THE BENEFITS OF LONG-TERM STEWARDSHIP OF RESEARCH DATA

Dr Matthew Addis is Co-founder and CTO of Arkivum. Matthew previously worked at the University of Southampton IT Innovation Centre. Over the last fifteen years, Matthew has worked with a wide range of organisations in the UK, Europe and US on solving the challenges of long-term data retention and access.

There is substantial value in making research data open and accessible [1]. Benefits include: science that is higher quality and more productive; faster development of new products and services; and increased impact for research when addressing societal challenges. Substantial benefits can be derived from making other forms of data open too: for example, the Open Data Institute reports [2] on the potential held by data from the public sector. From an economic standpoint, the value of making data open is estimated to be as high as 4 per cent of GDP [3]. Or, as a McKinsey report [4] put it, "Open data – public information and shared data from private sources – can help create $3 trillion a year of value in seven areas of the global economy."

Nowhere are these benefits more apparent than in the response to the current COVID-19 pandemic and in the efforts of the international community to rapidly share data in an open and trusted way. The Research Data Alliance COVID-19 working group is at the forefront of establishing guidelines to ensure that open data on COVID-19 engenders the maximum benefit both today and in the future too.

THE NEED TO PLAY FAIR

These benefits can only be fully realised if research data is Findable, Accessible, Interoperable and Reusable – otherwise known as FAIR [5]. It is not enough simply to put data online and hope for the best. FAIR encourages and supports high-quality research that follows good research practice which produces results that are repeatable, verifiable and re-usable. Only under these circumstances can research data be used with confidence and exploited to its full potential. These drivers for FAIR data are embodied at an international level in statements made by the G8 science ministers in 2013 and last year in the Beijing Declaration from the Committee on Data of the International Science Council (CODATA).

There are substantial costs if good practice is not followed and data is not made available in a FAIR way. A recent PwC report containing a cost-benefit analysis of FAIR data stated that: *"We estimate the annual cost of not having FAIR data to be a minimum of €10.2bn per year. The actual cost is likely to be much higher due to unquantifiable elements such as the value of improved research quality and other indirect positive spill-over effects of FAIR research data."*

Crucially, FAIR data also needs to be made available in a trusted and reliable way for the long term, often many decades – only then will the value of open access to be fully realised. For example, a study [6] shows that academic and industrial innovators cite biological data resources in their patents decades after the data was originally published.

IT'S A MATTER OF TRUST

The need for long-term trustworthy research data is embodied in the TRUST principles, namely that stewards of research data should consistently consider Transparency, Responsibility, User Focus, Sustainability and Technology. Or, as the TRUST article in Nature [7] puts it: *"to make data FAIR whilst preserving them over time requires trustworthy digital repositories (TDRs) with sustainable governance and organizational frameworks, reliable infrastructure, and comprehensive policies supporting community-agreed practices"*. The article goes on to point out that *"Consensus on 'good' data management practice is beginning to form, but there is still insufficient implementation in some scientific domains."* This is where there is much work still to be done – work to ensure not only that research data is made available today, but that it is also properly managed and stewarded and made available for those who can, should and will benefit from it in the future.

THE UK LEADS THE WORLD

The long-term stewardship of research data is an area in which the UK is well placed when it comes to cementing and extending its current position as a world leader.

The European Bioinformatics Institute at Hinxton in Cambridge

presents an excellent example. For decades, the EBI has been providing access to a wide range of life-sciences data. An independent report [8] found that "EMBL-EBI services contributed to the wider realisation of future research impacts conservatively estimated to be worth some £920 million annually, or £6.9 billion over 30 years in net present value." As the report notes, this is equivalent to 20 times the operational cost of running the EBI. Furthermore, the report findings note that "45% of survey respondents stated that they could neither have created/collected the last data they used themselves, nor obtained it elsewhere."

UK institutions such as the EBI already lead the field in providing long-term open access to research data, but the UK is also home to world leaders in important and related fields. For example:

- The Digital Curation Centre (DCC), the Digital Preservation Coalition (DPC) and the Open Preservation Foundation (OPF) provide expertise in digital curation and digital preservation.
- Memory institutions such as the British Library and The National Archives, and of course the Parliamentary Archives, continue to break new ground in digital preservation put into practice.
- The Jisc Open Research Hub is a new innovation in hosted and service-oriented solutions for research data management.
- Arkivum's digital preservation and data archiving solution is an example of new commercial services for long-term data management.

Together this means that the UK punches well above its weight and is ideally placed to create and deliver new solutions and new commercial services for the long-term stewardship of research data – and to offer these solutions on the international stage.

EOSC: RESEARCH DATA AT AN UNPRECEDENTED SCALE

A major multi-national initiative is the European Open Science Cloud (EOSC) [9], which exemplifies the scale of the challenge and the opportunity. EOSC seeks to store, share and re-use research data across European borders and scientific disciplines, and to provide access to an array of related services, bringing together institutional, national and European initiatives and developing a shared pool of scientific knowledge underpinned by FAIR data. To put that in context, EOSC targets 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences. The scale is huge – and so are the volumes of data being produced.

Whilst EOSC has embraced FAIR, challenges still need to be addressed when it comes to ensuring that the initiative's FAIR data is properly stewarded and managed for the long term. Even the largest organisations in EOSC, such as CERN and the EBI, recognise that stewarding data on the scale required by EOSC needs new approaches and solutions.

ARCHIVER

ARCHIVER [10] is a new €4.8m European Commission-supported project, led by CERN, which will start addressing this challenge. ARCHIVER recognises that commercial services for digital preservation now need to be reliably and certifiably scaled to the "petabyte region and beyond" in order to address the specific complex data requirements of many scientific disciplines. The ARCHIVER project aims to introduce radical improvements in the area of archiving and digital preservation services by combining multiple ICT technologies – including extreme data-scaling, network connectivity, service inter-operability and business models – in a hybrid cloud environment. Its aim is to deliver end-to-end archival and preservation services that cover the full research lifecycle.

Arkivum, in partnership with Google Cloud, has been chosen as one of five consortia for the design phase of the three-year ARCHIVER project, which launched in June. Spun out from the University of Southampton nearly a decade ago, Arkivum now provides specialist software and services for long-term data management and digital preservation to major institutions and commercial organisations in a diversity of sectors, including life sciences and pharmaceuticals, research and higher education, and culture and heritage. Arkivum and Google together will be tackling the challenges that the ARCHIVER buyer group (which in addition to CERN and EMBL-EBI includes DESY in Germany and PIC in Spain) has laid down. The end goal is providing new services for long-term data archiving and digital preservation to the whole EOSC community.

The vast volumes of data that are now being produced around the world, in so many areas of endeavour, hold major long-term opportunities for producing substantial economic, scientific and societal benefits. ARCHIVER is just one example of the opportunity for the UK to consolidate and extend its position as a world leader in the long-term stewardship of research data – all the way from new governance and policy-making through to the development and commercialisation of innovative new services and infrastructures.

References

1 https://ec.europa.eu/info/sites/info/files/research_and_innovation/funding/documents/ec_rtd_he-partnership-open-science-cloud-eosc.pdf

2 https://theodi.org/article/ research-the-economic-value-of-open-versus-paid-data/

3 https://www.huffingtonpost.co.uk/jeni-tennison/economic-impact-of-open-data_b_8434234.html?guccounter=1

4 https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

5 https://www.go-fair.org/fair-principles/

6 https://pubmed.ncbi.nlm.nih.gov/27092246/

7 https://www.nature.com/articles/s41597-020-0486-7

8 https://beagrie.com/static/resource/EBI-impact-summary.pdf

9 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

10 https://www.archiver-project.eu/ □