blog

William Kilbride, Director Digital Preservation Coalition and
Ingrid Dillo, Deputy Director DANS and project coordinator FAIRsFAIR
31 Jan 2022 .

Why long-term preservation is important in relation to FAIR

Ingrid

In April 2019, so right in the first month of our FAIRsFAIR project, one of the Champions of FAIRsFAIR, Barbara Sierman, wrote a message to what she called in her blog “the FAIR-ists”. She stated that there was no need to freshly start wondering how to cope with keeping data FAIR over the years, because this is daily work for digital preservationists. We all would benefit from a better interaction between FAIR and digital preservation. She ended her blog by saying that it is high time that the worlds of FAIR and digital preservation embrace each other, to avoid costly reinventions of wheels. At that time, early 2019, keeping data FAIR was not yet a topic that was widely debated.

In FAIRsFAIR we have put quite some effort in supporting repositories, both with the improvement of their trustworthiness as well as in helping them to enable FAIR data. We also produced a working paper on this topic called FAIR plus Time, together with the SSHOC and

EOSC Nordic projects. By now I think we acknowledge that both FAIR and long-term preservation are needed if we want to keep the data in the EOSC FAIR over time and not only at the point of delivery. Now, acknowledgement of the importance of the topic is of course an important first step. But now we need to bring this into action and put time, focus and effort into really enabling researchers, repositories and all other stakeholders involved to make and KEEP data FAIR.

What is your take on the current situation in the research data sector?

William

I have two answers to this: it is probably better than we sometimes fear; but not as good as it needs to be. I look on from the fringes and I can see the research data landscape has a lot of insight into the digital preservation challenge. Data is valued and actors at all levels are aware of the issue, even if far too many still confuse preservation with backup. That puts the research data sector way ahead of other sectors of the economy: the creative industries, finance, engineering, healthcare, public administration, justice, transport – are all miles behind. Our own work in the decommissioning of nuclear power stations, for example, means we have been trying to package and translate the best thinking from research into a sector which has an awful lot of data and a concern for the (very) long term.

At the same time, you could be fooled into thinking research science has solved digital preservation. Frankly too many senior managers still confuse preservation with backup. The Fair Forever study showed that responsibilities and accountabilities are not properly aligned. Data is 'born fragile' and there is a risk that repositories, at the end of a complicated supply chain line, are being set up to fail.

There are opportunities to learn from other industries. I am particularly thinking of aviation where there are long and complicated supply chains in the construction of products that have a long maintenance lifecycle. The value of product information is widely understood as is the professional obligation to maintain it, share it, and pass it on. It is also increasingly true in the construction sector where building information management has emerged as a specialism. These are only two examples and there are others: but it is telling that in neither of these examples are memory institutions in the regular sense, nor even are they expecting to maintain data forever. They are just doing what they need to do to ensure the success of their product over the long term and meeting their professional obligations. That's very close to what FAIRsFAIR is also hoping to achieve.

So, a key theme for me is how to share knowledge and expertise gained in projects like FAIRsFAIR with other sectors; and perhaps also how to learn from them about the data continuum to fill the gaps that have been described. There is a risk research data management has become detached from the wider economy and that the wider economic impact you could have, is diminished.

Long-term data preservation involves multiple stakeholders

Ingrid

Let us dive a little deeper into this topic of keeping data FAIR. In the Figshare State of Open Data report 2021 one of the main findings is that repositories have a key role to play in helping make data openly available and FAIR. In their survey thousands of researchers around the globe were questioned and over one third of them say that they rely upon repositories to help them making their data FAIR and openly available. This finding confirms that the work we have undertaken in FAIRsFAIR to support and strengthen FAIR-enabling and trustworthy digital repositories was a right choice to make. But is also clear that they are not the only ones. The State of Open Data report also mentions publishers and institutional libraries as central players.

This means there is a shared responsibility to enable the researchers to comply with open and FAIR data policies. But I also know that this shared responsibility between those who provide assistance to researchers is not yet widely acknowledged. Which in its turn again leads to a corresponding lack of coordination between these enablers. Repositories obviously have a crucial role to play, but your report 'Fair Forever' also identifies many other stakeholders with different responsibilities throughout the whole research data life cycle.

Could you provide us some insight into these findings?

William

The 'Fair Forever' study examined the state of the art in digital preservation in EOSC. We reported just as the EOSC Association was being formed. We found some really great things in the EOSC landscape – which is very big and complex. The obvious highlights are the data repositories which have commitments to continuous quality improvement codified into their work towards accreditation, like DANS. There are other strengths too, such as the use of persistent-identifiers or data management plans, which are well understood, well-developed and widely used.

Even so, the data is immensely complicated. It is not always clear who is responsible for ensuring its integrity or usability through some very long supply chains. So, even when the skills exist – and they don't always – and even when the policies exist within institutions – and they do not always, and even when the funding exists – it almost never does - the supply chains from creation to repository invite conflicting responsibilities. No one is accountable if data is lost. Repositories expect to be audited but no one expects data management plan to be audited. That probably needs to be addressed. It is also why I was really pleased to hear earlier today from others in the FAIRsFAIR project about the role of researchers, the role of data in research assessment frameworks as well as the reform of higher education

Put that in context. We publish the annual 'Global List of Digitally Endangered Species', the BitList. If you measure research data against BitList criteria you get a stark answer: data which is structurally complex and where there is uncertainty about responsibilities is termed 'Critically Endangered'. That's the second highest alert. So, despite all the progress that is been made, parts of the open science estate are one click away from extinction.

Thinking of future challenges

Ingrid

When we think about the main challenges for the future and we acknowledge that many stakeholders need to fulfill different roles, it is clear that the cost element is an important one. Quite some work has already been done in this area. I am thinking of the 4C project many years ago where we focused on clarifying the costs of curation and where we also tried to make the point that digital preservation is not only complex and costly, but also realizes a benefit. I am also thinking of the work done within the context of OECD on business models for sustainable repositories.

That we are still struggling with the cost aspect, is clear from the fact that this topic is now also highlighted within the EOSC Association and will be one of the strands of work of the newly created EOSC Association Task Force on Long Term Data Preservation.

Which other challenges for the future come to your mind?

William

There is a lot to pick up here. Yes, cost is an issue which we must address as are accountabilities. Anyone listening, preservation is not backup and you need to reflect deeply and carefully if you think storage is enough. In too many cases preservation is an unfunded mandate. We spend fortunes on other things without thinking twice. I also want to think about carbon costs. I do not think we are as far down this line as we ought to be. We have worked together on cost models in the past, especially the 4C project – Collaboration to Clarify the Costs of Curation. I wish I could assemble the team again around a 5C project which is '4C plus Carbon'.

Carbon and the wider debate about environmental sustainability is inherently complicated. There is also a lot of misdirection and, quite frankly some bad actors too. So, it would take a real effort to understand the carbon costs, and an even greater one to communicate the findings and act upon them.

What we need for the coming years in order to keep data in EOSC FAIR

Ingrid

If we agree that repositories have a key role to play in keeping data within the EOSC FAIR, I believe it is important that these repositories collaborate. Seeds of this collaboration have already been sown through the repository support programmes of FAIRsFAIR, SSHOC and EOSC Nordic. Also, through other national, regional initiatives and of course through CoreTrustSeal.

I am a strong advocate of the creation, based on these initial repository communities, of a European network of trustworthy digital repositories. Such a collaborative network could unite and strengthen the voice of the repositories in the EOSC policy debates. It could provide support and training and an opportunity to learn from one another, for the repositories. And

there is also a technical component. In an ideal world it should be possible for researchers to push through the data they want to preserve for the long term from the EOSC services to a safety net of repositories, with more or less a single click on their dashboard. This is something we are working on e.g. in the EC funded DICE project.

What other vital elements do we need to safeguard our FAIR data in Europe?

William

I strongly agree about collaboration between repositories being vital, and maybe I can make a brief pitch for membership of the DPC as a vehicle for collaboration. I have already mentioned costs and carbon as emerging issues. I would also like to add a thought about the nature of data, and it reminds me to mention change.

We still seem to make assumptions about data as a form in its own right, distinct from hardware and software. I understand why; but it can be hard to distinguish data from application. That is even harder in the cloud where the long chains of interdependence result in small, subtle and often unannounced changes.

Reproducibility is the emerging challenge.

This is the 20th anniversary of the DPC. If you look at our early plans we were supposed to be around for a couple of years, solve some problems and return to our real jobs. But change is constant. If we want to make a success of preservation, we have to recognize this: preservation is not a project or an app; it is a commitment. It needs continuing dialogue and insight into our changing needs. That’s where collaboration can be most effective.

This commitment can only be taken on if we learn from each other and work together: when FAIR-ists and preservationists unite!

3,142 Read

<?php// print render($content['links']); ?>