JISC OPEN RESEARCH HUB

Integrate, Deposit, Discover and Preserve

Paul Stokes
Jisc
UK
Paul.stokes@jisc.ac.uk
0000-0002-7333-4998

John Kaye
Jisc
UK
John.kaye@jisc.ac.uk
0000-0002-4400-4252

Dom Fripp
Jisc
UK
dom.fripp@jisc.ac.uk
0000-0001-5352-4666

Tamsin Burland
Jisc
UK
Tamsin.burland@jisc.ac.uk
0000-0002-5129-979X

Tom Davey
Jisc
UK
tom.davey@jisc.ac.uk
0000-0002-0319-5574

Matt McGowen
Jisc
UK
matt.mcgowan@jisc.ac.uk

Abstract – Jisc's Open Research Hub (JORH) integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for researchers and research managers. The service offers both open source and proprietary systems and allows data and metadata to be shared openly if required. The platform has been developed through years-long consultation with the UK HE research sector and sector bodies, along with contributions from both in-house Jisc and third-party experts.

The need for such a solution has arisen from the sector's desires to achieve several, shared aims, including: greater collaboration; tackling the reproducibility crisis; enabling better research; and meeting funder requirements.

Jisc's custom-built repository—the Jisc Research Repository—is part of the Jisc Open Research Hub. It's built upon an extensive data model and rich messaging layer, providing users with a clean, simple, and easy-to-learn interface for the deposit, approval, and discovery of a range of outputs. In particular it allows for a seamless end to end experience for the user; from deposit straight through to preservation.

Jisc's position in the UK higher education / research sector, as well as the scale of the service provides us with many domain-specific insights to share with iPRES delegates, ranging from the broad methods mentioned above, down to individual design decisions informed by our research and domain expertise.

Keywords – Research Data, Integration, Preservation, Shared Services, Repository to Preservation

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation;

Designing and Delivering Sustainable Digital Preservation.

I. INTRODUCTION

Jisc's Open Research Hub, integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for all users of repository and preservation systems. The service offers a range of systems—both open source and proprietary—and a range of potential integrations, both in terms of data (and metadata) sources and endpoints. It allows data and metadata to be managed, preserved and shared as openly as possible and as securely as needed.

This paper explores the design and development philosophy of JORH and presents the features that make it well placed to fulfil many different types of repository and preservation use cases.

This paper will be of interest to content generators, developers, integrators, vendors, repository managers, curators, research data managers, support staff, and data end-users.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

II. Origins and Drivers

Jisc has been involved in development of research data management projects for over a decade. Between 2009-11 Jisc funded work that developed a number of important tools, technologies and services within universities. Programmes that run between 2011-13 built on earlier work and also supported the development of skills and expertise for specific disciplines and support staff. After this, the next step was to broaden the implementation of previous work

This led to the Research at Risk programme (2014-2016) led by Jisc in partnership with RLUK, RUGIT, SCONUL and UCISA, and informed by numerous stakeholder consultation events. The stated aim of the programme was *"To provide infrastructure, advice and tools to support universities in establishing good data management practice as a core part of their research function."* An additional aim was to take the lead in developing a sector-owned direction of travel for Research Data Management (RDM) and develop a community governed infrastructure.

All this work led to the Jisc Research Data Shared Service project—the output of which transitioned to service last year.

A. Sector sanctioned

In 2015 a report [1] by Jisc and a number of other sector bodies including SCONUL, RLUK and UCISA looking at RDM and universities identified 5 key areas for action in research data management

1. Policy development and implementation
2. Skills and capabilities
3. Infrastructure and interoperability
4. Incentives for researchers and support stakeholders
5. Business case and sustainability

Identified within the Infrastructure and interoperability actions strand the following themes were identified *"…there is clear demand for national shared services for research data management… …the potential economies of scale are attractive to the community… …there is also demand for a national approach to data perseveration…"*

B. Drivers

Research active HEIs now operate in an environment which requires them to address issues around research data management and 'open research'. There are also mandates from publishers and research funders regarding making data findable and accessible to others, and preserving the data for a given length of time.

In addition, the recent changes in the laws around storing personal data means that universities need to take more interest in how researchers manage datasets containing personal information.

There are also increasing concerns around research integrity and reproducibility, calls for greater access to the original research data.

Finally, there is an increasing recognition of the academic and societal benefits of open research, and initiatives to openly share both the findings of the research and the data these were based on.

Given all these issues, it's increasingly important that universities, carefully manage, store, share (if appropriate) and preserve their digital research outputs.

Jisc was tasked with building a service primarily for Institutions to address the emerging requirements and challenges of RDM

It's a big challenge for institutions to take on individually, especially small institutions without much expertise in the area of research data and digital preservation. Hence the concept of a shared service.

Given their position in the sector, and previous work we had been involved in, Jisc was perfectly placed to lead such a project.

III. Development process

A. Shared service and co-design

The platform has been developed through years-long consultation with the UK HE research sector (over 70 universities took part in the consultation) and sector bodies, along with contributions from both in-house Jisc and third-party experts. The functional requirements fell out broadly into 4 categories: Capture & reuse; Preservation; Reporting; Advice and best practice.

The requirements collected posed a challenge for building the shared service. To meet these requirements ultimately we needed:

- A multi-tenant system—which enabled it to be an affordable and scalable cost effective solution for the sector.
- Multi content types—institutions wanted to be able to store all the outputs of research—data, articles, theses, software and method—in one place.
- A flexible system—one where institutions used only the parts of the system they needed and which allowed for interoperability with other systems.
- A good user experience—one that eased the pain of preservation and reduced the need for expertise in digital archiving.
- Reporting
- Most importantly community governance, a national service build for the UK sector, by the UK sector.

To build a service which could meet these challenges, we mapped out the work flow shown in Figure 1 - Service workflow summary.

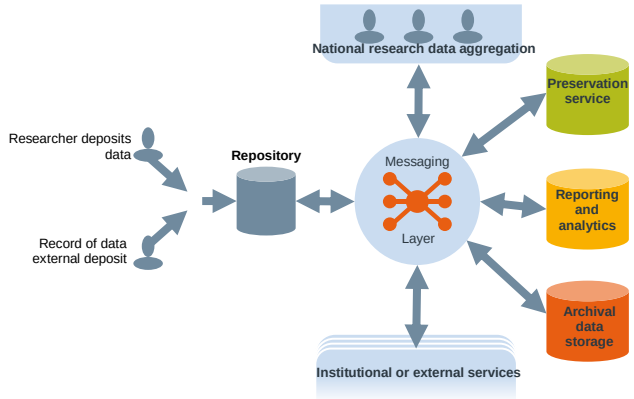

Figure 1 - Service workflow summary

To achieve this workflow and interoperability, we developed a publish-subscribe messaging layer, based on an open, canonical data model and open APIs.

Our alpha Minimum Viable Product (MVP) demonstrated automatic ingest of data files into the preservation system from two different repositories (one open source and one commercial), into two different preservation systems (again one open source and one commercial).

B. Agile development and the current Open Research Hub

The project has been developed using agile techniques throughout. It has also been extensively tested with a group of Pilot institutions through a co-design process. Inevitably this has meant an evolution from the original proposed architecture devised to fulfil the initial set of requirements. In the early stages of the project many more than the current two repositories and two preservation components were incorporated into the overall system. However, in order to achieve a working MPV is as fast a time frame as possible, these were winnowed down to the current two of each. On the other hand, other systems can still be used with JORH through the use of connectors and the open APIs.

Figure 2 - System architecture shows the current version of the Open research Hub

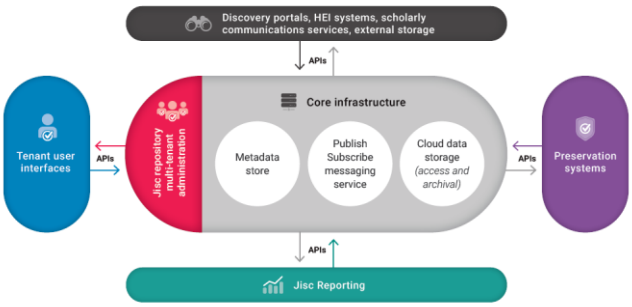

Figure 2 - System architecture

The core infrastructure is where the messaging layer sits, along with the metadata store and data storage.

Initially we were developing an existing open source repository to sit alongside this, but we couldn't find a product which match all our requirements. It became clear that we could get the functionality required from the Core infrastructure, and just needed to build a light weight front end for our tenants and our administration purposes. These front ends (the admin interface and the tenant interface) make up the Jisc Research Repository, a multi-tenant, multi-content, interoperable repository.

The infrastructure interacts with the two other components of the system—Preservation and Reporting—via open APIs

Interoperability with other systems, such as other repositories, Current research Information Systems (CRIS), scholarly comms services, external

storage, and discovery portals is also facilitated by adaptors based on the open API specification.

The service as a whole forms the Jisc Open Research Hub. It allows users to deposit data quickly and easily (with some of the required metadata being pulled in from other institutional systems), and have that data automatically preserved, reported upon and exposed to discovery portals.

C. What are the next challenges in research and are we prepared for them?

Jisc built JORH around a national need for affordable, accessible and compliant research data management infrastructure in higher education. Now the platform exists, how can it further address the needs of the wider UK research sector and the international open science agenda?

JORH has the capacity to support and apply metadata standards for publishing various stages of research in a standardized and robust fashion. This could include anything from publishing a hypothesis, to methods, to null and significant results, bringing visibility to the vast array of findings that are never published. JORH can bring interconnectivity to these component pieces of research within by integrating with tools from across specific research domains and catalyse new habits in research that supports openness, and verifiable science.

While the Hub was built to address the needs of research data curation, its adoption of open, best practice standards means it has the potential to allow the service to handle a much wider range of digital research objects, including Open Access articles, theses and software. The data model, rich messaging layer and an open API facilitate interoperability with other institutional and scholarly communications systems. This provides the potential for the Hub to underpin infrastructure capable of meeting the requirements of an ever-evolving open research agenda.

Artificial intelligence enables great opportunities for the automation of discovery, computation and analytics for data driven research at scale. Yet training effective algorithms takes effective training data. JORH can help by providing a place for well described and verified datasets to be made openly available. This is a first but important step to a more transparent and interpretable version of AI.

IV. KEY DESIGN FEATURES

A. Open and extensible

JORH is a cloud-based, community governed, multi-tenant solution for universities and other research institutions to manage, store, preserve and share their published research data. Based on existing open standards, the service's open and extensive data model incorporates best practice from across the sector, including DataCite, CrossRef, CERIF, Dublin Core and PREMIS.

B. User Experience

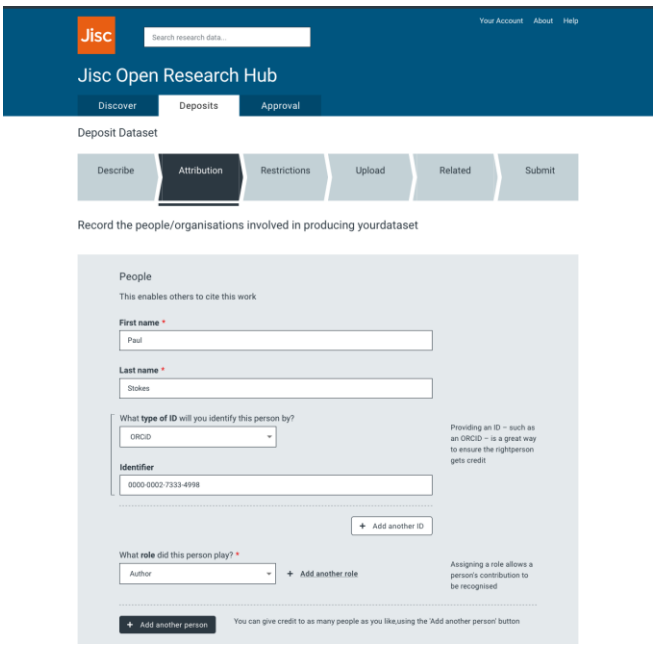

Figure 3 - User interface example - capturing attribution metadata

Jisc recognised the importance of delivering this service with a compelling user experience, and invested greatly into achieving this aim.

A key challenge for the viability and effectiveness any digital solution is achieving engagement from users. This is particularly true in the research arena, where people are required to engage with a number of complex systems in order to achieve their goals, as well as the aims of their institution, funders, and other stakeholders.

To this end we engaged a user-centered-design approach. Beginning by researching to develop a rich understanding of the people who will use the system (their motivations, their pain points, their skills and knowledge) we were able to set about designing a solution that is meaningful, compelling, and engaging. This informed a range of design

decisions, some fundamental (for example, how to break complex tasks into smaller but meaningful ones); and some much smaller ones (for example,. which of the myriad words we could use for this field will people most likely understand).

The ability to deliver a good quality and compelling user experience requires an iterative approach to problem solving: a willingness to generate, test, and improve upon solutions; and to do so through ongoing engagement with real users. Several rounds of user acceptance testing (UAT) at key points in the design process have enabled Jisc to have confidence in the fundamentals of the service, and (through benchmarking) allowed the team to demonstrate continual improvement in usability, as well as identifying areas that require improvement.

C. Data model

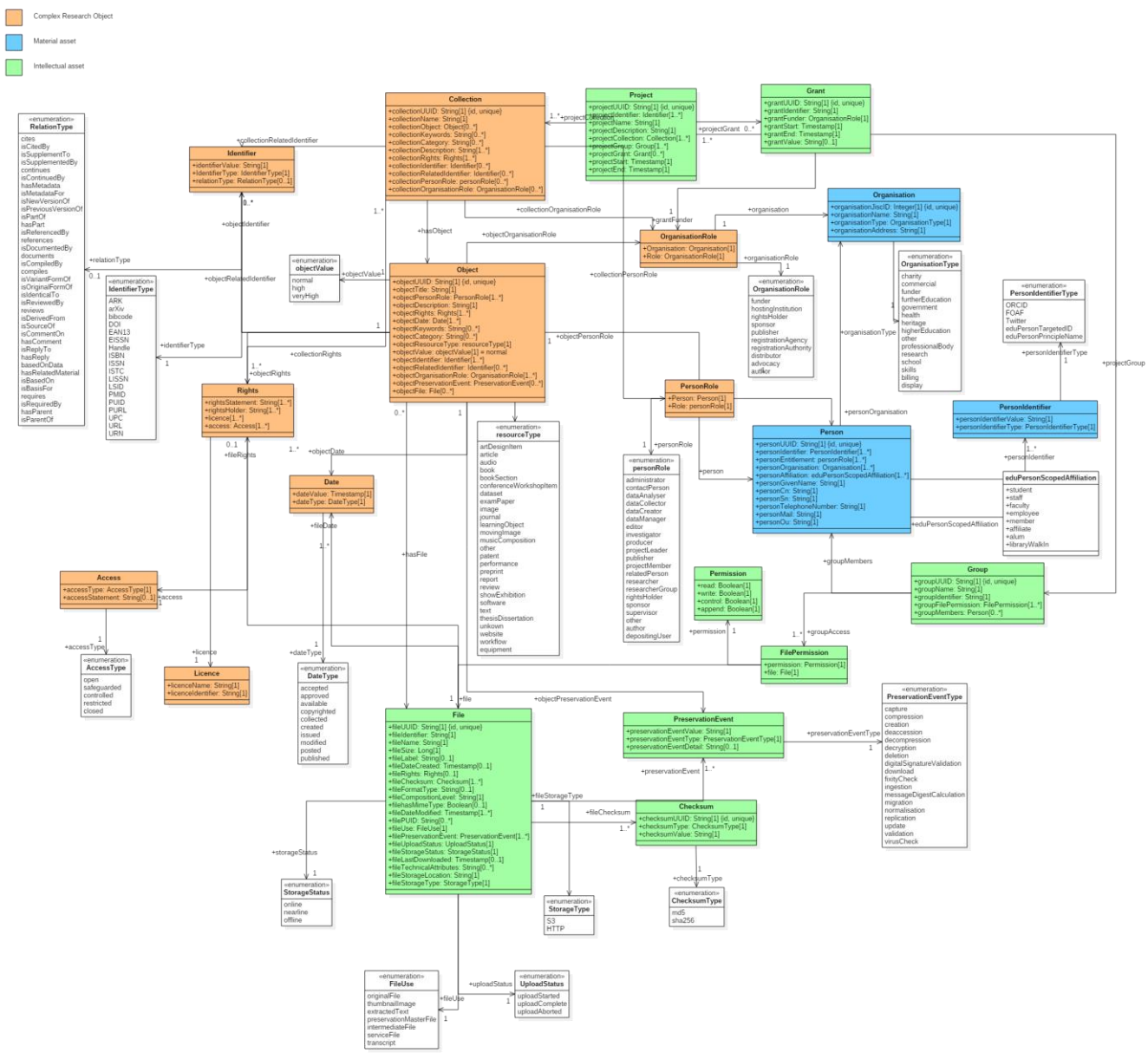

Figure 4 - JORH canonical data model

The canonical data model shown above in Figure 4 underpins the message specification used throughout (and in particular in the messaging layer and APIs). Built through consultation with the UK HEI sector and utilises best practice in popular metadata standards, schemas and ontologies, such as Datacite, Premis and Cerif. The data model is a living document, iterates based on the requirements of the service and the third-party integrations it supports. The documentation and the model (including every version) are openly available in a Github repository [2] and can be used as the single source of truth for

metadata mapping between the service and any other domain. The interoperability framework the data model provides is critical for JORH to integrate with other services and systems, and why it is essential to make this document public and transparent.

V. CONCLUSIONS

In developing this service Jisc have taken an innovative approach to a problem—that of preserving research data—which is often approached purely from the technical angle. It's not a trivial task, even for specialists in preservation, but it is a task that is relatively well understood.

But in the real world preservation is in the hands of non-specialists, for whom this isn't their day job. They have little or no digital curation experience, would rather someone else did it and would rather be doing their research. They need to just be able to do it without re-training to be a digital archivist and with very little effort on their part.

So JORH has been deigned with this goal in mind; to be as cutting edge as possible, to be as open as possible in order to be as integrated as possible, to have the potential to be metadata rich, but without the need for manual entry and curation expertise.

In a nutshell; Integrated, innovative, extensible and user friendly.

REFERENCES

[1] S. Brown, R. Bruce, D. Kernohan "Directions for RDM in UK universities" https://www.fosteropenscience.eu/sites/default/files/pdf/1240.pdf

[2] https://github.com/JiscRDSS/Canonical-data-model/tree/master/Data-Model