of data that were actually included with primary research publications – chiefly the tables and figures used to support particular arguments. The vast majority of material has remained ‘dark’ – firstly in notebooks, now often on desktops or within institutional repositories or on USB sticks. Until recently, funder mandates did not incentivize researchers to curate or preserve the data that their activities produced, so consequently researchers did not spend time on these activities. However, this is now changing profoundly. ‘Data-driven science’ has been earmarked as critical to our abilities to solve various so-called ‘grand challenges’ around managing scarce environmental resources, optimizing new technologies for health and industrial development, and raising the general level of human knowledge. This is reflected by major funding initiatives such as the European Union’s Horizon2020, and the USA’s Office of Science and Technology Policy (OSTP) directives around increasing public access to the results of federally funded research which specifically include research data within their definitions (EU 2013, OSTP 2013).

Whilst funders are increasingly prioritizing research data as valued scholarly output (NSF 2011, University of Sheffield ND), the competitive nature of the research industry still incentivizes investigators to restrict access to data, through the fear of their work being used by others without accreditation or otherwise
improperly, or because of time or other constraints. In other words, the issue of research data and publication is affected by social, as well as technical and political factors. This complicated mix is proving a challenging environment within which to effect material change in behaviours.

Similarly challenging is the publications piece of the overall system: over the last century and more, scholarly publishers have operated a broadly content-licensing business model, managing production and, to a lesser extent, peer-review systems, collecting subscriptions, and charging for access. Initial responses to the increase in ‘data needs’ have resulted in a number of piecemeal solutions: increased capacity to include data within the article, host supplementary information, implement manual cross-links, and so forth. Many of these solutions have arisen on a journal-, editor- or publishing house-basis rather than with either the bigger picture or long-term situation in mind.

However, a combination of technological breakthroughs – including the Internet – together with changing policymaker and funder priorities (NSF 2011, University of Sheffield ND) are coalescing towards a new research communication model. The Royal Society published a key report in 2012, morphing the then-hot debate around open access into the more inclusive ‘open science’. In this they defined ‘open science’ as ‘open data (available, intelligible, assessable and useable data) combined with open access to scientific publications and effective communication of their contents’ (Royal Society 2012: 16).

These developments in the paradigm in turn profoundly affect the needs of the publishers’ audience (authors and readers), pushing the business proposition away from access and licensing towards discoverability and query facilities, visualizations, storage services, and community hubs.

So far there has only been fragmented progress amongst the key stakeholders. Procedures for producing and making available high-quality data (in terms of potential for reuse, interoperability, etc.) need to be simplified and the case for how much it benefits the producer (in terms of professional credit) clarified. There are very real preservation issues – formatting, longevity of institutions, and funding.

The question of who pays has not yet been settled.

In short, in order to remain relevant and valuable within the scholarly communications system players need to understand how the relative relationships between research data and publications are transforming, and to be open to developing new skills, partnerships, and products in order to continue serving the knowledge economy. It is predicted that some existing products and revenue lines – in particular print- and copyright-based – will diminish in prominence, whilst others – research data publishing and management, services linked with discoverability and reusability – will grow.

The publishers’ perspective

Apart from a relatively small number of data centre managers, digital librarians and informatics research specialists, many potential actors within this transforming landscape are themselves relatively unembedded in its recent history, partly because the last decade or so has seen an escalating amount of activity that has been challenging to keep up with. The first real milestone was the 2007 STM Association Brussels Declaration, which stated:

Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars.

Whilst asserting this high-level ideal was clearly a necessary step, there were no guidelines given as to how this might be encouraged, achieved, funded, or policed. Its lack of teeth resulted in few or no material changes in publisher or researcher behaviours and further investigation was required before meaningful action could be taken.

Project ODE (Opportunities for Data Exchange)

A key exemplar was this ground-breaking study which ran between 2010 and 2012 and was funded by the European Union’s Seventh Framework Programme. Its key objectives were to consider the impact that data sharing,
reuse, and preservation is having on scholarly communication and identify incentives for researchers and other stakeholders to help optimize the take-up of future e-infrastructures. The investigation was carried out by a cross-functional group including researchers, librarians, data centre managers, and publishers, and drew together previous studies (such as PARSE.Insight 2010) as well as new evidence from its own questionnaires, interviews, and desk research.

One of ODE’s key achievements was to verify and express a critical disjunct in scholarly research and publication. It found that while researchers value data as first-class research objects, existing systems for data production, registration, storage, accreditation, and reuse opportunities have not supported optimal practice. In fact, there is a pervading sense that there is simply too much data to be managed effectively: ‘[W]e all experience it: a rising tide of information, sweeping across our professions, our families, our globe’ (Wood et al. 2010: 7).

In order to express the situation as clearly as possible, the ODE team adapted Jim Gray’s data pyramid (Hey et al. 2009; and see Figures 1–3).

In addition, as at mid-2012, ODE paints a picture of a few very organized initiatives doing the bulk of the work of building links, promoting reusability, devising standards for metadescription, and building crosslinks. In the meantime, most other institutes, disciplines, and organizations have been rather passive, and the landscape is still fragmented.

However, as both this paper and the special issue as a whole hope to show, since ODE a huge amount of progress has been made.

2012—the present

In June 2012, the STM Association issued another data statement which announced a huge step forward in terms of progress. To begin with, this time it was issued jointly with DataCite, a global organization that co-ordinates the allocation and metadata standards for datasets to be allocated DOIs. Importantly, there were more details and a considerably clearer mission for publishers:

To improve the availability and findability of research data, DataCite and STM...
encourage authors of research papers to deposit researcher validated data in trustworthy and reliable Data Archives.

DataCite and STM encourage Data Archives to enable bidirectional linking between datasets and publications by using established and community endorsed unique persistent identifiers such as database accession codes and DOIs.

DataCite and STM encourage publishers and data archives to make visible or increase visibility of these links from publications to datasets and vice versa.

So, how are publishers responding to the new landscape?

New products

In recent years, a number of data-related titles have been launched. These include Nature’s Scientific Data, Wiley’s Geoscience Data Journal, the F1000Research journal, and BMC’s GigaScience. All of these are running slightly different models in terms of workflow, archiving policies, and peer-review requirements, which illustrates the current lack of consensus around such critical issues and points towards the need for further debate and consolidation. At the same time, they represent genuine attempts to push the publications envelope by supporting the longer-term scientific infrastructure objectives.

New policies

Publishers, often alongside learned societies, are also adapting publication policies in support of the evolving new norms. Increasingly – at publisher, subject, or journal level – authors are being given clear guidelines as to data availability requirements in order for their paper to be accepted (see, for instance, announcements by Nature 2013, PLOS 2013, and the American Geophysical Union 2013). A key example here is the innovative work being done within the neuroscientific community. The Resource Identification Initiative (https://www.force11.org/resource_identification_initiative) consists of a group of researchers, journal editors, publishers and biocurationators who have joined together to endorse and support a system to identify antibodies, model organisms, databases, and software tools in a machine- and human-readable format across publications. Amongst others, Wiley, Elsevier, F1000, and PeerJ have co-operated (see, for instance, Bandrowski et al. 2014) and the hope is that neuroscience as a discipline will eventually be strengthened.

New partnerships

Publishers and learned societies already have a long history of largely symbiotic association. However, this new world order is beginning to open up fresh challenges and opportunities for the two to work together. Societies are potential loci of deep understanding about specific disciplines, their communities’ needs and researcher behaviours, whilst publishers are in the potential position of offering economies of scale in terms of technological support, access to policymakers and adaptation of journal behaviours.

At the same time, new relationships are starting to develop. Ties between publishers, funding bodies, and policymakers are no longer confined to a few key personnel; instead, new dialogue points are beginning to open up across organizations more generally – see for instance the list of speakers for the symposium ‘The Now and Future of Data Publishing’ (http://nfdp13.jiscinvolve.org/wp/programme). Data publishers, data centre managers and digital librarians are also emerging as key contacts and colleagues in forming new relationships and workflows. See, for instance, the PANGAEA data maps in Elsevier publications (http://www.elsevier.com/about/content-innovation/pangaea-data-maps-in-articles). Entities such as DataCite, the Research Data Alliance, and the Belmont Forum (http://www.bfe-inf.org/) are also emerging as increasingly important partners, both in their own right, as well as for their ability to foster meetings and collaborations amongst the stakeholder groups.

The Research Data Alliance/World Data System, Publishing Data Interest and Working Groups deserve particular mention here (see https://rd-alliance.org/groups) With membership made up of a variety of publishers, data centre managers, librarians, and other agencies, these groups are working to make data
publishing a practical ‘business as usual’ reality. Currently four groups are each concentrating on specific issues: workflows, bibliometrics, publishing services (cross-linking), and cost recovery. Regular updates are published, and there are also opportunities to attend webinars and other real-time events to learn more and input to the thinking on these topics.

Conclusions

As increasing progress is made towards understanding and managing the publication of research data, the landscape is becoming more interesting and diverse. Whilst there is greater clarity of message from funders, a large proportion of researchers have not systematically adopted an ‘open science’ mindset, which indicates this message has still not fully permeated all the relevant communities.

At the same time, despite increasing signs from publishers, learned societies, and other key entities working with research data, that they are absorbing and attempting to respond to the changing scholarly communication environment, it is clear that issues such as ‘best practice in data publication’, standards, scalability, and, critically, overarching business models are still largely undecided. The underlying implication of the Royal Society report is that ultimately research communications will look very different from even the online-only, enhanced format papers we are increasingly used to seeing. Yet, while researcher assessment and reward systems are still predicated upon existing scoring systems, and the sheer workload involved with proper data curation, uploading and peer reviewing remains so high in relation to the rewards involved, it is difficult to foresee how this potentially transformative issue is ultimately going to play out.

The upside of this continued uncertainty is that there is currently no infallible monopoly or product that has so far achieved the ultimate ‘global grab’ in this sphere. Those publishers and societies concerned about their achievements and current strategies within the data publishing field still have a critical part to play in the development of the overall landscape. Many of the organizations mentioned in this paper — and indeed the STM Association, ALPSP, SSP, and other publishing bodies — are also keen to co-operate and to build membership and participation in future ventures. The future, whilst unwritten, does potentially contain publishers.

References

American Geophysical Union. 2013. AGU Publications data policy. http://publications.agu.org/author-resource-center/publication-policies/data-policy/ (accessed 2 August 2014).
Bandrowski, A., Tan, S. and Hoé, B.R. 2014. Promoting research resource identification at JCN. Journal of Comparative Neurology, 522: 1707. http://dx.doi.org/10.1002/cne.23585
DataCite and STM Joint Declaration of 14 June 2012, http://www.datacite.org/node/65 (accessed 15 August 2014).
ELI. 2013. Europea Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (accessed 2 August 2014).
Hey, T., Tansley, S., and Tolle, K. (eds) 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Corporation.
NSF. 2011. National Science Foundation Data Management Plan Requirements. http://www.nsf.gov/eng/general/dmp.jsp (accessed 11 August 2014).
Nature. 2013. Availability of data and materials. http://www.nature.com/authors/policies/availability.html (accessed 2 August 2014).
Opportunities for Data Exchange Project: http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf, dated 17 October 2011 (accessed 11 August 2014).
OSTP. 2013. White House Office of Science and Technology Policy (OSTP) memorandum. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed 2 August 2014)
PARSE.Insight. 2010. Permanent Access to the Records of Science in Europe, http://www.parse-insight.eu/ (accessed 11 August 2014).
PLOS. 2013. Data Access for the Open Access Literature: PLOS’S Data Policy. http://www.plos.org/data-access-for-the-open-access-literature-plos-data-policy/ (accessed 2 August 2014).
Royal Society. 2012. Science as an Open Enterprise. The Royal Society Science Policy Centre Report 02/12. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
STM Association Brussels Declaration. 2007 http://www.stm-assoc.org/brussels-declaration/ (accessed 15 August 2014)
Wood, J., et al. 2010. Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data, Final Report of the High Level Expert Group on Scientific Data, A Submission to the European Commission, October 2010, http://cordis.europa.eu/ipf7/ict/e-infra-structure/docs/hlg-sdi-report.pdf (accessed 11 August 2014).
University of Sheffield. ND. Overview of RCUK definitions of a data management plan. http://www.sheffield.ac.uk/psychology/research/groups/dmgspsych/rcukdefs (accessed 11 August 2014).

Fiona Murphy
Publisher
Earth & Environmental Sciences
Global Research, Wiley
The Atrium, Southern Gate
Chichester
PO19 8SQ, UK
Email: f.murphy@wiley.com