Getting it together:

realising the value of museum collections data

This report describes a research project led by Collections Trust between August 2020 and March 2021, one of seven supported by the Open Data Institute's Stimulus Fund via their Innovate UK funded R&D programme.[1] The Stimulus Fund aimed to explore approaches that enable trustworthy and ethical sharing of data to help citizens and businesses lower their impact on the environment, improve public services, and save lives.

The researchers on our project included consultants associated with a data-sharing initiative known as *Preservation to Preservation* (P2P), involving four national museums and Cisco.

The research involved: a selective review of relevant literature; investigation of comparable initiatives that offered possible lessons; and 29 semi-structured interviews with experts from a range of sectors. Insights from this research informed an outline 'framework for change' that was used as the basis for consultation and discussion with key stakeholders and the wider museum sector. The research also identified many strong use cases for open/shared data from outside the sector, which will be the focus of a further report by the P2P team.

The Open Data Institute (ODI) would characterise the scenario we are working in as an example of a *data access initiative*:

1. Has a *clear challenge* that is the focus for the collaboration.
2. Involves *multiple stakeholders* actively working together to solve the problem.
3. Includes a strong focus on *collecting, using and sharing data*.[2]

Support from the Stimulus Fund gave us the time and opportunity to reflect on why past efforts to tackle this problem have failed, question many existing assumptions and re-think what an open ecosystem for collections data might look like. The ODI's involvement piqued the interest of our stakeholders and gave a fresh impetus to our discussions with them.

1 What is the challenge we want to address?

Collections Trust's mission is to help museums capture and share the information that gives their objects meaning. That information goes to the very heart of what it means to be a museum. In the words of the Museum Association's code of ethics, museums 'preserve and transmit knowledge, culture and history.'[3] Yet almost every museum we know struggles with the information management needed to fulfil this responsibility and it is holding them – and the wider sector – back.

1.1 Bringing data together in the first place

The first problem is how to connect all the millions of existing object records – not as an end in itself, but as the prerequisite for any and all further use of that data. Millions of digital records about museum collections spanning almost every discipline remain fragmented in

[1] https://theodi.org/article/join-us-as-we-embark-on-the-fourth-year-of-our-rd-programme/
[2] https://theodi.org/article/what-do-we-mean-by-data-access-initiatives/
[3] https://www.museumsassociation.org/campaigns/ethics/code-of-ethics/#

the databases of at least 1,700 museums. It is impossible for human researchers or AI tools to search across this data, let alone use it at scale for any purpose whatsoever.

Many – perhaps more than half [4] – UK museums *do* share their collections data in the sense that they publish at least some of it on their own websites. Some larger, tech-savvy institutions also offer data for re-use via APIs. But it remains true that, apart from the oil paintings and sculptures on Art UK's national platform,[5] the only way to search across all online collections is to visit hundreds of museum websites, one by one.

It might seem strange to raise this as a problem when the Arts and Humanities Research Council (AHRC) is spending a very welcome £19m on the ambitious research programme *Towards a National Collection*.[6] However, its focus is on collections data *already published online*, especially by those larger institutions classified as 'independent research organisations'.

From our perspective, there is also the more basic need to help *all* UK museums share collections data currently sitting in standalone computers up and down the country, so it can be used by anyone, not least by museums themselves in ways beyond their own individual capacity.

1.2 Demonstrating the value of sharing collections data

The second problem our research addressed was how to convince the museum sector of the value of sharing its collections data openly. This considered some of the bigger strategic opportunities offered by the digital revolution transforming so many aspects of our lives, and the lessons that museums might learn from other sectors further down the road of opening up their data.

These opportunities include:

- Democratising the interpretation of collections, allowing anyone to help shape our understanding of objects and their meanings.
- Reimagining learning and research based on collections.
- As a sector, achieving the critical mass to work at scale with private-sector innovators and make an impact in an increasingly crowded 'attention economy.'

The findings of this more expansive enquiry are not easily summarised in a brief report such as this, which focuses on a number of priority actions. The P2P researchers plan to disseminate this wider work elsewhere.

2 A framework for change

Over the course of the ODI-funded project we arrived at a potential way forward that was much looser than we might have imagined at the start. The ODI would call what follows a *data infrastructure* (comprising *data assets* supported by *people, processes* and *technology*), but in talking to stakeholders we have found it better to talk about a *framework for change*.

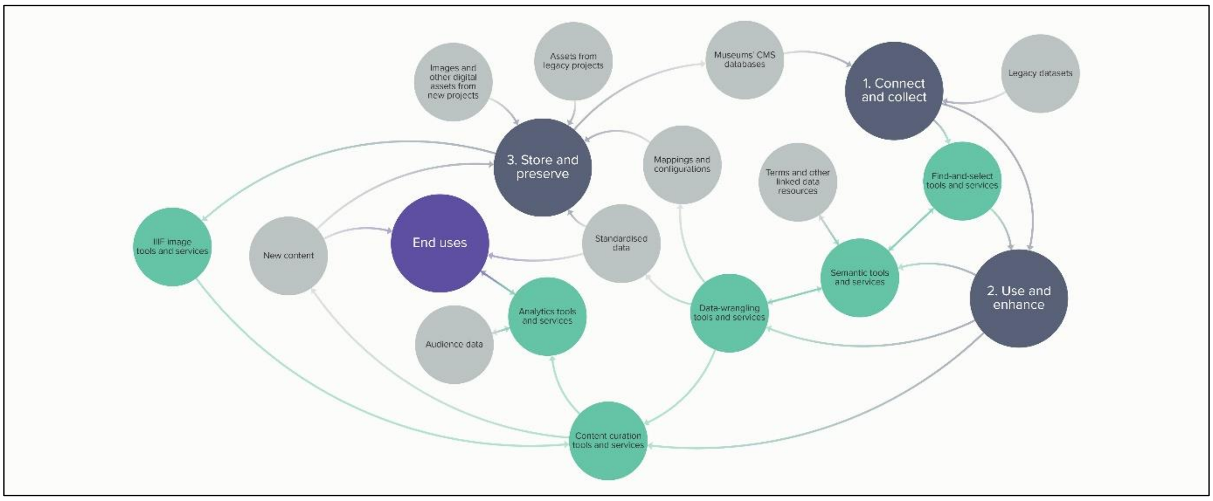
Given the large number of stakeholders involved in various aspects of working with museum collections data, one aim of the framework is to help everyone see and understand where their activities fit within the overall picture, or where they might usefully focus their efforts in future.

[4] https://collectionstrust.org.uk/blog/remotely-possible-access-to-collections-data-during-lockdown/
[5] https://artuk.org/
[6] www.nationalcollection.org.uk

There are three main activities within the framework:

- **Connect and collect**: the process of simply gathering data from museums however they can provide it and making it available – unprocessed - as the raw material for any potential use.
- **Use and enhance**: firstly, being able to find and select the right raw material; from there, any and all onward uses of it – processing it into standardised formats, using raw data for research, or curating digital products.
- **Store and preserve**: not only the obvious storing of image files, etc, from digitisation projects, but also capturing and futureproofing digital outputs such as interpretive text, research notes, data mappings and AI configurations.

The diagram below shows an overview of the framework. The large blue circles represent the three main activities of the framework. The grey circles represent various kinds of data, while the green ones indicate the tools and services that might form part of the 'use and enhance' ecosystem. An interactive version is available on the Collections Trust website.[7]


Images and other digital assets from new projects
Assets from legacy projects
Museums' CMS databases
1. Connect and collect
Legacy datasets
3. Store and preserve
Mappings and configurations
Terms and other linked data resources
Find-and-select tools and services
IIIF image tools and services
New content
End uses
Standardised data
Semantic tools and services
2. Use and enhance
Analytics tools and services
Data-wrangling tools and services
Audience data
Content curation tools and services

The framework includes four key interventions that Collections Trust proposes to focus on, in partnership with others:

- A '**connect and collect**' **service –** a minimum viable solution for bringing together collection records from all museums as the raw material for any and all uses.
- In the 'use and enhance' part of the framework:
  - A '**find and select**' **tool** allowing users to find and select the data they want to work with, as the first step in countless scenarios using digital collections.
  - A generic '**content curation**' **tool** that demonstrates how museums might capture and re-use knowledge generated in the course of projects and collaborations.
- Underpinning the 'store and preserve' part of the framework, a sector-wide **digital preservation strategy**, making better use of existing funding to improve the digital storage arrangements of many hundreds of museums and futureproof the benefits of short-term projects.

[7] https://collectionstrust.org.uk/tapping-our-collections-potential/framework/

Below, we describe the activities that would take place within the framework.

2.1 'Connect and collect'

Addressing the first of the problems that make up our challenge, we propose an enabling **'connect and collect' service** that would, as its core function, simply gather a lake of source data from museums in whatever form they could provide it. Then, with minimal processing, make it available online as the raw material for *any* potential use. Of course, most end uses would need the data to be standardised in specific ways, but that would happen within the 'use and enhance' part of the framework as described in the next section.

This approach avoids two key problems faced by 'traditional' cultural heritage aggregators that bring together data from different institutions and present it in a standardised way. Firstly, standardising such data is time-consuming and expensive for the contributors or the aggregating service or both. Secondly, the harmonising process imposes standards that might suit one end purpose, but not others. The nuances and richness of the source records can get lost in translation. In our stripped-back proposal the raw data would remain available to those who needed it that way.

Only non-confidential data would be collected and only metadata about images, etc, rather than the files themselves. The original field names of the source data would be retained and, beyond identifying the object number field, there would – *at this stage* - be no mapping to any metadata schema, nor any other processing to standardise the data in any way.

Metadata would be added about the source of the data, together with licensing information for onward usage. We would strongly encourage open licensing of data, but not insist on a public domain declaration.[8] Working on projects within the Europeana ecosystem we have found this to be a step too far for museums that would otherwise have taken part.

One technical option would simply be to hold the records as JSON[9] within Elasticsearch.[10] Conceptually, this is little more than museums agreeing to use the same file-sharing platform.

Such a service could start now and start small, working with existing initiatives that share a need to collect data from various UK museums – such as Art UK's planned data harvesting service [11] and the *Towards a National Collection* research projects.[12] This would go a long way towards building a critical mass of raw collections data available for use elsewhere.

2.2 'Use and enhance'

This part of the framework is deliberately broad and loosely defined, since it potentially includes anything that anybody might want to do with collections data, from routine collections management tasks to creating highly innovative digital products aimed at consumer audiences. In the framework, these and all other potential outputs are grouped together under the single heading 'end uses'.

[8] Eg https://creativecommons.org/publicdomain/zero/1.0/
[9] JavaScript Object Notation – see https://en.wikipedia.org/wiki/JSON
[10] https://www.elastic.co/elastic-stack
[11] https://youtu.be/glwdp4xaE6Y
[12] https://www.nationalcollection.org.uk/projects

Our focus here is on the tools and services that would help those working towards whatever end use they have in mind. We envisage a broad-based ecosystem – both commercial and non-commercial – including the following:

- The **'find and select'** tool proposed as one of our key interventions. It would not aim to present curated content to the general public but would be a tool to help anyone developing such content to find and select their raw material. It would allow users to search across the raw data gathered by the 'connect and collect' service, and/or processed datasets held by others. Users would be able to select relevant records and either send them onwards via an API to other tools and services or download them for offline use. We see this as a community-owned tool, starting simple and improving collaboratively through incremental project-based development (eg research into using AI – artificial intelligence – tools and services to increase the precision of searches across the data).
- **Semantic** tools and services to develop and maintain resources such as classifications and terminologies, and to apply such resources to mitigate the inconsistencies of the source data. Includes AI tools and services as well as ones to help manual annotation of source data. Here, the museum sector mainly needs to keep abreast of wider developments across the semantic web, avoid duplicating effort, and contribute content and curatorial expertise where helpful. A proposed (but currently unfunded) update to the Reference Data Manager of the open-source Arches platform would be a useful addition to this part of the framework.[13]
- **Data-wrangling** tools and services to produce standardised versions of inconsistent source data, as needed for various end purposes. Art UK, for example, would want to map the raw records it needed to its required format (based on the LIDO metadata schema)[14] and standardise artist names, etc, using linked data resources. Its proposed data harvesting service includes tools to do both those tasks, and these tools will be potentially configurable for other uses by other people (eg mapping natural history records to the Darwin Core standard for GBIF).[15]
- **Content curation** tools and services that allow a wide range of users to create new content from the raw material of source data - in such a way that improves workflows and anticipates the need to capture additional content for potential future re-use. To get the ball rolling, Collections Trust hopes to develop a generic tool that demonstrates the principle and can be optimised for specific use scenarios.
- If all museums could store their images in digital repositories that supported the International Image Interoperability Framework, **IIIF** tools and services could boost even small museums' online content.
- **Analytics** tools and services could support sophisticated (and legally compliant) tracking of content use, with benchmarking and analysis drawing on wider data about audiences and their access to online culture, with the potential to cross-refer to attendance data from physical venues.[16]

Much of this work is already happening to some extent, or will be taken further by initiatives such as *Towards a National Collection*. This might include improved cross-collection discovery, inter-disciplinary research, crowdsourcing, and linking collections data to other datasets to enable new forms of engagement and analysis. However, across the whole sector it lacks coordination. The trick will be to pull together the many different strands so

[13] https://www.archesproject.org/roadmap/
[14] http://cidoc.mini.icom.museum/working-groups/lido/
[15] https://www.gbif.org/darwin-core
[16] Eg https://audiencefinder.org/

that the whole is greater than the sum of its parts. Crucially, we propose that these tools and services are designed in such a way that enhanced data and new content can be captured for re-use as seamlessly and easily as possible.

2.3 'Store and preserve'

In this part of the framework, we would like to see expanded provision of trustworthy digital repositories available to any museum, sector organisation or collaboration that needs them. These do exist, along with well-established standards that allow users to have confidence in the service provided.[17]

The fourth key intervention Collections Trust proposes is a **digital preservation strategy** for the UK museum sector. As noted already, the scope of this would extend beyond the obvious assets resulting from digitisation projects; a key aim would be to preserve a wide range of outputs that too often gather digital dust in long-forgotten spreadsheets and documents.

For once, the problem is not really a financial one: museums around the country already spend money each year on *ad hoc* digital storage arrangements, while many millions in grant-funding continues to be available for projects that create new digital assets of various kinds. Economies of scale could fund better solutions for preserving legacy digital assets, while funders could require project budgets to include enough to endow the preservation of newly created assets to agreed standards.

3 Leadership and governance

A 2020 survey of national aggregators by the Europeana Common Culture project found that two-thirds had a legal mandate from their government.[18] Some were created as the result of government initiatives, while others reported some kind of official recognition within the cultural heritage ecosystem of their countries. In the survey report this is cited as a critical success factor.

In the UK, one barrier has undoubtedly been the patchwork nature of the museum sector, not least the devolution of cultural policy across the home nations and the arms-length nature of the relationship between government and leading sector bodies. No single museum has a mandate to lead, unlike the British Library or The National Archives.

3.1 Aligning with developments in research infrastructure

The two organisations leading the way at the moment are AHRC and its umbrella body, UK Research and Innovation. The *Towards a National Collection* programme is already funding eight small 'foundation projects' and currently shortlisting bids for five 'discovery projects', each of which will receive up to £3m over three years.[19]

Moreover, this research activity is happening at the same time as more ambitious, longer-term proposals are being developed in the context of UKRI's infrastructure roadmap. The potential investment opportunities identified by UKRI include some highly relevant to the framework we propose, including shared digital storage and aggregation of cultural heritage data.[20]

[17] Eg https://www.coretrustseal.org/

[18] Landscape of national aggregation in Europe and establishment of emerging national aggregators

[19] https://www.nationalcollection.org.uk/funding-calls

[20] https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf

The interest and investment in our sector by AHRC-UKRI is hugely welcome. They understand data and its infrastructure needs, their UK-wide remit cuts across the fragmented museum landscape noted above, and they have access to enviable levels of funding. The whole museum sector needs to engage with these important initiatives, and vice versa.

3.2 A museum data service?

To help the academic sector work more effectively with 1,700 museums, it would be useful to have an intermediary partner with a foot in each camp and a mandate supported by key stakeholders. The Museum-University Partnership Initiative funded by Arts Council England,[21] and Leicester University's AHRC-funded *One by One* projects,[22] hint at what might be achieved by more sustained engagement between the two sectors.

In our research, we also looked at the Archaeology Data Service (ADS), based within the University of York.[23] This offers several useful lessons to the museum sector, not least a business model that has stood the test of time over 25 years and is considered in more detail in the next section. Here, it is worth noting that the ADS not only provides a practical service to its sector, but also leadership at national and international level on the development of relevant standards, policy and practice.

We think the time is ripe for a comparable data service to meet the needs of the UK museum sector. Given the interest of the research funding councils noted above, it would make sense for the proposed museum data service to be based within an institution eligible to receive UKRI funding. One option might be one of the national museums and galleries classed as independent research organisations (IROs). Better might be a university with existing links to all kinds and sizes of museums, and also with a track record of collaborating with sector bodies. This would secure buy-in across the whole museum landscape, reinforced through advisory boards that represented the full range of stakeholder interests and expertise.

At its core, the museum data service would maintain the proposed 'connect and collect' service. In partnership with others, it would also develop the proposed 'find and select' tool.

Within the 'use and enhance' part of the framework the museum data service would be well placed to show leadership in the field, using its convening and fundraising power to broker partnerships, projects and other collaborations, including international ones.

Finally, the proposed museum data service *could* provide a trustworthy digital repository, following the 'store and preserve' business models described below, but these could equally be delivered by other providers within the overall context of a sector-wide strategy for digital preservation.

3.3 Guiding principles

As stated already, one aim of the framework suggested in this report is to help everyone involved in sharing and using collections data to see how their work fits into the bigger picture. The scope of potential activity is huge, as is the number of potential stakeholders.

We think that some guiding principles will be needed to turn the framework into sustainable practice. These might be developed and agreed by the representative stakeholders advising the proposed museum data service. Existing guidance relevant here includes the FAIR principles for findable, accessible, interoperable and reusable data,[24] and the over-arching values of the Digital Culture Charter.[25]

[21] https://www.publicengagement.ac.uk/nccpe-projects-and-services/completed-projects/museum-university-partnership-initiative
[22] https://one-by-one.uk/
[23] https://archaeologydataservice.ac.uk/
[24] https://www.go-fair.org/fair-principles/
[25] https://digitalculturecompass.org.uk/charter

Below we suggest some more specific principles. They are based on hard-won experience of many data-sharing initiatives in the UK and elsewhere over recent decades.

Be helpful

- Make life easier for museums, not harder.
- Make state-of-the-art tools and services available to all museums, not just the big ones.

Be open

- Encourage data to be licensed openly, but don't insist on it.
- Build core infrastructure with standards-based, open-source tools.

Be flexible

- Accept data however museums want to supply it.
- Take a modular approach, with core services kept as generic as possible.

Be sustainable

- Unless it is intended to be ephemeral, don't create digital content without a plan to preserve it long-term.
- Keep core services to the minimum likely to be affordable over the long term.

4 Business models

In this section we consider some of the insights gained through the research about business models that might be appropriate across different parts of the framework. Above all, we accept that the short-term nature of funding for museum digital activity is unlikely to change. The challenge is to ensure that the benefits of time-limited projects remain available over the long-term.

4.1 Connect and collect

One of the main insights to result from our research is that the fundamental problem we are trying to address has a relatively simple solution, and a correspondingly straightforward business model. Our proposed 'connect and collect' service would stop short of the kind of data processing that forms much of the cost of a running a 'traditional' aggregator such as Swedish Open Cultural Heritage (SOCH),[26] which we looked at as a case study. Moreover, in the case of SOCH there is an additional cost to be met by museums that want to contribute, because they must already publish their data online before it is aggregated.

In the UK context, we suggest that the most sustainable business model for bringing together collections data from museums is one that keeps the core service – and therefore the core costs - to the absolute minimum and pushes all the costs of processing that data into the 'use and enhance' part of the framework, as needs and funding opportunities arise. While further scoping work is needed, we believe that the annual cost of running the core 'connect and collect' service would be less than £100,000.

We propose that this core service should be grant-funded and should be free to all users. In order to gain traction, we argue there needs to be an initial commitment to the service of at least five years, and that it needs to be set up in such a way that it could be transferred to another home if the need arose in future.

[26] https://www.raa.se/in-english/digital-services/about-soch/

4.2 Use and enhance

In the course of our research, we looked beyond the museum sector to consider how it might scale up its engagement with digital innovators. Some potentially useful case studies, drawn from many more, are noted below.

Case study: 'Government as a Platform'

One of the UKs most successful digital institutions over the past decade has been the Government Digital Service (GDS)[27] which recognised that the efficient delivery of services meant building and managing an ecosystem of common platforms and easily updatable components. "Every service rests upon an ecosystem of components that can be snapped together or pulled apart whenever needed… In a world of platforms, you find out what users need earlier in the process, so you know sooner whether or not you're building the right thing. When it's so simple to create services, you can create them as experiments. They can be almost disposable." [28] This approach has influenced some early experiments with 'Charity as a Platform' [29] and the modular nature of our proposed framework for museum data.

If the cost of experimenting is brought down by adopting this platform-based approach, public sector bodies could stimulate innovation at relatively low risk, as in Helsinki.

Case Study: Forum Virium

Forum Virium is an innovation organisation owned by Helsinki City Council.[30] It has the flexibility to move more quickly and absorb the risk of experimentation involved in innovation. If projects work, the city can then adopt them. This focus on innovation has attracted millions of euros of development investment into the city for injection into the growth of local business and international profile. Forum Virium champions an ecosystem of openly licensed assets created in pilot projects which anyone is free to utilise and turn into business (and jobs) for their company.

OpenActive is a good example of sector lead body stepping in to enable a service that helps meet a strategic goal.

Case Study: OpenActive

The OpenActive data access initiative has created data standards and common tools to help people to book activity sessions using a variety of website and apps, solving a problem that many are put off becoming active because it is hard to find suitable options nearby. The project evolved out of a venture-funded startup trying to build an app, but the commercial model did not work out. The ODI and ukactive steward the initiative, and receive National Lottery funding through Sport England, which sees it as a key tool of its remit to create an active, healthy nation.[31] As well as a steering group drawn from these three organisations, there is also a sustainability working group drawn from the sports sector, providing recommendations on the future model and direction of the initiative.

[27] https://gds.blog.gov.uk/about/
[28] https://gds.blog.gov.uk/2015/10/07/government-as-a-platform-for-the-rest-of-us/
[29] https://medium.com/doteveryone/charity-as-a-platform-a-prototype-81f0cfa567a5
[30] https://forumvirium.fi/en/
[31] https://www.openactive.io/about/

4.3 Store and preserve

In this part of the framework, we see scope for multiple providers of digital preservation solutions, each meeting recognised standards. Some may be standalone services, others based within existing cultural heritage or academic institutions with spare capacity to sell. The question is how to pay for the level of digital preservation the museum sector needs.

Economies of scale

For legacy assets, including the results of digitisation programmes dating back over more than two decades, our proposal is that museums should combine the purchasing power that is currently spread between countless *ad hoc* solutions. Museums are already spending money on digital storage; by coordinating their procurement, it seems highly likely they could get better, proactively managed digital *preservation* for the same or less. Further research is needed to build a solid business case for sector-wide collaboration.

Micro-endowment

For newly created assets resulting from grant-funded projects, we propose micro-endowment, a tried-and-tested model for buying long-term digital preservation. The University of York's Archaeology Data Service (ADS) provides a compelling case study. ADS preserves the digital archives created during archaeological fieldwork and charges a one-off fee up front that goes into an endowment fund, income from which is part of a mixed business model that also includes participating in a range of research projects. The main funders of museum digital projects could usefully require grant recipients to allocate enough budget to endow the long-term preservation of the outputs in a trustworthy digital repository.

Case study: Permanent Legacy Foundation

The US-based Permanent Legacy Foundation, also operates a micro-endowment business model.[32] The current calculation is around $10 per GB (gigabyte) of storage. Users are charged a one-off fee depending on the size of the archive they wish to upload, and the Foundation adds this lump sum to its endowment investments. The annual interest generated covers the cost of the multi-cloud storage negotiated with commercial providers, and also the Foundation's operating expenses, such as data migration, integrity checking and format conversion.

5 Next steps

Collections Trust will continue the process of consensus-building after the end of the current project. In particular, we will explore with Art UK and a potential academic partner how to develop the 'connect and collect' service and 'find and select' tool to meet the needs of the whole museum sector. We also hope to demonstrate the proposed content curation tool as part of AHRC-funded work with smaller museums.[33] We have learned a great deal from the ODI during the project, and hope to build on the relationship over coming months and years.

Inevitably with such a wide-ranging study condensed into a brief report, much useful material remains on the cutting room floor. This is true of many conversations with interviewees about how – if the basic building blocks were in place – opportunities for innovation would open up and bring museums up to speed with other sectors that have discovered the value of sharing data in an open ecosystem, mixing community enterprise with entrepreneurial spirit. The P2P project will continue to develop this thinking in its own work, itself just such a mix.

[32] https://www.permanent.org/
[33] https://archaeologydataservice.ac.uk/research/makingItFair.xhtml