Unlocking the potential of digital collections

A call to action

TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

Publication date:
21 November 2024

Report authors:

- **Rebecca Bailey**, Programme Director, Towards a National Collection
- **Dr Javier Pereda**, Senior Researcher, Towards a National Collection
- **Dr Chris Michaels**, Human Economics
- **Tom Callahan**, Human Economics

TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

About the Arts and Humanities Research Council
The UKRI Arts and Humanities Research Council (AHRC) funds internationally outstanding independent researchers across the whole range of the arts and humanities: history, archaeology, digital content, philosophy, languages and literature, design, heritage, area studies, the creative and performing arts, and much more. The quality and range of research supported by AHRC works for the good of UK society and culture and contributes both to UK economic success and to the culture and welfare of societies across the globe.

Contents

Recommendations

Abbreviations

A

AI	artificial intelligence
AHRC	Arts and Humanities Research Council
API	application programming interface

B

BBC	British Broadcasting Corporation
BL	the British Library
BM	the British Museum

C

CARE	collective benefit, authority to control, responsibility and ethics
CC	Creative Commons
CGDC	community-generated digital content
CIDOC	International Council for Museums International Committee for Documentation
CLIP	Contrastive Language-Image Pre-training
CMS	content management system
CoStar	convergent screen technologies and performance in real time
CRM	Conceptual Reference Model
CV	computer vision

D

DaaS	data as a service
DAM	digital asset management
DiSSCo	the Distributed System of Scientific Collections
DMA	data maturity assessment

F

FAIR	findable, accessible, interoperable and reusable

G

GDPR	General Data Protection Regulation
GIDA	Global Indigenous Data Alliance
GLAM	galleries, libraries, archives and museums
Grad-CAM	gradient-weighted class activation mapping

H

HES	Historic Environment Scotland

I

ID	interactive design
IDGov	indigenous data governance
IDSov	indigenous data sovereignty
IIIF	International Image Interoperability Framework
IML	interactive machine learning
IP	intellectual property
IRO	Independent Research Organisation

J

JUSP	Journal Usage Statistics Portal

L

LIDO	Lightweight Information Describing Objects

M

MARC	machine-readable cataloguing
MIDS	Minimum Information about a Digital Specimen
MDS	Museum Data Service
MIC	Małopolska Institute of Culture
ML	machine learning

N

NGS	National Galleries of Scotland
NHM	Natural History Museum
NIMOZ	National Institute for Museums and Public Collections
NLHF	National Lottery Heritage Fund
NLP	Natural Language Processing
NLW	National Library of Wales
NMS	National Museums Scotland

O

OA	Open Access
ODI	Open Data Institute
ODMM	Open Data Maturity Model
OHOS	Our Heritage, Our Stories

P

PIDs	Persistent Identifiers

R

RCAHMW	Royal Commission on the Ancient and Historical Monuments of Wales
R&D	research and development

S

SaaS	software as a service

T

TaNC	Towards a National Collection
TNA	The National Archives

U

UKRI	United Kingdom Research and Innovation
UNIDRIP	United Nations Declaration on the Rights of Indigenous Peoples
UX	user experience

V

V&A	Victoria and Albert Museum

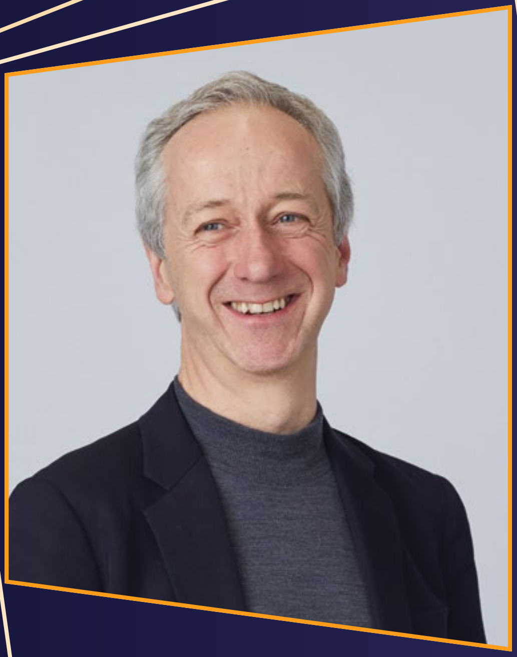
W

WHG	World Historical Gazetteer

Foreword

Sir Roly Keating

The collections held in museums, galleries, archives and libraries across the UK are one of the great assets of the world. Together they comprise an astonishing resource of knowledge – one that has the potential to change lives, bring people together, and drive innovation and growth.



At the moment, however, we are failing to make the most of this great national asset. Through a historic lack of investment in skills, digitisation and common infrastructure, the remarkable collections held in the UK remain fragmented and, in many cases, hard to access, even for the most dedicated researchers.

This powerful set of recommendations presents a call to action for cultural heritage organisations and funding bodies to come together in pursuit of a critical goal: the development of an inclusive, unified, accessible, interoperable and sustainable UK digital collection.

In my own career – across both broadcasting and the cultural sector – I've seen the power of digitisation to transform the reach and impact of content that would otherwise be inaccessible. It's a power which can make a real difference to people's lives, by connecting individuals of all backgrounds with experiences and ideas they would never otherwise encounter.

Taken together, the collections which Towards a National Collection seeks to unite encompass a vast span of data, information, knowledge and human experience: from science, technology and the natural world to art, culture, history and society.

Getting this right holds benefits for so many different parts of UK society, as well as for those using our collections from around the world: for researchers building new knowledge; for entrepreneurs looking to innovate and contribute to economic growth; for communities who may previously have felt excluded from the 'national' narrative; for any one of us who wants to find out more about who we are and where we've come from.

One of the most inspiring achievements of the Towards a National Collection programme has been to forge an unprecedented coalition of museums, galleries, archives and libraries – big and small, scientific and cultural, national and regional – all stepping out of their individual specialism or locality and agreeing to work together for the common good.

That working together is what this document is all about. Building on lessons learned in the course of the programme so far and drawing on all the evidence from the newly commissioned research, it sets out a roadmap for best practice in the digital transformation of collections, with practical guidance and lessons to learn for all of us.

It has benefited immeasurably from advice from 50 institutions across the UK who took part in consultation, as well the expertise of our dedicated Steering Committee. I am deeply grateful to everyone who contributed so generously and helped to shape these findings.

If you are involved in the collections sector, whether as a collections holder or a funding body, I strongly encourage you to adopt what you can of the thoughtful recommendations in the pages that follow. It's a way that each of us can play a role in bringing the long-term vision just a few steps nearer.

It's important to stress, though, that this document is just one milestone in the Towards a National Collection journey. In the months ahead, the full findings of the five major Discovery Projects will, in different ways, provide crucial insight into key aspects of the wider vision, from inclusivity to interoperability, and much else besides.

All of the findings and evidence from the programme will be presented by the Steering Committee, on behalf of the sectors we represent, to the Arts and Humanities Research Council (AHRC) who funded this work.

I know that AHRC wants to work with leaders across the sector, devolved nations, relevant government departments and major funders of digital collections to build on this work. We need to come together to develop the case for an inclusive, large-scale national investment in digital collections and the research infrastructure that underpins and delivers benefits from them.

That investment – in the people, the skills, the systems and the software so desperately needed to ensure the UK's national digital collection is truly sustainable and secure – has the potential to make the UK a global leader in collections-based research. In so doing, it will contribute not just to economic growth in our own country but to fostering, through increased international access to culture, science and history, a better, more humane world.

Roly Keating

Sir Roly Keating, Chair, Towards a National Collection Steering Committee

Executive summary

This document presents the policy recommendations of Towards a National Collection (TaNC), a five-year, £18.9 million investment in the UK's world-renowned museums, archives, libraries and galleries, with funding provided through UK Research and Innovation's Strategic Priorities Fund and delivered by the Arts and Humanities Research Council (AHRC).

The recommendations detail an end-to-end process, with case-study examples and supported by detailed training materials, which we ask UK cultural heritage institutions and their funders to adopt to help build a unified UK digital collection.

The adoption of these recommendations will take place as the AHRC develops options for a digital research infrastructure that would create the investment necessary to fund the elements of these recommendations beyond the capability and capacity of individual institutions.

TaNC believes that a unified UK digital collection will achieve transformational outcomes by breaking down the barriers that exist between the UK's outstanding cultural heritage collections and unlocking their full potential for our cultural, social and economic good.

If the sector can adopt more common ways of working, ensuring their digital collections are created, stored and organised using agreed technical standards; protected with appropriate cybersecurity provisions; and preserved in ways that keeps them available for generations of users to come, they could be shared across a UK-wide technology infrastructure that would enable the public, specialist researchers and key communities of interest to access and engage with our cultural heritage in ways not yet even imagined.

A UK digital collection will enable collaborative R&D, bringing together different disciplines to create new technological innovations and new knowledge. It will create spillovers, unforeseeable at this point, in the sciences, helping to bring the humanities and sciences closer together. And as we enter the age of artificial intelligence (AI), it will help both improve the quality of AI models and data – providing world-class training data with appropriate ethical controls – and unlock the value of cultural heritage assets and knowledge for what may prove to be the 21st century's key technology.

To build a UK digital collection, we need to:

1 **Selection**
Broaden our approach to what we include in our digital collections and expand who participates in the process of creating them

2 **Production**
Accelerate how data is produced and allow new technologies to accelerate its enrichment

3 **Skills**
Build upon the skills in data creation already within digital collections organisations through a scalable, sustainable approach to accessing advanced technology skills

4 **Reuse and rights management**
Adopt a coherent, consistent approach to data and rights management

5 **Access and engagement**
Make the platforms and infrastructure through which digital collections are used accessible to everyone

6 **Security**
Ensure our data and technology infrastructure, and the way they are used, are protected by common standards and legislation, as governed by good practice

7 **Preservation**
Take a long-term view on the preservation of data to ensure we can access it despite forces of change

8 **Impact**
Understand how our digital collections are used so we can fully harness their cultural, social and economic value

9 **Models and frameworks**
Treat digital collections as first-class research objects that allow us to transform our understanding of collections and the world

10 **Experimentation**
Continuously expose these collections to new technologies and new ways of thinking, while prioritising research into their environmental impact

Introduction

What are digital collections?

Our policy recommendations are focused on the development of digital collections and the digital research infrastructure that helps produce, manage and facilitate engagement with those collections. Building digital collections depends upon processes both of digitisation and digitalisation, terms we use throughout this paper, and explain below for clarity.

A digital collection is a group of files or databases preserved digitally and made accessible via the Internet or specific software. A digital collection will include digitised materials of both tangible and intangible content and the related knowledge necessary for its preservation, analysis, management and engagement. It also includes born-digital material like 3D scans, photographs and text documents.

A digital research infrastructure for digital collections is the ecosystem of tools and human and technology systems that open up the knowledge held in digital collections. Some of the key components of a digital research infrastructure include computational resources, software and access, and the human expertise of skilled professionals that use, maintain and implement these infrastructures.

Digitisation is the process of transforming physical objects or analogue information into digital formats. The process of digitisation generally includes the systematic enhancement and representation of these digital forms to ensure their accuracy and utility.

Digitalisation is the use and implementation of digitised content in a richer and broader context than its original sphere.

About the UK's digital collections

The size of digital collections in the UK is substantial but unevenly distributed.

In 2021, the Collections Trust was commissioned by Towards a National Collection (TaNC) to undertake a digital collections audit of 230 collections-holding cultural heritage institutions – the largest ever attempt to survey and benchmark the state of digital collections in the UK.

Between them, these 230 institutions hold nearly 146 million item-level records. A quarter of these item-level records – 37 million – have associated images or other digital media. However, just over 50% of these digital records and associated assets are held by only six institutions.

What do researchers and the public want from digital collections?

In late 2023, Claire Bailey-Ross of Portsmouth University was commissioned by TaNC to gain a comprehensive understanding of the needs and requirements of different research users across academia and Independent Research Organisations, and what they would like to see included in a future UK digital collections infrastructure. The consultation used six focus groups, 40 interviews and a survey with almost 200 responses to understand digital-infrastructure needs across various research fields and career stages.

There was strong support for connections to be built between and across collections, and across institutions. Researchers felt that standardisation was key to sustainability and interoperability. Researchers want a digital collections infrastructure to have true interoperation. Researchers felt it was important to balance technological advancements with the preservation of human expertise, and fostering community engagement in and across digital platforms.

Collaborative efforts were recognised as essential to address challenges and leverage opportunities in developing a digital collections infrastructure. By prioritising sustainability, enhancing search and discovery, establishing standardised frameworks for interoperability, and fostering collaboration, Bailey-Ross concluded that we can create a more inclusive and interconnected digital cultural heritage research environment.

How would users value a future UK digital collection?

The UK government has begun adopting an economic valuation technique called Total Economic Value to help better understand what estimated value users place on potential new types of digital service in hard-to-value areas such as arts, culture and heritage. We have used this method to help understand what value people would place on a UK digital collection and the digital collections research infrastructure that would be required to support it.

In 2024, Alma Economics undertook a study to calculate the Total Economic Value of a future unified digital collection of cultural assets in the UK. This involved the deployment of surveys across three distinct groups: the general population, formal researchers and those with a special interest. Over 8,000 individuals from across the general population participated and reported an average willingness to pay for this unified collection of £8 per person per year.

Willingness to pay in this context does not infer that we propose to create a 'pay to use' digital platform. Rather it indicates that they would be willing for their taxes or other means of public investment to be used, equivalent to £8 per person per year, to fund such a venture.

Extrapolating this to the full UK adult population leads to an estimate of the Total Economic Value for a UK digital collection and digital collections research infrastructure of around £425 million per year. The most popular reasons given for this willingness to pay from the general public were 'I believe this service would contribute to preserving our cultural heritage for future generations' and 'I believe this service would make it easier for people around the world to understand UK cultural heritage'.

In addition, a survey questionnaire issued by Alma Economics was completed by 326 academics and researchers, who reported a willingness to pay for the proposed service of around £150 per person per year. They reported that they wanted to use a unified collection in their education and research, and that it would open new lines of research. More generally they welcomed having access to assets they would not otherwise be able to see.

A further 267 survey responses were received from those with a special interest in cultural digital heritage. Their average willingness to pay was around £36 per person per year. These respondents, like those in the general population group, valued the service because they saw benefits in using it, as well as perceiving value in the existence of the service regardless of their level of use of it.

Across the work around Total Economic Value and consultation with research users, the message is clear: both the general public and researchers would highly value the creation of a UK digital collections infrastructure.

How we developed this paper

These policy recommendations have been prepared by Rebecca Bailey, TaNC Programme Director, working with Chris Michaels, Associate at Human Economics and Javier Pereda, TaNC Senior Researcher, supported by Tom Callahan.

The recommendations draw on nearly five years of research and development undertaken by the TaNC programme, as well as on the extensive feedback and advice provided by a broad range of stakeholders as part of our consultation process.

The paper consists of ten recommendations, supported by case studies and sample training materials plus an appendix with information on all of the research that has been grant-funded or commissioned by the TaNC programme.

Each recommendation is presented as a core proposition, alongside a series of steps which institutions and funders can take to work in alignment with that proposition.

Case studies are presented to support and expand upon the core recommendations. The case studies were chosen both from TaNC's own research programme and from the wider digital collections' community.

Sample extracts from the training materials commissioned by TaNC to help small and medium-sized institutions through the digital journey of managing their collections are also included in this document.

These training materials, which will be supported and made available for free, are included to illustrate how we can help adopting institutions and funders to act in line with the guidance provided here. Whilst the policy recommendations are intentionally high-level and overarching, the training materials provide considerably more detail and guidance on the implementation of these recommendations.

This paper is not a summary of the full research undertaken by the TaNC programme, although it draws from it. In addition to the policy recommendations and training materials, significant further programme outputs will follow, as each of our large-scale Discovery Projects draw to a close in the coming months and report their own detailed findings.

The paper's development was based on extensive consultation with the digital collections community. Consultation was undertaken with 50 organisations in two rounds in a process guided and endorsed by the TaNC Steering Committee chaired by Sir Roly Keating.

Access the training materials:
towardsdigitalcollections.org

TaNC Steering Committee

Advice on the development of these recommendations was kindly provided by the following individuals as part of our expert steering committee:

- **Roly Keating**, Chief Executive, British Library (Chair)
- **Maria Balshaw**, Director, Tate
- **Gemma Brough**, Deputy Director of Museums and Cultural Property, Department for Culture, Media and Sport
- **Caroline Campbell**, Director, National Gallery of Ireland *Gailearaí Náisiúnta na hÉireann*
- **Gus Casely-Hayford**, Director, V&A East
- **Tom Crick**, Chief Scientific Adviser, Department for Culture, Media and Sport
- **Kath Davies**, Director of Collections and Research, National Museum Wales *Amgueddfa Cenedlaethol Cymru*
- **Richard Deverell**, Director, Royal Botanic Gardens Kew
- **Catherine Eagleton**, University Librarian and Director of Collections and Museums, University of St Andrews
- **Andrew Ellis**, Director, Art UK
- **Liz Johnson**, Director, Museums and Collections Development, Arts Council England
- **Chris Michaels**, Independent Consultant
- **Sabyasachi Mukherjee**, Director General, Chhatrapati Shivaji Vastu Sangrahalaya (CSMVS) Museum, Mumbai
- **Ross Parry**, Director, Institute of Digital Culture, University of Leicester
- **Laura Pye**, Director and Chief Executive, National Museums Liverpool
- **Amina Shah**, National Librarian and Chief Executive, National Library of Scotland
- **Allan Sudlow**, Director of Partnership and Engagement, Arts and Humanities Research Council
- **Kathryn Thomson**, Chief Executive, National Museums NI
- **Johannes Vogel**, Director General, Natural History Museum Berlin
- **Esme Ward**, Director, Manchester Museum
- **Sohair Wastawy**, President, The Information Guild Consulting Group

The organisations consulted

Extensive advice on the development of these recommendations was kindly provided by 50 organisations and institutions, including the following:

- Alan Turing Institute
- Art Fund
- Arts Council England
- Art UK
- Association of Independent Museums
- British Film Institute
- British Library
- British Museum
- Collections Trust
- Congruence Engine Discovery Project
- Creative Commons
- Creative Informatics/Design Informatics, Edinburgh College of Art, University of Edinburgh
- English Heritage
- Historic England
- Historic Environment Scotland
- Historic Royal Palaces
- Imperial War Museums
- Jisc
- National Gallery
- National Library of Scotland
- National Lottery Heritage Fund
- National Museum Directors' Council
- National Museums Liverpool
- National Museums NI
- National Museums Scotland
- National Trust
- Natural History Museum
- Our Heritage Our Stories Discovery Project
- Research Libraries UK
- Royal Armouries
- Royal Botanic Garden Edinburgh
- Royal Museums Greenwich
- Royal Shakespeare Company
- Science Museum Group
- Sloane Lab Discovery Project
- Tate
- The National Archives
- Transforming Collections Discovery Project
- Unpath'd Waters Discovery Project
- University of St Andrews, Libraries and Museums
- Wikimedia

Our recommendations

The ten areas we make recommendations on are:

1 **Selection**
How to select materials from which to build digital collections

2 **Production**
How materials should be produced, the standards that should be applied to data and technology used for production

3 **Skills**
What key skills should be supported, developed and extended, and how we bridge the technology skills gaps in digital collections

4

Reuse and rights management
How to create sharable collections data

5

Access and engagement
What types of tools can be used to help people discover and engage with collections data

6

Security
How to secure digital collections data against digital harms

7

Preservation
How to preserve digital collections data for the long term against different forms of technical, organisational and economic change

8

Impact
How analytics, evaluation and audience insight can help you understand and amplify the value of digital collections

9

Models and frameworks
How to standardise approaches to building digital collections to help others

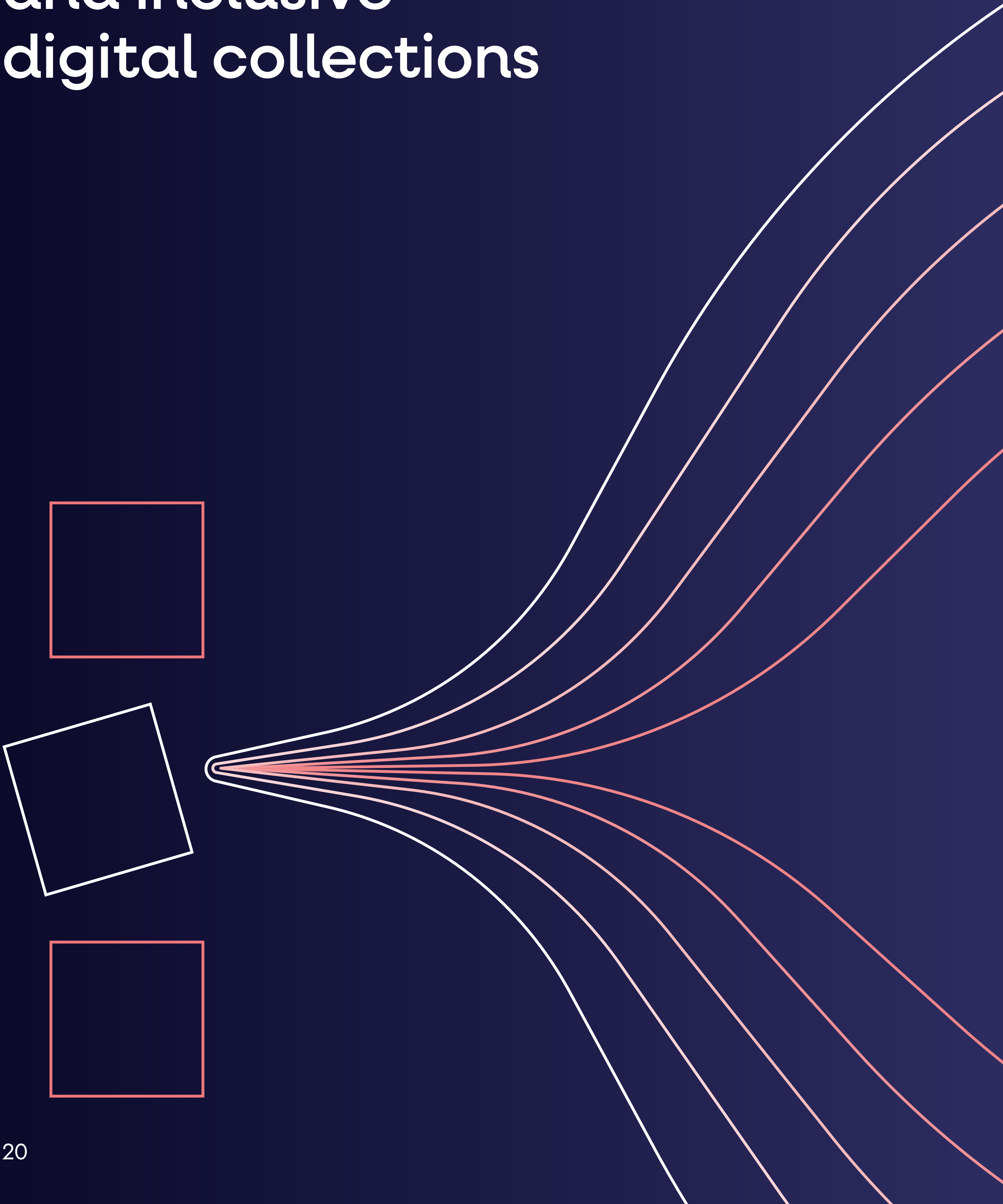
10

Experimentation
How we should systematically experiment with new technologies and new ideas to evolve our digital collections for the future

The following sections of this report deal with each of these recommendations in turn. Each section provides a more detailed explanation of our recommendation, along with example case studies and summaries of training modules.

1. Selection

Building diverse and inclusive digital collections



Our recommendation

To build a UK digital collection, we need to broaden our approach to what we include in our digital collections, and we need to expand who participates in the process of creating them. By doing so, we can create collections with more relevance to diverse and different audiences.

What do we mean by this?

Who selects what to include in digital collections? How is that selection made? What voices should be heard when we add to or create new digital collections?

Hundreds of millions of records already exist across the country, in many silos across many institutions. The prioritisation of, and rationale for, what to select to include in digital collections has many different motivations and circumstances at an institutional level, whether driven from demand and informed by audiences, building on work already done, or by issues such as the vulnerability to irreversible loss, and awareness of culturally sensitive heritage. And sometimes, it must be acknowledged that the best ethical decision is not to digitise some things at all.

However, if we are going to work towards creating an inclusive, unified, secure and sustainable UK digital collection, we need to do four things:

1. **Broaden our approach to what we include in our digital collections**, recognising the incredible variety of UK collections and the diverse types of knowledge that comes with them. TaNC has begun to unlock the potential of both new types of collections and of collections previously on the margins, such as intangible heritage, that needed new ways of understanding to realise their value. But huge varieties remain still barely touched.

2. **Diversify who participates in the process of making them**. The creation of digital collections is not a neutral activity and could embed and perpetuate historical biases. The teams who create data should be interdisciplinary and diverse, where possible representing communities from whom collections originated. The creation of digital collections can address biases by creating more equitable and diverse data, more representative of both contemporary society and the nuances and contested cultural perspectives from which it emerged.

3. **Ensure the data we create for digital collections is relevant for more diverse audiences**. By providing warnings of potentially harmful content, or where appropriate by using more current vocabulary and including new voices, data can represent a wider range of perspectives. Free from racist and other derogatory language, and after consideration of guidelines and frameworks such as the CARE principles (collective benefit, authority to control, responsibility and ethics), our data can be more fit for use and equitable, use respectful terminologies and help to build trust between communities.

4. **Reflect that evolutionary change is happening**. Whatever we do is building on decades, sometimes centuries, of materials that already exist. Much of this data describing our collections will be imperfect and represent the world views of different times and groups. We should not fear this but positively embrace that progressive change is happening, whilst understanding that change takes time and resources.

Case study 1:

Preserving and Sharing Born-Digital and Hybrid Objects From and Across the National Collection

This TaNC Foundation Project, led by Natalie Kane from the V&A Museum, alongside the University of London (Birkbeck) and the British Film Institute, highlighted the critical importance of including born-digital and hybrid digital objects for a broader and more inclusive UK digital collection.

Born-digital objects are diverse and complex and can range from virtual- and extended-reality experiences to social and digital platforms, such as Instagram, mobile applications and games, and the electronics and imaging technologies used to create them. The project focused on the governance systems and standards needed to support the management of these types of objects, as well as their digital preservation and conservation.

The project highlighted the importance of developing new cataloguing standards, vocabularies and data models to make digital collections relevant to diverse and contemporary audiences. To address this, the project extended the Linked Art Data Model, based on the CIDOC Conceptual Reference Model (CIDOC CRM), and mapped it to a decision-making model that showcases the diverse workflows when acquiring born-digital objects. By building on CIDOC CRM through the Linked Art Data Model, the project can provide a robust framework that enhances how art-related data is represented and shared, thus leveraging linked data principles to create more interconnected, accessible and reusable datasets. This approach helps ensure that the complexities and dynamic nature of digital objects and the platforms or technologies where they exist are adequately addressed.

This project highlighted the unique challenges of preserving born-digital content. It identified that digital objects can change and, in many cases, are made by other multiple digital objects. Digital collections need to evolve to accommodate the complex and dynamic nature of digital objects and the platforms where they exist. Addressing these changes requires innovative preservation strategies and ongoing efforts to maintain the accessibility and relevance of digital objects over time.

Case study 2:

Transforming Collections

The TaNC Discovery Project Transforming Collections, led by Professor susan pui san lok from University Arts London, addresses the structural and systemic biases inherent in the production, organisation, classification, categorisation and description of artists and artworks in collections.

Transforming Collections engaged with 15 project partners and collaborating organisations across the UK, including Tate, Arts Council, Art Fund, Art UK, Birmingham Museums Trust, British Council, Contemporary Art Society, iniva (Institute of International Visual Art), Jisc Archives Hub, Manchester Art Gallery, Middlesbrough Institute of Modern Art, National Museums Liverpool, National Museums Scotland, Wellcome Collection and the Van Abbemuseum, Eindhoven (NL).

The project brings together a diverse team of art historians, museum professionals and creative computing technologists to develop an interactive machine learning (IML) tool. Used critically and self-reflexively, the tool has the potential to help produce a more inclusive and representative cultural heritage collection by challenging and enriching existing data. The project underscores the importance of questioning how decisions around data are made – what is captured, classified, deployed, how and by whom. Critical approaches to ML development and research is essential to ensuring that the knowledge held in digital collections is relevant and meaningful to a wider audience. The IML tool was developed through an iterative process that emphasised equitable collaboration. The technology was shaped to serve the critical research needs of the team, focusing on identifying patterns in text and image descriptions. The development of this tool was informed through case studies that included critical analyses of institutional language, exploring texts related to various artists, and offering alternative interpretations of artworks that acknowledge problematic texts, methods and contexts.

The project identified patterns in the enduring rhetoric of benevolence, entreaty, wealth and health, which are rooted in philanthropic, paternalistic, and colonial ideologies and practice. By evidencing the recurrence of problematic or euphemistic language and habitual narratives, the project highlighted the under-resourcing of research within collections and underscored the need to engage global majority scholars and practitioners in relevant fields, ethically and equitably. The project also revealed how the privileging or exclusion of certain types of information based on racialised, gendered and ‘othered’ identities of artists perpetuates their marginalisation or omission from UK heritage and history. Moreover, the resurfacing of overlooked artworks within and between collections pointed to the need for long-term human resourcing to integrate individual and embodied knowledge into data records, supporting institutional memory and digital cultural heritage.

In addition to the research and development of the IML tool, the project has engaged four artists-in-residence whose work was showcased during a week-long public programme at Tate Britain and Tate Modern. Their practice research highlighted critical and creative approaches to challenging and examining existing collections knowledge. They pointed to potential histories and multivocal narratives, emphasising the opportunities and risks of machine learning in transforming ‘art’, ‘nation’ and ‘heritage’.

Case study 3:

Creative Commons Open Culture Platform

The Creative Commons (CC) Open Culture Platform is a community platform facilitated by Creative Commons' Open Culture Program, which has facilitated the sharing of best practices and collaboration on projects and advocacy for Open Access, which are pivotal in shaping a digital collections landscape that is inclusive, dynamic and accessible.

The platform plays a critical role in disseminating best practices for making digital collections available online and ensuring that diverse cultural expressions are captured and preserved. Advocacy for policies that promote Open Access to cultural resources has been a cornerstone of the Creative Commons Open Culture Platform. This ensures that digital collections are accessible and relevant to more diverse audiences. This was done by pushing for fewer restrictions and ensuring that the data within these collections catered to diverse cultural, educational and social needs.

Creative Commons enables policy-advocating initiatives through participation in legislative and policy discussions alongside policymakers, cultural institutions and key stakeholders. Furthermore, by collaborating with global institutions, it has produced partnerships that often result in pilot projects that further serve as case studies that support policy reform. This has also included the implementation of participatory approaches for content curation and the continuous incorporation of new voices into decision-making processes.

For example, the Los Angeles Contemporary Archive is notable for its approach to making archives more accessible and involving the community in the archiving process as active participants. They also invited community members to contribute their own materials to the archive. A further example is the *Festival Iminente*, a contemporary art and music festival that has developed and maintained an archive that travels with the festival whilst involving the diverse communities it visits. The festival's archive is both physical and digital, and, similarly to digitally born collections, is able to capture the ephemeral nature of the installations and the art created in the temporary spaces where it takes place.

Through the Creative Commons Open Culture Platform, the Creative Commons has opened access to cultural heritage, allowing free interaction with the collective history of culture and heritage and fostering a globally inclusive environment. Engaging with a diverse array of global voices can facilitate the enrichment of archival practices and expand the understanding of copyright and open licensing, which can empower groups and organisations to document and preserve unique cultural traditions.

Sample training module descriptions

Relevant module 1: Who is your digital collection audience?

The purpose of this training module is to help practitioners identify key audiences for their current or future digital collections and evaluate the effectiveness of different forms of digital collections for those audiences.

Beginner

Beginner learners will identify differences between analogue and digital audiences, and experiment with some existing tools to understand the latter. They will consider the differences between researcher and non-specialist audience needs for their collections. Two case studies will illustrate different approaches to online collections – one with a searchable online database and another with curated highlights of the collection.

Intermediate

Intermediate learners will review some of the sector learnings from the pandemic's digital shift and consider how this relates to their organisation. Case studies will demonstrate new approaches some organisations took in the pandemic and whether they have become part of 'business as usual' for their digital collections offer.

Advanced

Advanced learners will review and reflect in more detail about their own organisation's digital audiences framework or steps they would like to take to implement such a framework.

Relevant module 2: Biases in collections

This module encourages learners to consider the historical formation of their collections, and how the cultural attitudes of the time have contributed to the size, shape and nature of GLAM collections.

They will evaluate the impact of these historical biases within their own institution and consider how digital collections and digitisation could contribute to diversification of collections. Participants will also learn how to evaluate the risks of public engagement with debates over biases in collections such as negative press, against benefits of diversifying the representation of people within online collections, such as greater engagement from previously unreached communities.

Beginner

Beginners will gain an understanding of the Enlightenment and rationalistic roots of collecting across GLAM institutions and consider the types of tangible and intangible heritages in their own institutions and other organisations. They will consider historic injustices such as the collection and storage of human remains or objects considered sacred to some communities, and the implications for digitising and making accessible such collections which may require the application of indigenous knowledge structures. A case study will illustrate a difficult case in which the needs of researchers and ideals of open access were balanced against source community restrictions in the digital sphere.

Intermediate

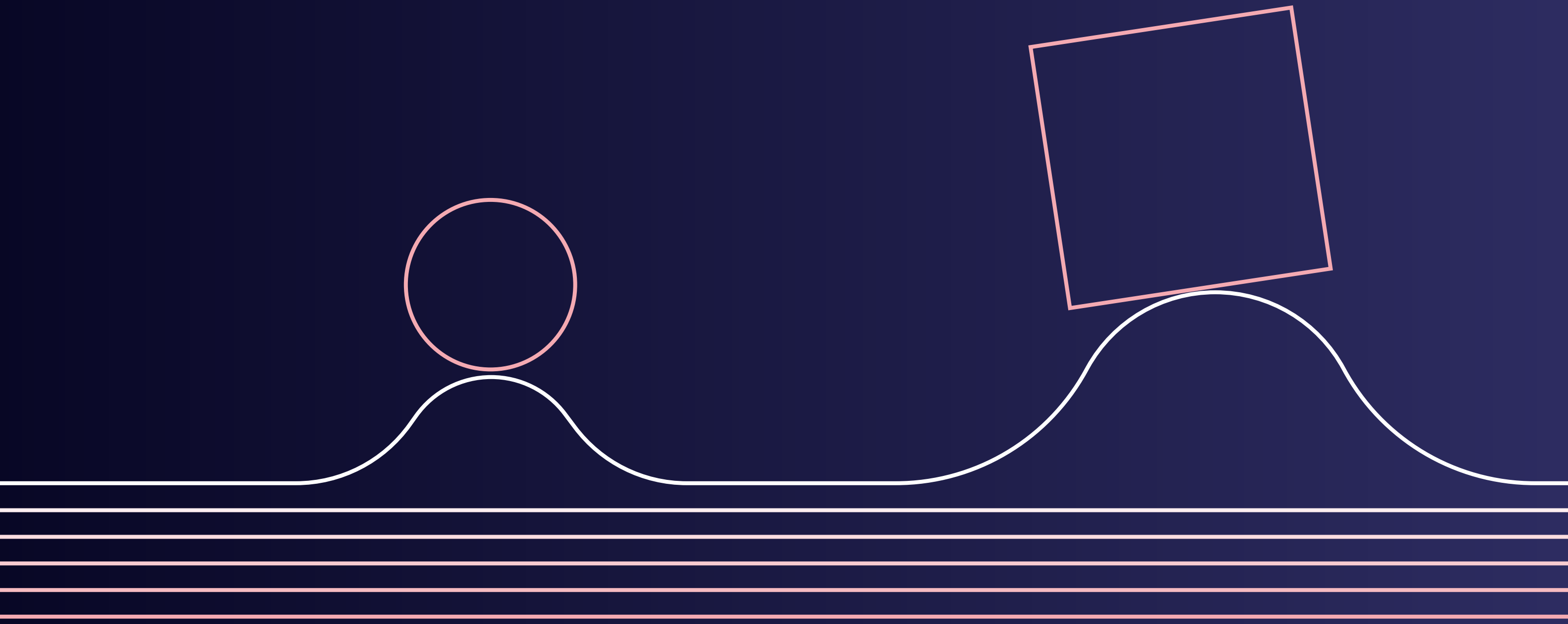
Intermediate learners will evaluate their own personal and institutional concerns over dealing with contested objects, intangible heritages and biases in their collections. They will be presented with evidence and case studies of recent cases of contested digital collections. An information sheet will be available outlining this evidence, so that learners can present and articulate the evidence to senior leaders and trustees so they are able to make informed decisions.

Advanced

Advanced learners will investigate methods of digital community engagement. This will include and link to a later step detailing crowdsourcing with digital collections.

2. Production

How to make the data for a UK digital collection



Our recommendation

To build a UK digital collection, we need to accelerate how data is produced and allow new technologies to automate its enrichment.

What do we mean by this?

Historically, the archiving and cataloguing tasks required to build digital collections have been slow activities, undertaken by specialist teams with a wide range of knowledge and skills. Everything we do builds on the legacy of their work.

However, the rate at which we are currently increasing the size of our digital collections means we would never digitise the majority of the objects we hold today if we wanted to – let alone those yet to be created or acquired. And this slow rate of building digital collections has contributed towards a narrowness of perspective, and sometimes bias in representation, that we should correct.

If we are to build a more comprehensive and inclusive UK digital collection, we need to consider our digital collections like our national collections – as a shared public good – and change our approach to how digital collections are created, establishing them on sound principles and adopting a view across their whole lifecycle from initial creation, through enrichment and to renewal.

Doing this will require us to do four things:

1. **Establish digital collections according to FAIR and CARE principles.** FAIR principles relate to the 'FAIR Guiding Principles for scientific data management and stewardship' published in *Scientific Data* in 2016. The authors intended to provide guidelines to improve the findability, accessibility, interoperability and reuse of digital assets. They are a critical set of ideas from which new projects, data, software, tools and workflows should follow. CARE principles for indigenous data governance make it possible to adopt a respectful, ethical and careful approach to collections based on an acknowledgement of differing understandings of the world we live in.

2. **Create a minimum viable volume of data at the first point of digitisation and data production that we can then build on over time.** By creating essential core records – using well established approaches such as Dublin Core, LIDO and Spectrum metadata standards – we can create a digital collection with the core set of information required to build upon in the future. Developing and adopting an equivalent to the Minimum Information about a Digital Specimen (MIDS) system, which is used in natural science collections and offers a mapping of the quality of digital records, will help us better assess the quality of core datasets.

3. **Open these records to various forms of automated and collective enrichment.** We need to enable and empower both communities and new technologies to help enrich our records. Work to increase the ability to access data, and to transform and enhance datasets is a continuous process. New technologies such as artificial intelligence (AI) have the potential to help us accelerate that enrichment, linking records at speed and scale. Communities – both of specialists and from the public – are also critical participants in the act of enrichment. By giving them the power to add to the depth and breadth of knowledge and tools, we can build highly specialist datasets at lower cost, at greater speed and at greater scale.

4. **Unlock the generative potential of AI to create new data in real time.** AI's ability to work across multiple data records and datasets at speed can allow us to create new tools for managing digital collections data of a kind impossible before. There are many substantial risks to consider first. AI will need to be put at the service of cataloguing and digitisation teams who will need to be trained to take an ethical approach to diverse cultural heritage data. Perhaps most critically, we will need to understand that AI systems often replicate and amplify biases, so digital collections should be built on high-quality training data to correct for these, promoting transparency and accountability. If these risks can be addressed, there is unsurpassed potential.

Case study 1:

Creating lightweight data for digital collections

LIDO (Lightweight Information Describing Objects) design for object metadata documentation aligns with CIDOC's (International Council of Museums International Committee for Documentation) principles, ensuring interoperability in cultural heritage data management. This makes LIDO a practical tool for professionals in the field, facilitating the integration and exchange of museum and heritage information within the International Council for Museums International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM) ecosystem.

As an example, the Virtual Museum of the University of Barcelona houses various non-museum institutions that span diverse cultural categories including movable, immovable and intangible cultural heritage. By adopting LIDO, the University of Barcelona has streamlined its data interoperability, enabling the efficient sharing and integration of its collections within broader cultural heritage networks. LIDO's flexibility and ease of use have been instrumental in this context, especially considering the resource constraints that are often present. This has not only increased the visibility of the university's collections but has also facilitated collaborative research and enhanced public engagement with these cultural assets.

In addition, LIDO's adaptable framework has allowed the University of Barcelona to customise metadata fields to their specific collection needs, helping ensure a detailed and accurate digital representation of unique and specialised items. This attribute of LIDO is particularly beneficial for institutions that house a wide variety of collection types. It underscores LIDO's role in promoting broader access to and democratisation of cultural heritage, proving it to be an effective solution for digital collections management in academia.

Case study 2:

Unpath'd Waters

The Unpath'd Waters TaNC Discovery Project led by Barney Sloane from Historic England is a pioneering maritime heritage initiative. The project recognises the need to foster a symbiotic relationship among major and independent research organisations, commercial bodies and seaside communities.

The project embraces the varied conceptual and geographical nature in which maritime collections are produced to facilitate community participation and collective enrichment.

Unpath'd Waters recognises the role of involving communities at diverse levels of data contribution. Whilst both smaller research organisations and individuals contributed either at metadata level (e.g. avocational divers providing their data, and independent historians contributing to Isle of Man records) or at the data level (e.g. Mary Rose Trust), the most successful exchanges occurred when these communities collaborated with professionals to produce and engage with the data and metadata, as in the case of CITiZAN (intertidal surveys).

The project used advanced digital standards and AI methodologies, such as Low-Shot Learning, to facilitate the enrichment and tagging of collections where minimal data is available for analysis. This combined process helped to combine disparate inventories, enhance controlled vocabularies to produce more accurate categorisations and identify new, missing ones, and thus enhance conservation and preservation efforts of both maritime landscapes and their related historical artefacts.

Data handling and processing were centralised through the Unpath'd Waters Portal, ensuring cohesive aggregation and connections to broader datasets. This portal unites the four UK home nation datasets, and additionally that of the Isle of Man, and links them with wider European aggregators, facilitating research across diverse communities. Although the Unpath'd Waters Portal, as a proof of concept, will be a snapshot, fixed at a point in time, and will not be updated as national inventories are, creating an active link for regular data updates through APIs using OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) would be technically feasible and beneficial. Unpath'd provided a pilot for ADS in the development of multiple user interfaces to the European ARIADNE triple store, which will now be implemented for their own updated UK archaeology search interface, and for the RICHeS digital research service. Additionally, the North Sea simulation of now inundated prehistoric landscapes, while fixed in the Unpath'd Waters context, can absorb new data with appropriate resources, refining and growing the model.

Unpath'd Waters strategically incorporated collections critical for management or policy development alongside the diverse groups engaging with maritime collections and their public. This helped to produce data within pre-existing frameworks and workflows, including sustainability, physical access, ownership and legal rights. This helped ensure that data could be uploaded, assessed for validity and preserved securely, thus promoting future enrichment, reuse and exploration.

Case study 3:

Heritage Connector

Heritage Connector, a TaNC Foundation Project led by John Stack of the Science Museum Group, alongside the V&A Museum, provides an example of how to enrich digital collections and thus increase their access and discoverability. The project integrated cutting-edge AI technologies, such as Natural Language Processing alongside Linked Open Data, to interlink diverse cultural heritage data and produce vast networks of interconnected knowledge or Knowledge Graphs.

The project demonstrated how even limited initial datasets could be expanded into rich, interconnected digital resources through AI techniques such as Named Entity Recognition (NER) and entity linking. Furthermore, by using Linked Open Data approaches, the project offered a scalable and flexible system that can rapidly add, extend and modify data, helping automate these updates and include user contributions. The combination of AI and Linked Open Data helped link entities in real time across datasets and create new connections and data points, which expanded the depth and breadth of the collections. The vast generated networks of data helped expose hidden relationships and narratives that could serve and enrich the public's engagement with digital collections.

The project made evident the importance of utilising a robust data infrastructure that can handle large data volumes, as well as the tools and interfaces to make sense of and engage with it. Furthermore, it helped showcase the benefits of opening records to automated and collective improvements. Heritage Connector leveraged the collection catalogues by making use of core or lightweight (e.g. LIDO) metadata elements such as the item's creator, date, material and historical context to further enrich the knowledge about the object through Wikidata. This helped offer additional contextual information such as related events and broader significant historical context, thus offering external validation and the ability to produce rich narratives. This approach depends on the understanding and accessibility of the collection and sets a precedent for the implementation of advanced data integration methods for digital collections.

Case study 4:

Our Heritage, Our Stories

The TaNC Discovery Project Our Heritage, Our Stories (OHOS), led by Professor Lorna Hughes at the University of Glasgow, has produced workflows to integrate community-generated digital content (CGDC) into larger archival frameworks.

Collaborating with partners across the UK, including the University of Manchester, The National Archives (TNA), Tate, the Digital Preservation Coalition, the National Libraries of Scotland and Wales, and the Public Record Office of Northern Ireland, this project addresses the critical need to preserve and enhance CGDC by overcoming the social and technological barriers that have historically excluded these invaluable cultural assets.

OHOS facilitates CGDC discoverability and the ability to establish new connections between such data and further datasets by leveraging advanced computational methods, including artificial intelligence (AI), to establish novel approaches for CGDC post-custodial management. Central to this effort has been the creation of a public-facing Digital Observatory at TNA, designed to reduce obstacles to CGDC preservation, integration and accessibility.

A key element behind OHOS's methodology is the use of Natural Language Processing (NLP) to identify key entities within free-text CGDC metadata that can enable its integration into broader archival systems and linked-open-data networks, including Knowledge Graphs. This ensures that community archives retain control over their data while enhancing its accessibility and interconnectivity with other materials. The project has also established a Wikibase instance as a proof of concept, which enables community archives to create their own linked data repositories. This practical implementation of post-custodial principles serves as an example for future archives, allowing them to open their materials while maintaining control.

OHOS has developed an innovative human-centric perspective towards AI and computational approaches. By iteratively refining AI workflows with data from diverse communities and of varying levels of complexity, the project has ensured that their tools are developed collaboratively with those who use and generate the data and can be applied to the full range of materials that these communities curate. This unique undertaking has been possible through an extensive process of community engagement, working with the project's partner organisations and a UK-wide network of local, community archives and heritage organisations. The OHOS Digital Observatory makes this data even more approachable by offering a broad suite of tools for searching and exploring CGDC alongside and linked to collections held by TNA. This public-facing interface makes this content widely accessible and flexible, remixable in ways that cater to both expert and non-expert users through various search interfaces and visualisations.

The developments from OHOS showcase the advantage of open-source, decentralised data management tools. OHOS outputs empower communities by offering a suite of solutions to ensure they maintain control and custodianship of their data whilst promoting sustainability, scalability and active participation. By fostering sustainable connections between communities and archival institutions, OHOS establishes a precedent to safeguard the preservation and accessibility of CGDC.

Case study 5:

Congruence Engine

The Congruence Engine TaNC Discovery Project, led by the Science Museum's Dr Tim Boon, aims to enable users to explore data from diverse heritage items, including museum objects, archive documents, pictures, films and buildings. To achieve this, it has focused on creating new data from historical sources and linking it to collections metadata.

They have produced a sophisticated digital ecosystem where diverse heritage data is interconnected, allowing for rich multifaceted analysis and engagement. Congruence Engine uses an action research methodology responsive to its community's needs and interests. Early insights demonstrated that practical engagement in a 'social machine' of contributory data linkage requires an emphasis on the human experience in the past as a point of connection between object and archival records.

One example is its work with oral histories; the project approach to data linkage and interoperability has created multilayered annotations of records and metadata to enable users to explore connections according to interest. At a time of rapid development in Large Language Model technologies, it has prioritised defining a conceptual process and evaluating current best solutions while investigating ethical implications. It has experimented with training AI systems on diverse cultural heritage data.

Their workflow involves speech-to-text transcription (Whisper), Named Entity Recognition (spaCy, with training on specialist entity terms for e.g. objects), automated reconciliation to Wikidata Qcodes (spaCy-entity fishing), entity relation extraction (Llama to Neo4j graph database), key term identification and annotation (KeyBERT, surprising phrase detection), and topic modelling (BERTopic), with bespoke visualisations in development using Visx and D3.js. The entity/term identification is indexed to an interim informal knowledge base of aggregated taxonomies and gazetteers (of people, locations and roles, objects, materials, records and activity types). This is improved by human involvement at each stage.

The project has shown the digital infrastructures need to be flexible and adaptive to enable contributions from diverse stakeholders with specific challenges and obstacles. Technical mechanisms should mediate data resources and pipelines, whilst respecting and adapting to local systems and ways of working.

Collaborating with local partners, Congruence Engine has accessed oral history data alongside other collections and primary source materials along with their metadata. The project aims to understand how work with innovative technology for data linkage may become self-initiating and sustaining at a grassroots level. Technical capacity and training are important for enabling the adoption of new technologies, but so too are the dynamics of local control over access and sharing. By inviting confident ownership of the tools and processes, the project has refined the specification of a social machine that can generate widespread participation in transformational digital work on industrial heritage and enable productive dialogue around openness and interoperability.

Sample training module descriptions

Relevant module 1: Metadata basics

This training module provides users with an overview of what metadata is and how it relates to the management, curation and discovery of collections materials.

At an introductory level, users will learn how metadata relates to the object (data) it describes, how it can be structured and stored, and why it is valuable. They will recognise the relationship between traditional forms of cataloguing (hand-lists, index cards) and digital methods as an evolution of methods. Examples will be given of how simple metadata might appear in different formats, and some of the ways different materials might be described using different approaches to metadata.

Taking a simple 'describe this object' task, users will be introduced to ideas such as structured and unstructured metadata, faceting and how a metadata description can fundamentally change our perception of an object.

This will be used to demonstrate why it is valuable to have a consistent approach to metadata collection, and introduce schemas as a means of making metadata as regular and controlled as possible.

Relevant module 2: Data wrangling

In this module, users will learn about messy data and how this can be both improved or avoided. Users will be introduced to some of the typical and often simple ways in which data can become messy or 'noisy', and why this causes so much trouble.

By showing examples of a dataset that has been cleaned, it will highlight the differences between a clean, well-curated dataset and a messier one.

Users will then be walked through some of the tools used to do the work and how they have been applied in cleaning the data. This will be done in two stages, with worked examples of manually cleaning data at a very small scale in a spreadsheet and cleaning it using a more specialised tool such as OpenRefine.

3. Skills

The skills we have and the skills we need for a UK digital collection

Our recommendation

To build a UK digital collection, we need to build upon the existing deep skills in data creation already within digital collections organisations and create a scalable, sustainable approach to accessing the advanced technology skills we too often lack.

What do we mean by this?

We have not always thought that the specialists who work on our collections are data creators. But as we work towards creating a unified UK digital collection, we must evolve our understanding of the critical role these individuals play: they are the basis of all data creation – their knowledge is the bedrock from which to expand, enhance and organise our collections.

The skills these specialists have are already scarce — the outcome of long-term underinvestment — and so capacity is already stretched. To build digital collections that meet our future need, these specialists' capabilities will need to be augmented by advanced technology skills. If we are to achieve our ambitions to create a UK digital collection, we need to both advance the capabilities of these data creators and change how we access advanced technology skills.

Doing this will require us to do three things:

1. **Broaden the skills of data creators and empower them as data owners.** We need to recognise that our cataloguers and the teams who create collections data are critical resources. This means investing in their understanding of not just data creation but of security, preservation, utilisation and standardisation. They must have an understanding of, and take responsibility for, the full lifecycle of the digital collections they build.

2. **Invest in national-scale technological capacity for digital collections.** Building a national infrastructure will depend upon building an array of technical research teams. Learning from the approach of CoStar and DiSSCo, it is possible that technical R&D, core systems, and specialist digitisation and digitalisation teams will all need to be established. We need to consider how this capability is distributed both geographically and thematically to ensure it best supports the needs of the many institutions that will need to draw upon it.

3. **Build large-scale partnerships with industry and academia to be able to work beyond our skills capacity and capability.** The potential technology changes of the next decades may be outside a reasonable boundary for what skills can sit within our infrastructure and its constituent institutions. Constructing partnerships that give sustained and meaningful access to the most advanced and highest potential new forms of technology and computation requires national-scale coordination and shared commitment.

Case study:

Digital Preservation Coalition Competency Framework

As digital collections grow, the need for creating robust institutions becomes paramount. The Digital Preservation Coalition Competency Framework offers a comprehensive approach to equipping staff with the necessary skills to address these challenges effectively. The framework can assist in the development of role descriptors and planning professional development.

The framework offers five high-level competency areas:

1. **Governance, Resourcing and Management**, which covers policy development and risk management to ensure that digital preservation activities are well supported and aligned with organisational goals.

2. **Communications Advocacy**, ensuring that there is effective communication, collaboration and teamwork among stakeholders, including user analysis and engagement.

3. **Information Technology Skills**, to ensure digital preservation practitioners have adequate general skills to select, use and manage technology effectively.

4. **Legal and Social Responsibilities**, covering legal compliance, ethical considerations, inclusion and diversity, and environmental impact.

5. **Digital Preservation Domain Specific Skills** that are directly related to the understanding of metadata standards, information management principles and approaches to preservation, digital preservation standards and models, and managing access to digital content.

Organisations can use this framework to conduct skills audits, identify gaps and plan necessary training in the field of digital preservation.

4. Reuse and rights management

How to create sharable collections data

Our recommendation

To build a UK digital collection, we need to adopt a coherent and consistent approach to data and rights management. As part of this, we need to prioritise making new data in open ways that is as reusable as possible. But we must recognise the vast range of different data that cannot be made available in this form and maximise ways to share it, whilst managing its constraints and recognising its unique cultural, intellectual and ethical properties.

What do we mean by this?

The development of digital collections has been held back by differing approaches to how we think about intellectual property (IP) and rights, and the different IP frameworks that institutions use, based on the individual risks and opportunities they face.

For us to grow an interoperable UK digital collection at speed and with quality, more open forms of data will need to be prioritised for future digital collections. More open data can help accelerate data production and open doors for a wide range of engagement.

However, for many institutions, such as in the visual arts, screen or archive sectors, vast volumes of data exist in ways that cannot be shared on an open basis, whether because it is legacy, orphan, third party, community-held, commercial or in copyright. A UK digital collection will be poorer without this material, and so it is critical that proper protections are put in place and the risks to this data are properly understood, so we can include it.

To manage the balance of risk and opportunity, we need to consider how best to take an integrated approach to reuse and rights management. This needs us to do four things:

1. **New data should be created and made available on as open a basis as possible.** The Creative Commons licensing framework provides great models to follow, as do Open Government Licences or equivalents where they can be sensibly adopted.
2. **We must maximise the ability for this data to be reused.** The data's true value comes from its interoperability and dynamic reusability between and across collections to find unexpected adjacencies from which new knowledge and understanding about the collections can be generated.
3. **Data that cannot be released on an open basis needs to be understood and protected.** Whilst we build a more generally open collection, the kinds of data that cannot be made open because of a range of commercial, rights or other imperatives must be understood, and planning and investment made into how to protect its value.
4. **The use of open-standards technologies in sharing data is critical.** Open-standards technologies should be adopted across the sector to help interoperability. These include the use of IIIF (International Image Interoperability Framework) technologies for images, audio, video and annotations, and methods such as the use of Persistent Identifiers (PIDs) to identify data and more. Open technologies and open data formats that allow for interoperability and scalable reuse are critical enablers of a UK digital collection, whether the data they hold is commercial, open licensed or in the public domain and so copyright free.

Case study 1:

The use of IIIF

The role of the International Image Interoperability Framework (IIIF) was investigated in the TaNC Foundation Project Practical Applications of IIIF as a Building Block Towards a Digital National Collection, led by Joseph Padfield from the National Gallery. This project explored the transformative role of IIIF in connecting digitised collections from separate organisations.

The project investigated how IIIF can be used to integrate and present digital resources to the wide range of audiences that engage with digital collections. This included lowering barriers to uptake and creating new opportunities for digital reinterpretation, scientific examination and display.

Efficient methods of using IIIF to build collaborative online resources was a key focus, demonstrating the potential for dynamic, cross-collection searching. The project highlighted how consistent advocacy for, use of and ongoing support for this type of established and mature interoperable open standard is one of the most effective ways of connecting, at a national level, the vast number of collections and sources of digital resources in an economically feasible and sustainable manner.

The project findings show that IIIF is a mature, established system and operates beyond a purely presentational layer. In addition to images, IIIF supports audio-visual content, annotation and other types of content, with 3D support currently in development. This extensibility means that IIIF can accommodate a wide range of digital content, offering further opportunities and benefits as new features are developed. Adoption of this shared standard should ultimately drive down delivery costs while providing a standardised base for a user experience that has the opportunity and potential to be built upon by various initiatives.

Case study 2:

The importance of PIDs

The TaNC Foundation Project Persistent Identifiers as Independent Research Organisation Infrastructure, led by Rachael Kotarski of the British Library, looked at the potential role and value of Persistent Identifiers (PIDs) in a future digital collection's infrastructure. It was identified that PIDs will offer a long-lasting reference to both digitised and digital resources.

Heritage organisations across the UK house many millions of physical and digital objects; having the ability to uniquely identify these objects is paramount for their discovery, use and curation. For example, accession numbers are a key component in all collection and library management systems, but these only cover selected objects within an individual collection. By contrast, PIDs can provide a long-lasting, reusable and clickable link to any given digital resource or object. Use of PIDs can also help organisations, including the UKRI, alongside cultural and heritage organisations, to produce metrics from citations and the use of their content. The project highlighted that organisations should support the wider use of PIDs across the wide range of digital assets in their collections, environments, specimens and related items in a dependable ecosystem that can be reliably accessed and referenced over time.

However, the project noted that PIDs are poorly understood across the heritage sector. To address this, the team engaged in developing skills for the sector by gathering evidence through a mixture of workshops, surveys, desk research and case studies to develop a toolkit for using PIDs in heritage organisations in the UK.

Case study 3:

Open Data Institute – Open Data Maturity Model

The Open Data Institute's Open Data Maturity Model (ODMM) provides a comprehensive framework for how to foster an organisational culture that prioritises open data sharing. The framework is based on five core themes: data management processes, knowledge and skills, customer support and engagement, investment and financial performance, and strategic oversight.

The ODMM addresses the foundations required for effective data management. This includes the establishment of data release processes, the development of technical standards and the governance of data, including the management of sensitive information. It advocates for digitisation and open sharing under FAIR principles, ensuring that data is not only made available in a structured manner but also regularly archived, enhancing its availability and utility for open use. Data management processes are crucial to streamline and update datasets under open licences such as Creative Commons or the Open Government Licences. Ensuring that processes are standardised across organisations can help support the scalability and efficiency of data handling. However, organisations must ensure that they have the internal capabilities to reuse data and are proficient in open-data practices. Knowledge and skills training can help reduce dependency on external expertise, which is vital for managing open data effectively. Furthermore, these skills relate not only to the management of the data but also to creating robust engagement systems with data users. This includes the production of documentation, establishing support channels and targeting more user-centric and responsive approaches to the needs of their communities.

Organisations will need to have a clear strategic oversight for data sharing and reuse. This includes the strategic approach to making informed decisions about which datasets to open up and under what conditions, as well as protecting sensitive data that cannot be openly released. This strategic oversight also needs to consider how the organisation can better support publication and reuse of open data, as well as align those processes into financial resources that offer the highest return on investment, and/or identify potential savings associated with open data.

Case study 4:

Use of open licensing frameworks

Creative Commons is an international non-profit organisation that provides the legal and technical infrastructure to support the sharing and use of creative work, including the set of legal tools or frameworks provided through the Creative Commons licences. This licensing framework helps creators to identify the wide range of conditions in which others can use their work.

On the one hand, licences can be more restrictive, as in the case of Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licensing, which prevents others from changing or commercialising the work. On the other, more flexible licensing, such as Attribution (CC BY) licensing, can enable others to distribute, remix, adapt and build upon the work, even commercially, as long as they reference the original creation. Finally, CC0 is a Public Domain Dedication that enables sharing without any attribution. Organisations need to be familiar with the diverse licensing frameworks, regardless of their level of openness, and identify the practical, legal and ethical applications. These can depend on the nature of the digital output, calls for identifying sensitive information, or third-party rights that might require special licensing conditions. This includes identifying the legal position of rights holders concerning all material published online, whether or not they are openly licensed.

Many UK organisations have opened access to their resources by implementing open licences, including Open Government Licences, CC BY and CC0 (public domain) licences. These have resulted in the ability to enable further groups, individuals and organisations to generate innovative ways of working, displaying, and engaging with metadata and digital objects. For example, the University of Edinburgh collaborated with Wikimedia UK to integrate a historical dataset on Scottish witchcraft into Wikidata, making it accessible and reusable under the CC0 license. The National Library of Wales employed a 'Wikimedian-in-residence' to engage with local and global audiences through Wikimedia platforms. Finally, the project A Street Near You, developed by James Morley, saw the Imperial War Museum (IWM) provide access to its collections, including information on millions of people who served in the First World War, not just those who died. By sharing their metadata and other digital objects such as images under an Open Government Licence (equivalent to a CC BY 4.0 license), the IWM enabled broader and more global use of their collections.

The National Lottery Heritage Fund (The Heritage Fund, HF) is the UK's largest funder of heritage and provides guidelines to help ensure that UK digital heritage is accessible and managed sustainably. This includes the use of open licences, which, along with availability and accessibility, are a requirement of its digital grant-making. Through its Digital Skills for Heritage initiative, HF has developed guidance to support knowledge and skills around copyright, data protection and open licences in relation to digitisation and digital collections. Their guidance is aimed at enhancing the capability of heritage organisations to manage their digital collections in line with FAIR data standards. Finally, HF's strategy commits to funding digital heritage resources that are 'freely and openly accessible to the public, now and in the future'.

Sample training module descriptions

Relevant module 1: Open vs closed

This module introduces users to the terminology commonly used in research institutions around open data, open research and open source. Learners will evaluate whether open tools are relevant and appropriate for their own organisation. Three levels are included:

Beginners

Beginners will learn that Open Access describes a cluster of initiatives intended to make research available. They will understand the different terminologies between Open Access, open source and open research, and why this open movement has become important in the sector. They will learn some of the push and pull factors contributing to Open Access, including UK government policy, funder compliance and societal benefit.

Intermediate

Intermediate learners will consider the technical skills needed to make more collections openly available and reflect on the availability of those skills within their organisation. They will also consider software as a service (SaaS) solutions and identify those in use in their own institution. A case study will illustrate the risks of bespoke builds and the sustainability benefits of open source or SaaS software.

Advanced

Advanced learners will consider how open-source and proprietary software can be used and are connected within their own or other GLAM organisations. They will be challenged to consider how their organisation could be more open, and what resources and skills would be needed to enable this.

Relevant module 2: Copyright and licensing

Learners will understand the basic concepts of intellectual property rights – particularly copyright and database rights – including terminologies such as works, usage and exceptions. Users will be encouraged to analyse and apply some of these concepts to their own collections, taking a risk management approach. Three levels are included:

Beginners

Beginners will learn about different intellectual property rights and how they apply in the UK (including differences across the four nations). They will learn what kinds of collections items and assets are subject to copyright, and the different usages of copyrighted works within their organisation and by their organisation's users (such as commercial use and educational use). They will also investigate and consider the different licences their organisation currently uses in relation to its digital assets.

Intermediate

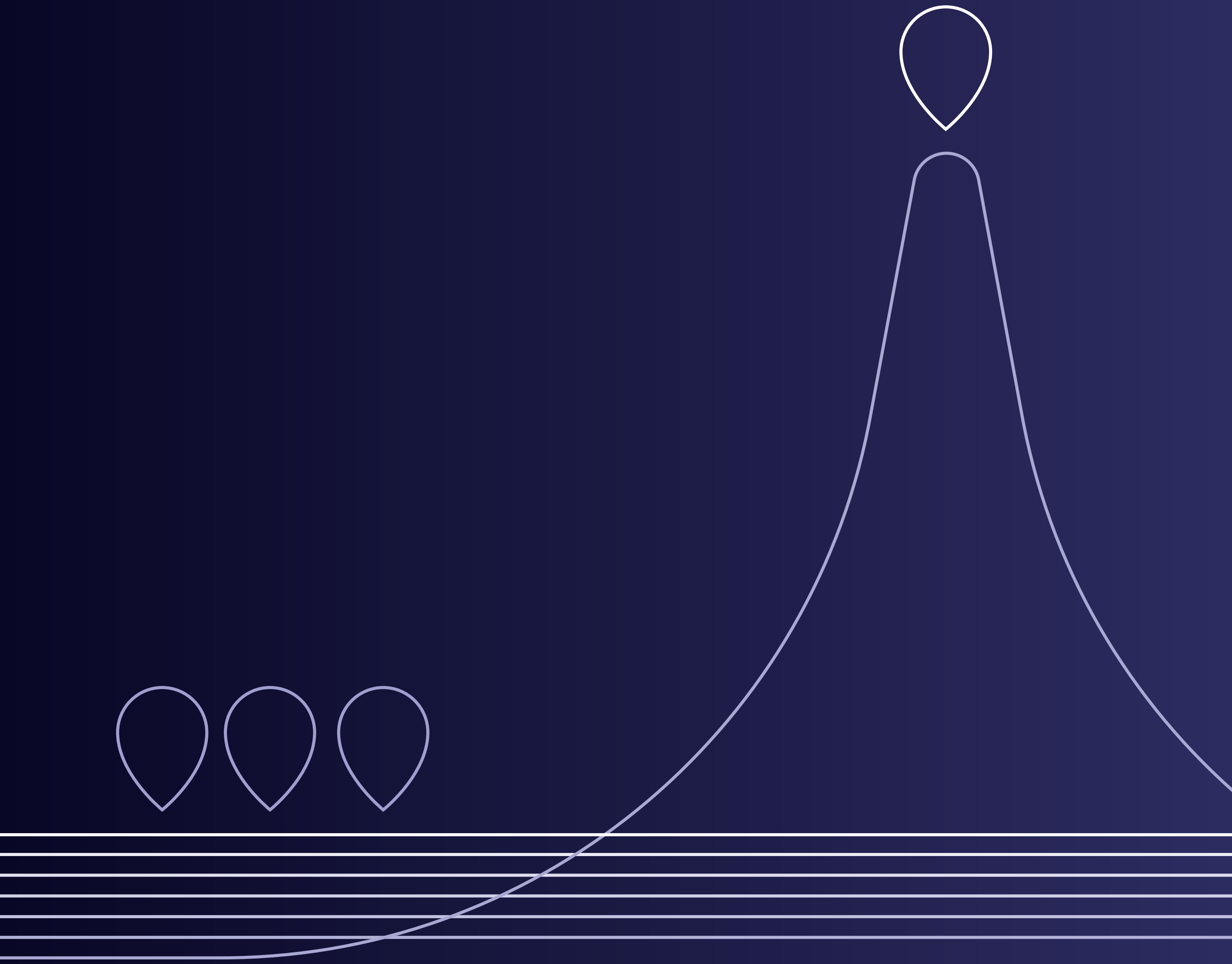
Intermediate learners will look at copyright exceptions through examples of more complex scenarios from a library, an archive and a museum. They will be encouraged to undertake a risk assessment that balances information relating to the works, usage and licences.

Advanced

Advanced learners will consider the differences between different legal regimes, and the changing nature of case law in the UK (based on recent court cases). Learners will be encouraged to review the implications for their own collections and within the context of their own organisations. A case study will show the impact of a recent court case on making digital collections openly available, whilst retaining the ability to generate income via an image library. A fact sheet will detail some key considerations relating to recent case law for senior leaders and trustees, highlighting how openness and income generation can still co-exist.

5. Access and engagement

How to make data accessible to everyone



Our recommendation

To build a UK digital collection, the data and the platforms and infrastructure through which they are used needs to be accessible and meaningful to everyone. Data should be easy to find. It should be made accessible through exploration and navigation tools and interfaces that enable engagement with it. And open data should be machine readable via direct data interfaces such as APIs.

What do we mean by this?

Everyone should be able to access the UK's digital collections. But unlocking the value from that access requires a combined series of steps.

For most, access will likely be via standard Web-based platforms and systems, such as websites, mobile applications or desktop applications. Audiences will have diverse needs when they first approach digital collections, and platform design will need to accommodate these.

When data is being made available to audiences, we need to do five things to make its access as useful as possible.

1. **Data should be discoverable.** Search engines, through simple or advanced text entry, are the primary way that people have been accessing digital collections, along with other methods such as web and mobile apps. New applications built to access digital collections should make use of the multiple forms in which both users and machines discover data.

2. **Data should be equitably accessible and fit for use.** A user-centred approach to data should be taken that recognises the diverse needs of different users, including neurodivergent and less able users, and communities including non-Western groups. We should keep in mind their requirements and expectations. Making access truly equitable is a challenge to be overcome.

3. **Generous interfaces should be provided that facilitate engagement by helping users make sense of, engage with, understand, compare and analyse data**, helping users identify insights at point of contact that exponentially multiply the value of the resources.

4. **All data that is made available to individual users through digital applications should be provided in machine-readable formats**, accompanied with the computational tools and interfaces to help produce value from the knowledge held within collections.

5. **Data, information and objects should be available and persist.** Users need to be able to trust that the resources they use will remain available in the long term. These should be stored in trusted long-term data systems or repositories that come with published sustainability plans so it can be used without concern.

Case study 1:

The Sloane Lab

The TaNC Discovery Project
The Sloane Lab, led by Professor Julianne Nyhan from University College London and TU Darmstadt, has focused on re-establishing the currently broken links between the collection amassed by Sir Hans Sloane during the early-modern period.

The Hans Sloane Collection is the founding collection of the original British Museum. It is currently split across the British Museum, the Natural History Museum and the British Library, with the digital information about the collection residing in various systems. The project has focused on digitally reunifying the collection and its records across these institutions, to mend the broken links between the past and the present of the UK's founding cultural heritage collection.

At the core of the unification processes resides the use of standard semantics for the cultural heritage domain derived from the CIDOC CRM ontology that ensures sufficient handling of complex data stratification and semantic interoperability. Automatic and semi-automatic data aggregation methods, along with long-legged data mobilisation processes, deliver a wealth of data to the project in various formats and serialisations. These are converted and aligned to a unified data model, facilitating cross-collection search, digitally augmented exploration and reuse of the Sloane collection. The data model extends CIDOC CRM with semantics that handles uncertainty, multivocality and modality of the collection (e.g. conflicting data from different records about the same object), data absences (i.e. gaps in the records), the difficulty of classifying objects and the fact that some of the objects described in the historical catalogues are now lost. Moreover, artificial intelligence (AI) and Natural Language Processing (NLP) are employed for the semantic enrichment and linking of data with references to entities of interest such as people, places, object types, techniques, materials and others. All these are realised under an advanced interactive environment known as the Sloane Lab Knowledge Base, which facilitates resourceful query, visualisation and fact-finding using Knowledge-Graph technology.

In addition, a participatory approach has been integral to the project. Working with diverse experts and interested communities, the project has sought to understand the many questions individuals and communities wish to ask of the project. New knowledge, tools and workflows have been developed and co-created through active collaboration with partners and a series of community fellowships, facilitating new kinds of creative engagements with the data. The project offers both human-centred access via an interface as well as programmatic or computational access through their Knowledge Base, as well as a SPARQL endpoint to query the complexity behind the semantic entities.

The project has developed the Sloane Lab Digital Platform, which offers researchers, curators and the public new opportunities to search, explore, and critically and creatively use and reuse digital cultural heritage. Their platform can also help users to navigate and interact with the data intuitively through the use of a Generous Interface. This approach enables the integration of sophisticated algorithms and filtering options to find data quickly and efficiently, as well as tools for diverse levels of technical proficiency and knowledge.

Case study 2:

Collections as Data – GLAM data labs

The perspective of Collections as Data focuses on making digital collections from galleries, libraries, archives and museums (GLAM) accessible for computational use. This involves providing clear licensing, detailed metadata and documentation, structured datasets, and utilising public platforms and APIs to facilitate these collections' reuse, analysis and exploration.

This initiative is championed by the International GLAM Labs Community, which has been empowering the GLAM sector to adapt to digital transformation by promoting tools like Jupyter Notebooks and data accessibility to enhance the utility of cultural heritage in the digital realm. Their work is important in modernising collections management, encouraging innovative research and expanding public engagement with heritage resources.

This perspective helps advance digital literacy and computational analysis within the cultural heritage sector. Some key participating organisations include the National Library of Scotland, the Australian GLAM Workbench, the Library of Congress, the British Library, Biblioteca Virtual Miguel de Cervantes, the National Library of Estonia and the Austrian National Library. Additional contributors are Det Kgl. Bibliotek (The Royal Danish Library) and KBR (The Royal Library of Belgium).

The GLAM Labs' approach greatly improves heritage management by enabling advanced computational analysis, fostering global collaboration and enhancing public access to digital collections. This results in enriched cultural understanding, research opportunities and a democratised approach to heritage preservation.

Case study 3:

Common European Data Space for Cultural Heritage

Common European data spaces are like secure online marketplaces where organisations can share data. They can help make more data available for use and for the development of new products and services. The European Commission has produced 14 different data spaces, including for manufacturing, health, media and cultural heritage.

The Common European Data Space for Cultural Heritage builds on the work of the Europeana Digital Service Infrastructure and the Europeana Strategy 2020–2025 to provide interoperable access to cultural heritage data. The rollout of the data space for cultural heritage is still ongoing, with data spaces created already within Europeana through Europeana PRO, AI4Europeana and DE-BIAS, as well as other multimodal dataspaces such as EUreka3D and 5Dculture.

The data space offers a centralised approach which ensures that all digital cultural assets are available via Europeana's platform, thus simplifying access for users, regardless of their geographic or institutional boundaries for their partners. This deployment is carried out through the development and operation of the data-space infrastructure, the integration of high-quality data, capacity building and fostering reuse amongst communities working with cultural heritage, and by providing examples of use and participation and expanding pan-European perspective and understandings. In addition to this, the Europeana Publishing Framework offers clear guidelines for implementing metadata standards and facilitates metadata enrichment by both manual contributions through crowdsourcing campaigns and automated tools.

The common European data space for cultural heritage also provides the Metis Sandbox, co-created with the Europeana Common Culture project, which enables data providers to test and validate their data on their own before it is delivered to Europeana. The system can provide feedback offering actionable insight to improve the quality of their data. In addition, Europeana has explored decentralised data aggregation models, which support the vision of a more connected and seamless data environment where changes and updates can be made efficiently and propagated without centralised bottlenecks. This infrastructure has been important for maintaining high-quality, interoperable data that is essential for the effective functioning of search and retrieval systems within Europeana.

All digital data made available through Europeana is provided in machine-readable formats. This standardisation supports accessibility for diverse uses, including academic research, education and creative projects. The infrastructure of the common European data space has been designed for long-term data sustainability, including the use of Persistent Identifiers (PIDs) and robust digital preservation strategies.

Case study 4:

Locating a National Collection

Locating a National Collection was a TaNC Foundation Project led by Dr Gethin Rees from the British Library in collaboration with the University of Exeter, the National Trust, Historic Royal Palaces, Historic England, English Heritage, Historic Environment Scotland and the Portable Antiquities Scheme. This project was a pioneering initiative that helped connect various geographic information with digital heritage records.

Through a strategic approach to the integration of geographic data, the development of web-based tools and the facilitation of user-centred interfaces, the project helped establish a process for how digital collections can meet the needs of modern audiences and researchers. These approaches help users explore collections through spatial relationships that add a valuable layer of context that was previously difficult to access. Furthermore, through tools such as Locolligo and Peripleo, they provide a workflow to transform traditional spreadsheet data into interactive maps, thus helping the community access and discover further collections. Locating a National Collection focused on delivering 'generous interfaces' that are designed to maximise engagement with the data and are crucial to making data more understandable to a wide range of audiences.

The project established some key technological innovations that facilitate the reuse and sharing of data across platforms and applications by transforming data into machine-readable formats such as JSON-LD Linked Places format. This approach allows for both people and computers, and can be processed automatically by computers, reducing time and effort.

Space, place and diverse depictions of geography are intimately connected with how identity is built. Therefore, providing novel interactions that help understand these relationships can help build on the public's values to motivate them to engage with the bread of digital collections. Geographic and spatial information can serve as a critical linking element across collections, helping to add meaning to digital collections and historical events and thus deepening engagement with them. Locating a National Collection offered a new way of adding meaning to datasets and collections whilst offering a richer and more layered sense-making process for both the public and researchers.

Sample training module descriptions

Relevant module 1: Introducing the application programming interface (API)

This module will help users understand what an API is, how it works, how to create one, and the benefits that having one can bring to users, researchers and organisations.

Three levels are included:

Beginners

Beginners will learn what an API is, based on examples from the GLAM sector and other commercial uses. A case study from audience research will demonstrate that some researchers want to analyse data at scale and access data through APIs.

Intermediate

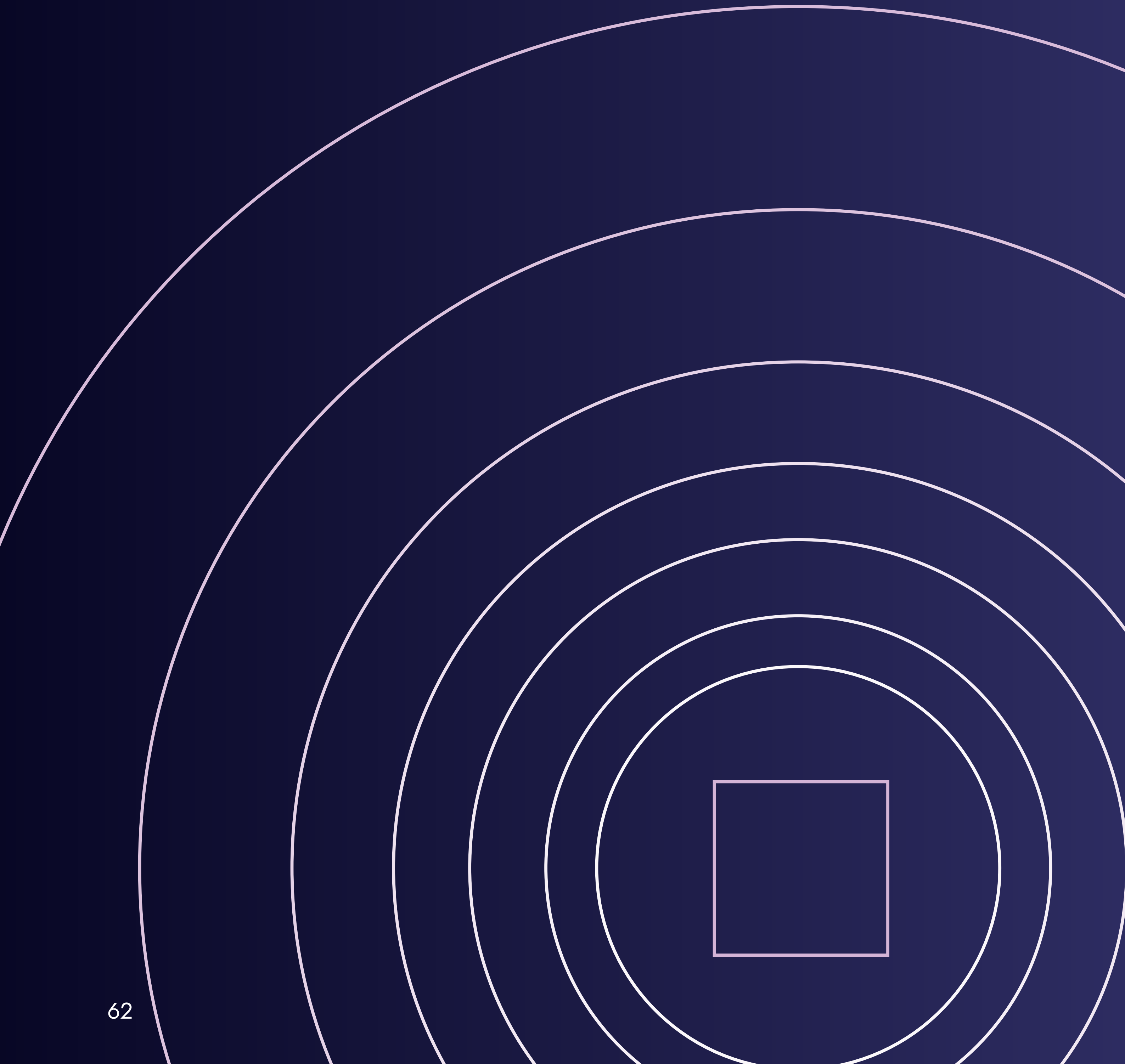
Intermediate learners will be able to identify institutions with data APIs, including whether their own institution has one.

Advanced

Advanced learners will look at the technologies and skills needed to make their organisational data available via an API.

6. Security

How to protect our collections from harm



Our recommendation

To build a UK digital collection, we need to ensure that the data, and the collections, are secure. Security means that the data, the technology infrastructure and the way data is being used are all protected by common standards and legislation, as well as good practice.

What do we mean by this?

Digital collections should be for everyone. But that will only happen if the data, technology and platforms, as well as the behaviours of the people who support them, are secure, and if there are high standards of information security applied to how data is created, managed and accessed.

Collections data has emerged in high-profile examples in the US and UK as a focus for hackers, with major service disruptions. However, this does not mean that we should shrink from digitising and sharing our collections for fear of cyberattacks. Rather, world-class cybersecurity should be in place to protect them.

There are four things which we need to do to make our digital collections as secure as possible:

1. **Digital collections data should be securely managed in isolated environments, with secure APIs facilitating safe, large-scale data access and extraction.** Data security must be prioritised with strong encryption and continuous compliance audits. Ensuring that software and systems are licensed and updated is critical.

2. **Tools to access and engage with data should operate with the highest-quality security standards.** Funding should be made available so that institutions with digital collections can meet Cyber Essentials cybersecurity certification at a minimum, whilst higher-level standards should be considered by medium and large institutions.

3. **Institutions and organisations should set clear policies on who can create, edit and publish data**, so as to manage internal threats. The ability to extract and share files directly with partners under both commercial and non-commercial licences from institutional systems should be restrained. All programmatic access should be instead via APIs, so that rigorous data security is in place throughout. General Data Protection Regulation (GDPR) and other information security training should be up to date for all of those who work on digital collections.

4. **Institutions should be prepared.** Cybersecurity attacks will happen. Data management plans should accompany every digital collection and include security and continuity plans. Collections data should become a regular part of institutional audit processes. Stress-testing vulnerabilities on a regular basis is a critical operating requirement for institutions with digital collections.

Case study 1:

The British Library

The British Library experienced a significant and high-profile cyberattack in October 2023, which disrupted its operations and highlighted many vulnerabilities in its systems infrastructure. This case study is based on the review of the incident published by the library.

During the incident, some 600GB of files were illegally extracted, which included data from users and staff. Major systems, including servers and applications, were destroyed, hindering access and the recovery of the systems. This resulted in the inability to access the data for both staff and the public for an extended period, severely affecting the library's ability to serve its educational and research functions. This lack of access further cascaded down to other government investments that relied on this data, such as the National Lottery Heritage Fund-supported Unlocking our Sound Heritage. It also temporarily halted digitisation activity, thus hindering partnership projects and commercial income.

The October 2023 cyberattack was successful for many reasons, including:

- **Outdated security measures** – the British Library's security infrastructure was outdated, poorly designed from a security perspective and included a large number of legacy systems across its IT infrastructure. The library also relied on multiple service providers, which complicated the management of the IT infrastructures and security protocols.
- **Insufficient staff training** – there was a lack of adequate cybersecurity awareness and training across the library and supplier teams that prevented employees from recognising phishing attempts or from undertaking secure data practices.
- **Lack of regular security audits and weak access controls** – even though the library had met Cyber Essentials standards before, it did not routinely conduct regular and comprehensive security audits and penetration tests and failed to comprehensively establish the most up-to-date and secure access-control and identity-management protocols across the digital environment.

Moving forward, the British Library's commitment to upgrading its security infrastructure showcases the need to create secure, isolated environments for data management and stronger identity management (particularly for privileged access) and encryption methods. It is important to acknowledge that people need to be well equipped alongside the systems and need to have the capability and understanding of standards of data creation, management and publishing. Importantly, all staff need regularly updated training to ensure their cybersecurity knowledge and practice keep pace with the continuously changing threat landscape.

Case study 2:

Historic England's digital strategy

Historic England, the public body dedicated to the stewardship of England's historic environment, has begun the process of becoming a data institution. With guidance from the Open Data Institute (ODI), the organisation has redefined its approach to managing and using data, with the purpose of enhancing its impact on public, educational and charitable goals through responsible data stewardship.

This transformation made data stewardship a key responsibility for Historic England. Data stewardship involves more than just the technical handling of data; it also considers the ethical implications of how data is collected, used and shared. Furthermore, as a data institution, Historic England is expected to manage the data not just for its own organisation but also to become a public asset that can provide value to the wider community, including identifying sustainable data management strategies.

Working alongside the ODI, Historic England was able to engage with targeted workshops and the deployment of specialised tools, which aimed to expand the understanding and implementation of responsible data practices. Firstly, the Data Ethics Canvas played a pivotal role in enabling Historic England to systematically address ethical considerations. Secondly, the Sustainable Data Access Workbook presented new ways for assessing and enhancing the revenue models and thus unlocking the value of their data.

Historic England's approach focused on elevating knowledge and quality standards through targeted workshops and specialised tools that enhanced their ability as data stewards. Running workshops alongside the ODI was pivotal in enhancing their preparedness, thus highlighting the importance of building resilient data infrastructures that protect historical data from cyberthreats and other risks, including ethical risks.

Sample training module descriptions

Relevant module 1: Security and your digital collection

This module encourages users to learn about the importance of digital security in a world of increasing cyberthreats and attacks, including key risks and some potential mitigations.

Two levels are included:

Beginners

Beginners will learn about recent cyberattacks on cultural, civic and educational institutions, including the evidence of attack, the initial response, and the impact such attacks had on staff and users. Users will consider what individuals can do to mitigate risks, such as undertaking information security training, thinking about personal IT use and knowing their organisation's IT policies.

Intermediate

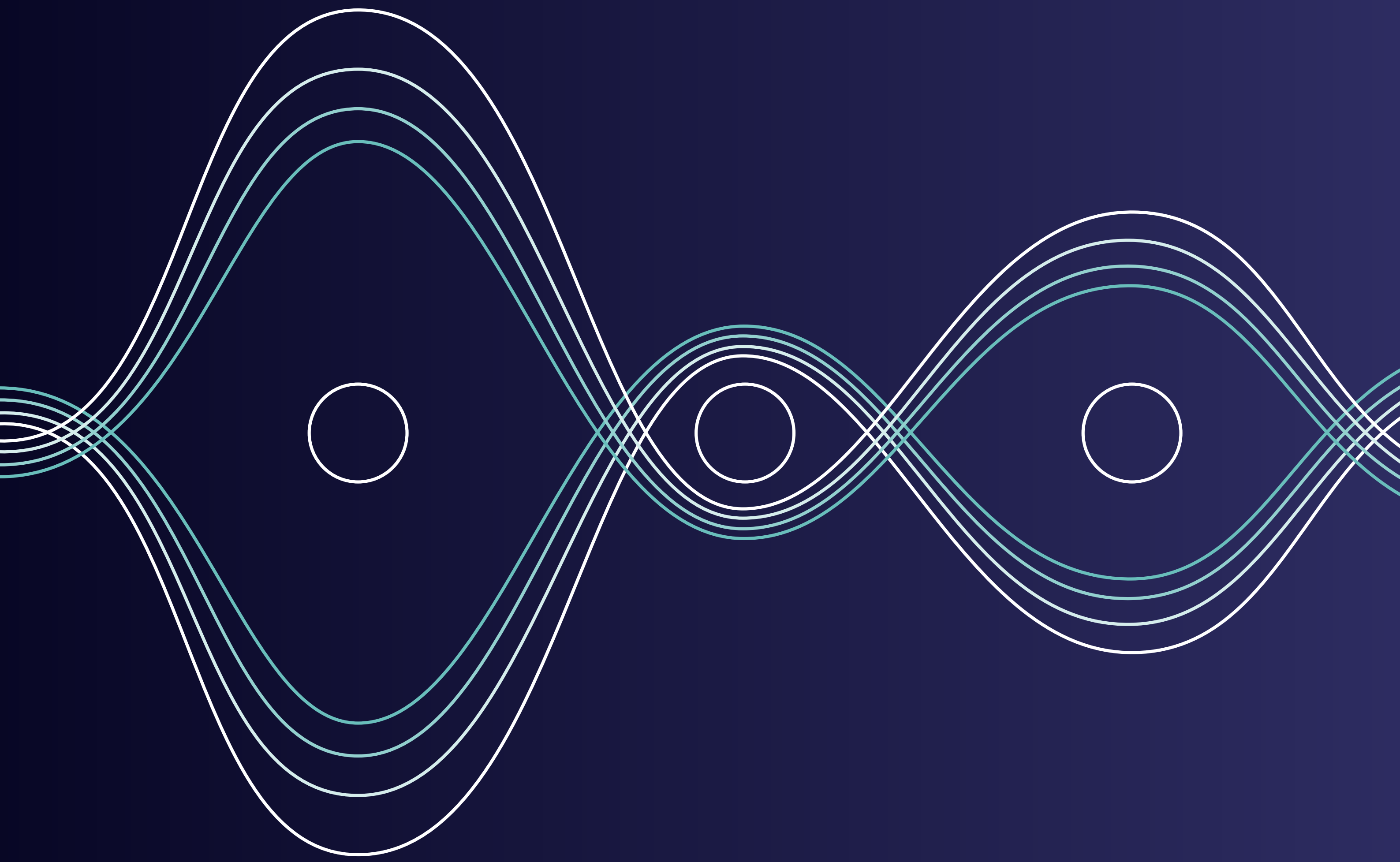
Advanced

Intermediate and advanced learners will think more about the organisational impacts and mitigations against cyberthreats.

The training module emphasises many of the key lessons from the recent British Library attack, such as the need for multifactor authentication, identifying and managing legacy sites, and cyber-risk awareness and training. A worksheet detailing the necessary steps to creating a digital salvage plan will be provided, alongside an information sheet that summarises some of the key findings from recently attacked institutions, including the British Library.

7. Preservation

How to ensure digital collections survive change



Our recommendation

To build a UK digital collection, we need to take a long-term view on the preservation of data. We should be able to access and use the data held in our digital collections beyond the limits of technical obsolescence, media degradation and organisational change.

What do we mean by this?

Hundreds of millions of pieces of data have already been created about collections held in the UK. There are many hundreds of millions still to make, transform and update. But how many have already been lost? How many could be lost as both physical technologies degrade and data standards change? How much valuable collections data in areas such as born-digital objects, with obsolescent file formats and technologies to contend with, has never been addressed because of their difficulty? What risk do the funding challenges faced by our institutions place on digital collections?

Today, we cannot answer those questions because we do not have a coherent, strategic, national, long-term view on digital preservation in the cultural heritage sector, and there is a shortage of the required skills and confidence across the workforce to develop one. The full potential value in digitisation will never be unlocked unless we develop our sense of the importance of preservation and the different kinds of technical preservation approaches that we could take – and act at speed and with thoroughness.

There are three things we need to do to ensure our digital collections will be preserved for the decades ahead:

1. **Make preservation strategy a key pillar of all future digital collections' development**. Build the strategic understanding, skills and technology resources to ensure that what we create cannot – through mistake or mismanagement – become obsolete.

2. **Consider coordinated investments in sector-wide preservation for legacy assets**. We should treat the digital collections data we have as under threat – and explore different mechanisms to ensure it can be preserved.

3. **Consider a future national infrastructure for preservation**. As part of a wider digital collections research infrastructure, this would be composed of policy, technology and people. Building on the work of organisations such as the Digital Preservation Coalition, making sure that there are collectively adopted standards, and that access to and use of repositories to ensure data is usable independent of its original platform, are areas to consider in this planning for the future.

Case study:

Museum Data Service

The Museum Data Service (MDS) aims to help connect and share the estimated 80 million object records held across the UK's 1,700 accredited museums, as well as further records from digital collections. It will also provide summary collection descriptions comparable to the 'finding aids' familiar to archivists.

To prevent data obsolescence through error or mismanagement as a preservation strategy, the MDS core system is built around Knowledge Integration's CIIM middleware, enhanced to meet specific MDS needs that include an innovative access permissions framework, a data transformation tool and a wizard for generating bespoke application programming interfaces (API). Museums can submit data in various formats, either via API or by uploading files, ensuring flexibility and adaptability to current standards. Data users can see when datasets were last refreshed. PIDs are minted for the majority of incoming records that do not already have them. Whatever the original field names of the source records, the data transformation tool allows mapping to any other schema. As part of the onboarding process, some or all of the fields in each museum's data will be mapped to corresponding Spectrum 'units of information'.

Due to cost considerations, MDS holds only metadata linking to externally stored media, fostering sustainable management of digital resources. Initially, access to this repository is restricted to MDS CIIM account holders, with a public interface planned for September 2024. This will be fairly simple to start with, but MDS will evaluate it and create a roadmap to develop it further, particularly to meet the needs of academic researchers.

The MDS is working with Art UK to enhance and transform records, aiming to double its online artwork listings to 600,000 by summer 2024. This initiative demonstrates coordinated investments in sector-wide preservation, ensuring that digital collections are preserved and adapted to current and future needs. MDS also assists small museums in sharing object records with local communities for content creation and knowledge management, exemplified by Wolverhampton Museum's community workshops and St Barbe Museum's collaboration with local history societies. It is important to stress that individual members of the public are not the target audience of the MDS. Rather, the MDS is a 'business-to-business' (B2B) service, making data available to people who will use it as the raw material for all kinds of end uses.

Sample training module descriptions

Relevant module 1: Storage

This module introduces users to the major storage options available to collections institutions, how they might be applied and how they might suit a range of different situations.

Three levels are included:

Beginners

Beginners will be introduced to the language and nature of different storage types, including on site, offline, networked, online, cloud and so on, so they can properly understand what each means as a first step to knowing how each might work for their organisation, as well as recognising that it might be beneficial for there to be a combination of different kinds of storage in place to serve different needs.

Intermediate

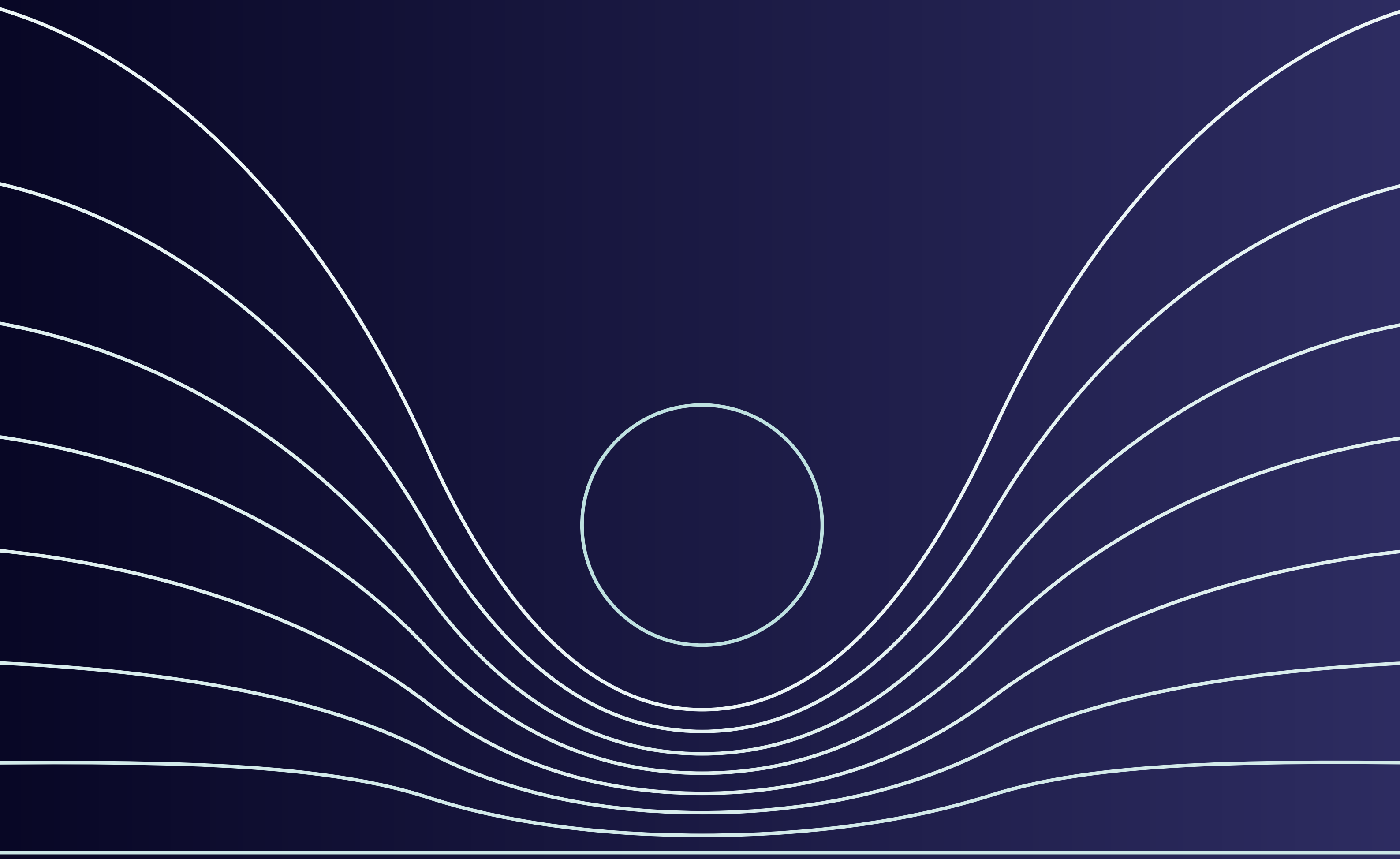
Intermediate users will consider the risks involved with each kind of storage, as well as the risks they can mitigate against. This will be supported by a storage possibilities decision chart, which will help learners to frame their questions and decision-making, including considering some of the possible costs.

Advanced

Advanced users will consider separating out storage for their data and their metadata, and the value of storing metadata in a system such as the Museums Data Service.

8. Impact

How to understand the usage of digital collections



Our recommendation

To build a UK digital collection, we need to understand how our digital collections are used so we can keep harnessing their cultural, social and economic value. Sustained audience research helps us understand this at a granular level, whilst evaluation allows us to understand the wider impact of what we have done on society. Both are critical ways to do this, and we need to make them a key part of how and why we build digital collections.

What do we mean by this?

The reason we create digital collections is for them to be used and engaged with. Because of this, the presentation of data online is not the end of a process – it is the beginning of a decades-long responsibility to understand how that data is used, and then to update, revise and improve the data to make it more useful to audiences.

When undertaking a digital collections project, investment is required to understand the quality of user experience being offered and how audiences will use it, both during its development and in the years that follow. There are different ways to do this – and a combined approach between three different methods will help to give a nuanced view:

1. **Audience insight should be gathered for both new and existing digital collections.** Both qualitative and quantitative insight should be collected, giving a balanced picture of usage. The creation of new data creates new opportunities to test and understand audience needs for that data.

2. **Audience usage data should be collected, subject to proper consents, whenever new data access points are launched.** Through tools such as PIDs and the licensing of data, and through digital analytics, we can develop a complete pattern of how our digital collections are used. This data is as much a part of a UK digital collection as data on the objects themselves – and so a common standard to help make informed decisions should be established.

3. **Thorough evaluation and impact measurement of large-scale digital collections projects should take place** after their launch to understand their wider impacts on society. These evaluation materials themselves are also critical data in the formation of a UK digital collection.

Case study 1:

Research user evaluation

In late 2023, Claire Bailey-Ross of Portsmouth University was commissioned by TaNC to gain a comprehensive understanding of the needs and requirements of different research users across academia and Independent Research Organisations, and what they would like to see included in a future UK digital collections infrastructure. The consultation used six focus groups, 40 interviews and a survey with almost 200 responses to understand digital infrastructure needs across various research fields and career stages.

Researchers expressed a desire for more digitised materials and recognised the need to balance shallow and deep digitisation approaches. Concerns were recorded about the long-term availability of collections data, highlighting the importance of a robust digital preservation framework to safeguard digital materials for future use. Increased support, Open Access initiatives and resources for staff training were all recognised as necessary to sustainably advance digital cultural heritage initiatives.

Respondents valued comprehensive search functionalities, advanced filtering and sorting options, and user-friendly interfaces for exploring and interacting with collections. Discoverability and serendipitous discovery were also highlighted as important aspects. Metadata accuracy, completeness and potential enhancement were raised consistently. Standardised metadata practices were recognised as being crucial for accurate description and discoverability.

There was strong support for connections to be built between and across collections and across institutions. Researchers felt that standardisation was key to sustainability and interoperability; researchers want a digital collections infrastructure to have true interoperation. Researchers felt it was important to balance technological advancements with the preservation of human expertise and fostering community engagement in and across digital platforms.

Researchers want sustainable practices to be integrated into a future digital collections infrastructure, and recognised the need for a fundamental shift in culture to acknowledge and actively work towards environmentally sustainable practices.

Collaborative efforts were recognised as essential to address challenges and leverage opportunities in developing a digital collections infrastructure. By prioritising sustainability, enhancing search and discovery, establishing standardised frameworks for interoperability, and fostering collaboration, Bailey-Ross concluded that we can create a more inclusive and interconnected digital cultural heritage research environment.

Case study 2:

Digital Footprints and Search Pathways

With the aim of understanding how people accessed cultural heritage content during the Covid-19 lockdown period, the TaNC Digital Footprints project analysed access logs to the digital collections of the National Museums Scotland (NMS) and National Galleries of Scotland (NGS) over a period of 12 months from April 2020 to March 2021 and compared those with the equivalent months for the previous three years.

NMS has over 12 million objects, with about 783,000 available to search online, and NGS's online collection has over 98,000 objects, of which about 80,000 are available to search. Both online collections have varying degrees of related content and associated data, and user journeys to collection items vary across platforms due to interface design and the data around the collection items.

Key insights gained from the project's data analysis included:

- Patterns of access during the year saw greater engagement with collections during lockdown compared to the previous years but with a very significant prevalence of new over returning visitors.
- Most access to the NMS and NGS sites came from computers (desktops or laptops), reflecting the conditions of the pandemic but also the continued existence of a broad device ecosystem.
- Despite the relevance of many NGS and NMS objects and collections to developing a richer understanding of contemporary social, political and other issues, users do not often search for such issues, and user studies demonstrate that users might not expect to find 'topical' or 'trending' content in cultural heritage collections.
- The return on investment for the use of external platforms is mixed and institution specific. Different collection items gain traction in different places, and the number of items available on the partner platforms does not necessarily translate to more views on the cultural institutions' website.

Case study 3:

Journal Usage Statistics Portal by Jisc

The Journal Usage Statistics Portal (JUSP) is provided by Jisc to offer a centralised service for libraries to access usage statistics of e-resources. Libraries often struggle with scattered and inconsistent data regarding the usage of e-resources. JUSP consolidates usage and data across diverse formats, providing libraries with insight into usage patterns, which helps them adjust their collections to better service their academic communities.

Libraries can make use of JUSP's analytics to make informed decisions about resource allocation, ensuring that investments in digital collections deliver maximum value and relevance. Library service managers can conduct comprehensive reviews of journal usage data and identify low-usage journals that could be cancelled and transitioned to subscription-based models. For example, Stranmillis University College in Northern Ireland cancelled 44 individual journal titles and reallocated those funds to databases that offered a wider relevance for their community.

Data usage collection can be integrated with library management systems, enabling automatic data harvesting and analysis. JUSP offers a server service called SUSHI (Standardized Usage Statistics Harvesting Initiative), where libraries can automate the retrieval of usage data reports. For example, the Open University integrated these services with Alma, their library management system, to enable data harvesting and analysis. This automation saved considerable staff time, reduced errors from data handling, and supported more frequent and accurate reporting.

By facilitating access to consolidated usage statistics, JUSP enables libraries to perform detailed analyses that inform strategic decisions about their digital collections. These insights can help libraries tailor their resources to better meet user needs. Providing analytics supports libraries in making informed decisions about resource allocation and investment. It can help with the automation of data-gathering processes, and standardisation through standard-based protocols to fetch consistent and compatible user data and resources. Finally, it also enables seamless integration with library management systems and other tools to manage electronic resource statistics and share those models or approaches with other organisations.

Case study 4:

Digital collections audit

The digital collections audit commissioned by TaNC and carried out by the Collections Trust between September 2021 and the end of January 2022 was the largest ever attempt to survey and benchmark the state of digital collections in the UK.

The aim of the audit was to understand the number, scale and attributes of digitally accessible collections across the UK cultural heritage sector that might form part of a future UK digital collection infrastructure. The study carried out a survey-based audit of 264 collections-holding cultural heritage institutions, of which 230 responded. Its main findings were:

- The total number of item-level records reported was nearly 146 million.
  - › *Of these, the British Library (BL), The National Archives (TNA), the BBC Archives and the Natural History Museum (NHM) together accounted for just over half (51%).*
  - › *A quarter of these item-level records, 37 million in all, have associated images or other digital media. Of these, TNA, NHM, BL, the BM and one other IRO accounted for 52%.*

- Records that describe a group of objects are held by 188 of the 230 institutions (82%). Not surprisingly, far fewer group records were reported than item-level ones.
  - › *BL, NHM, the National Library of Wales (NLW), Bradford Museums and Galleries and another IRO between them hold 54% of all group records reported.*

- Over 82% of the 230 respondents (190) said they publish at least some records on their own website.
  - › *These 190 organisations have made almost 98 million records available via their own websites.*
  - › *Nearly 25 million of these 98 million records (26%) have images or other digital media.*

- Forty of the 230 institutions (17%) only publish records online via third-party aggregators and other platforms (that is, not on their own websites).
  - › *These include national institutions (such as English Heritage, Historic Royal Palaces, National Trust for Scotland) and large civic services (including Glasgow City Archives, Museums and Galleries Edinburgh, Leeds Museums and Archives, and Nottingham City Museums and Galleries) as well as smaller institutions.*
  - › *A further 67 institutions reported that some of their records were only published on one or more third-party platforms.*

- Fourteen institutions said that some of their records were only available behind a paywall.
  - › *Around 5.3 million such records were reported, 98% from just four institutions, including: TNA (42%), Royal Artillery Museum (28%) and The Box Plymouth (Archives) (16%).*

- Asked whether their institution had 'an API that allows others to make use of your online collections', 41 respondents (21%) said they did.

This audit sets a standard we should follow. Further investment in understanding the path to a complete UK digital collection should be a priority.

Sample training module description

Relevant module 1: Collections as data

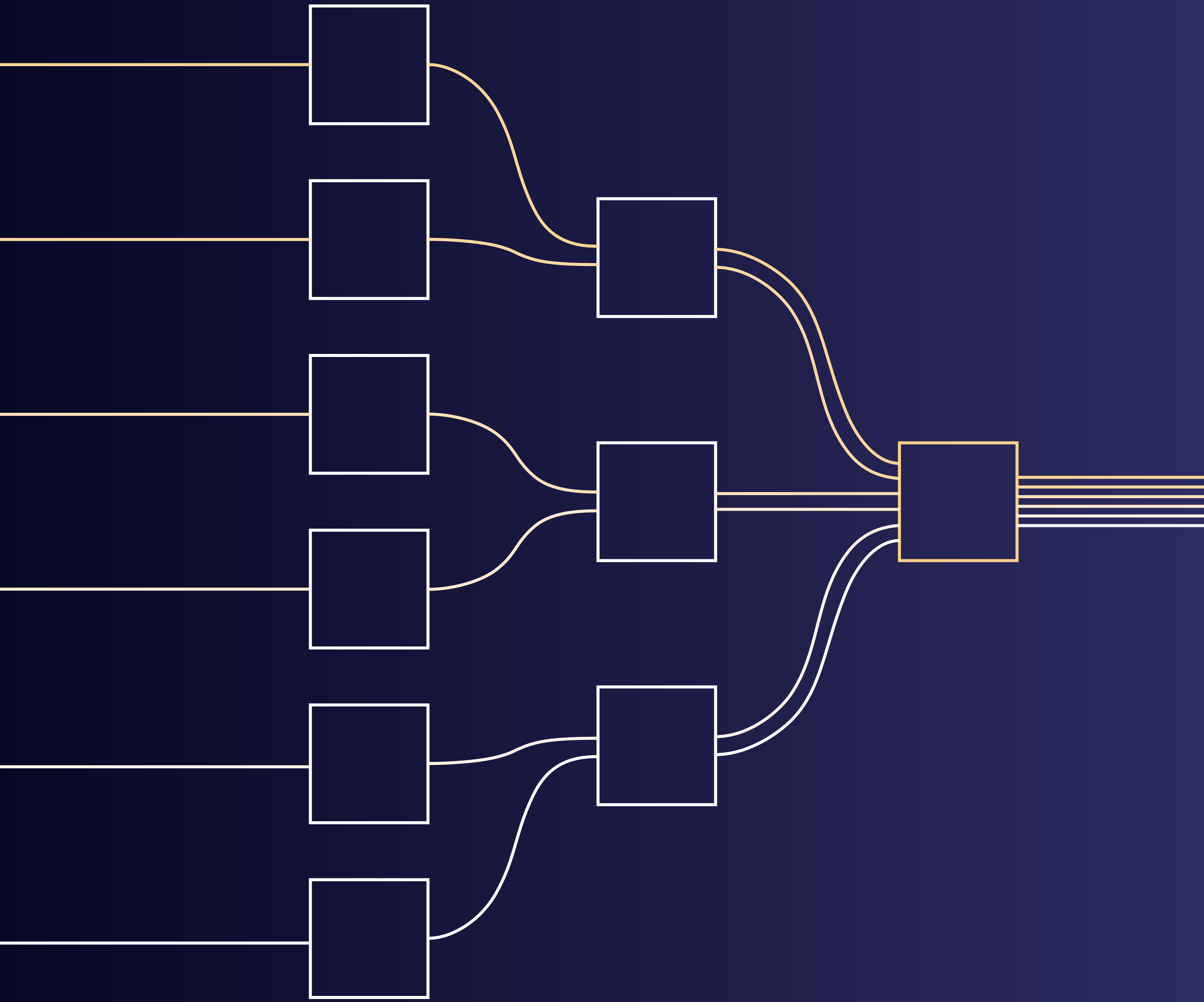
Learners will understand the aims of the 'collections as data' movement, and the opportunities and challenges of working across collections, including through linked data. Users will be able to practise some key collections data and metadata tasks using examples from GLAM datasets, which will enable them to understand the research potential of their own organisation's collections data.

Learners will discover the key concepts in the collections as data movement, and key examples such as the Living with Machines and Linked Art projects, demonstrating open datasets and open linking at scale. They will access data using a GLAM API and analyse this data to answer some basic data research questions. Learners will be introduced to key concepts in programming languages R and Python and use some basic coding to query their data.

For intermediate and advanced learners, additional activities will be available.

9. Models and frameworks

How to help digital collections evolve



Our recommendation

To build a UK digital collection, we need to treat digital collections as first-class research objects that will allow us to transform our understanding of collections and the world. To do this, we need to standardise our approaches to how we produce, manage and engage with digital collections. We must build on and adopt existing standards, and create scalable, long-term open standards models and frameworks.

What do we mean by this?

Neither digital collections nor a unified UK digital collection will be composed of static elements. Rather they will be sets and networks of dynamic, live digital entities composed of both people and technologies whose form is constantly expanding and evolving.

Our current digital collections are largely fragmentary because they have not been built to evolve. We need to change our approach to creating digital collections, maturing them in ways that are common elsewhere across the sciences and digital humanities.

To do this, we need to think of digital collections not just in terms of the datasets and the infrastructure that produced them. We need to think more holistically and recognise that the documentation and sharing of method, the use and reuse of software code libraries and widgets, and the creation of a sharable digital ecosystem are also key constituent parts.

To make this happen, we need to do three things:

1. **Ensure what we make is stored for reuse in open repositories.** We should promote and manage code libraries, applications and widgets so others working in similar collections domains can find them, understand them and build on them in their own projects. This will help speed up our digitisation and digitalisation projects and make them more efficient.

2. **Turn digital collections projects into open models and frameworks.** The design of new digital collections projects should start by considering existing open models and frameworks they could follow or adapt. After new data access points are launched in distinct collections areas, document them. Make sure those documents are openly accessible – and make them replicable frameworks for how to build digital collections in distinct fields.

3. **Manage and evolve both your digital collections projects and their documentation over the long term.** In the digital humanities, there are some standardised approaches to digital research around different kinds of knowledge that have retained authority for decades. We should learn from these and leading examples such as those in the Moving Image Archive sector. We need to mature our approach so our own investments fulfil their legacy and so they are genuinely helpful to the wider ecosystem.

Case study 1:

Making It FAIR

The TaNC Covid-19 project Making It FAIR was a response to the challenges faced by smaller museums struggling to engage online with audiences during lockdown and beyond.

These problems included low levels of basic digital capability, poor understanding of audiences, uncertainty over how to transfer real-world interpretive practice to the digital realm, lack of guidance about technical solutions, barriers to future-proofing digital assets and shoestring budgets. The difficulties faced by these smaller museums (and many larger ones too) would leave a huge volume of potential source material simply unavailable to researchers unless they could be overcome. In the project team's experience, too much museum activity relating to digital collections was resulting in outputs that did not meet the FAIR principles.

Between January and September 2021, the project team worked with a cohort of eight small museums as they navigated the challenges of staying connected with existing audiences, and reaching new audiences, through collections-focused digital content. The cohort received training, mentoring and technical support to plan and carry out digital storytelling experiments.

Making It FAIR pointed to the kind of collaboration between the digital humanities and the museum sector that would be of huge benefit to both, making available to future researchers museum-generated content that would not otherwise meet FAIR principles – or even survive at all.

Case study 2:

Global Indigenous Data Alliance – CARE principles for indigenous data governance

The Global Indigenous Data Alliance (GIDA) represents indigenous peoples' rights and interests in the domain of data governance. GIDA was formed as a response to the collective call for a framework that would protect and advance indigenous data sovereignty (IDSov) and governance (IDGov). This initiative arose from the 'International Law, The United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) and Indigenous Data Sovereignty' workshop in 2019, where scholars and practitioners emphasised the need for indigenous-designed legal and regulatory approaches founded on IDSov principles.

The CARE principles for indigenous data governance were formulated to supplement the FAIR principles, with the objective of shifting data governance into practices rooted in indigenous values and ethics. While FAIR principles aim to increase the usability and accessibility mainly of scientific data, their implementation can sometimes overlook indigenous people's rights and protocols, particularly regarding culturally sensitive data and ethics. In contrast, CARE principles extend these ideals by promoting indigenous self-determination and standards within data practices through collective benefit, authority to control, responsibility and ethics. From this perspective of indigenous peoples, data governance extends beyond knowledge systems and humans to encompass knowledge relevant to their cosmovision, ecosystem and other living things (beyond humans), which include the ecosystem, forests and landscapes essential for climate justice, within the perspective of well-being or *buen vivir*.

The CARE principles can help plan how data can benefit diverse communities. Combining the FAIR and CARE principles creates a more comprehensive strategy for establishing international standards for sustainable digital repositories. These standards can serve as a guide towards data that is fit for use and that meets the specific needs of different communities throughout the data lifecycle.

Following CARE principles ensures that indigenous peoples can delineate the use of their data in line with cultural norms and legal rights. It further engages with the responsibility principle by promoting respectful and informed use of indigenous data within academic institutions and research institutions. The principles also stress the importance of organisations acknowledging the need for collaborative models that grant indigenous peoples an active role in the stewardship of their heritage collections. Organisations can make use of these approaches to elaborate on fit-for-use data and workflows to strengthen and facilitate authority to control by allowing indigenous communities to annotate, manipulate and update data with clear indicators of indigenous ownership. The European Reference Genome Atlas, for example, has incorporated traditional knowledge within its metadata structure and ensures that indigenous knowledge remains within the governance frameworks established by the communities themselves, thus protecting and affirming their intellectual property rights.

GIDA's collaborative approach, involving diverse networks such as Te Mana Raraunga from New Zealand and the United States Indigenous Data Sovereignty Network, among others, engages with the effort to address the challenges faced by indigenous peoples in the data age through a unified global response. By facilitating international dialogue and sharing best practices, GIDA enhances the capacity of indigenous communities worldwide to govern their own data, reinforcing their autonomy and supporting their development aspirations. Through its work, GIDA is not just advocating for a change in data practices; it is fostering a movement towards a more inclusive and equitable data future, where the rights and aspirations of indigenous peoples are at the forefront of the digital age.

Case study 3:

World Historical Gazetteer

The World Historical Gazetteer (WHG), a project funded by the National Endowment for the Humanities and hosted by the University of Pittsburgh's World History Center, is an example of a digital humanities interface for transregional research. It is an innovative digital platform designed to collect, manage and share geographic data about historical places from around the world.

It serves as a valuable resource for historians, researchers, educators and the general public, providing tools to visualise, analyse and help understand historical connections through the lens of geography. The WHG aggregates vast arrays of historical place data, facilitating deeper insights into how historical events, and cultural and technological developments, have unfolded across different regions and time periods around the world.

The WHG provides a robust platform and API, and enhances its interoperability through the use of standardised data formats and protocols and linking its data with other datasets. Furthermore, it ensures that data governance models respect rights and cultural significance. This is achieved through the platform's capability to link multiple historical and modern name variants for a given place without privileging any particular one. Furthermore, this integration allows for a multifaceted representation of places, reflecting the complex histories and cultural significance attributed to them by different communities and groups.

The WHG offers storage for reuse in open repositories, ensuring that the data stored in the platform is not only preserved but available for reuse by the global community. The WHG operates as an open model itself, inviting contributions from a global community and supporting these contributions with tools and frameworks that facilitate the data integration and enrichment. For instance, the WHG's dual data store approach – comprising a relational database and a high-speed index – allows for the efficient management of data while supporting complex queries and data linking, which are crucial for turning individual collections into part of a broader, interconnected framework.

WHG draws significant inspiration from Pleiades, a pioneering project in the field of digital gazetteers, which has successfully curated and shared data on ancient places in the Mediterranean region. Pleiades' community-driven model – supported by volunteer editors and reviewers from the fields of classics, archaeology and history – demonstrates the potential for digital platforms to grow through active community participation and dedicated domain focus.

Sample training module descriptions

Relevant module 1: The MARC supremacy

This module provides users with an introduction to the importance of schema use in the context of connecting collections (and the pitfalls of going it alone with a metadata schema), as well as an introduction to some of the standards and standard tools that are used by collections professionals to describe their collections and manage their collections metadata. Users will be introduced to the most frequently encountered standards (such as MARC, CIDOC and DC), how these schemas relate to one another and how this reinforces the argument for using existing standards rather than developing new ones.

Relevant module 2: Towards Linked Open Data

This module introduces users to linked open data, and to some of the standards and technologies that allow this to occur. The module explains how curating data openly and according to standards makes it possible for data to be used with other datasets and collections. Learners will get to know the different standards of openness, as well as that linking data is, to some extent, an active process and that releasing data online does not automatically integrate it into the linked open data ecosystem.

Users will be introduced to the types of work required to reach each of the standards and some of the infrastructure each might require; they will be made aware that the underpinning of all linked data is the URI and the ability to consistently find and access data. Using this information, users will be able to understand what linked data is and how it relates to collections. More advanced users should gain a greater degree of confidence in identifying an approach to data openness that might suit their own organisation.

10. Experimentation

Research, development and innovation for digital collections

Our recommendation

To build a UK digital collection, we need to continuously and thoughtfully expose our collections to new technologies and new ways of thinking about the world. To do this, we need to scan the horizon for novel examples of each. We need to undertake early-stage technical R&D to trial new technologies for digital collections, and we need to use our data in near-to-market industrial innovation to help our collections shape tomorrow's technologies. Vitally, we need to focus on minimising the environmental impact of digital collections as an area in critical need of R&D.

What do we mean by this?

The act of building a unified UK digital collection is not a single task with a single outcome. Rather, whilst we continue to increase the total volume and quality of data we publish about our collections, it is through a continual process of evolution and change to the ways that digital collections can be accessed and understood that we will continue to generate new value from them.

Generating this new value means taking an experimental and innovative approach. There is a long history of using digital collections in both technical and creative R&D and acting with industry to discover the reciprocal value new technologies can create from digital collections, and the value digital collections can add to new technologies. We must now act in a coordinated way to maximise the shared benefits that this creates.

To help our digital collections benefit from, and add benefit to, emerging technologies and new ideas, we need to do four things:

1. **We need to listen – systematically scanning the horizon for emerging technologies and new ideas.** It is a collective responsibility of national importance to be aware of what is new and emergent in technology and ideas. We need to define a method for scanning and mapping the future and define a means of prioritising what has the richest potential.
2. **We need to scale up our investments in early-stage technical and creative R&D.** We need to systematically work with new technologies earlier and faster to find their potential – and then invest again in those which can make the biggest difference.
3. **We need to act with industry close to the market to help shape new innovations.** Our digital collections can help resolve the technical, ethical, intellectual and economic challenges business faces when trying to accelerate innovation. Our national-scale datasets are of enormous potential value to a data-driven technology market, and we need to channel our engagement to maximise its impact.
4. **We need to focus on environmental sustainability as a first priority.** One impact of doing more around our digital collections will be increasing the level of carbon they are responsible for producing. There is little understanding as yet of this issue – or what we can do to reduce the impacts we are creating. Substantial R&D is required, building on a shared vision for change that must become part of the future of digital collections.

Case study index

Appendix

Towards a National Collection research

Grant-funded research

All of the research grant funding in the Towards a National Collection programme was awarded by AHRC/UKRI. The awards were made in three phases – for eight Foundation Projects (2020–22), three Covid-19 Projects (2021–22) and five large-scale Discovery Projects. The Discovery Projects all started in 2021 and will finish between late 2024 and early 2025. Reports of the work of all of the projects are available in the TaNC community on Zenodo. The final reports of the Discovery Projects, with their findings and recommendations, will be published on Zenodo as each project finishes.

PI: principal investigator
Co-I: institutions providing co-investigators
Partner organisations: project partners and collaborating organisations

Foundation Projects

Deep Discoveries (February 2020 – July 2021)

PI: Lora Angelova, The National Archives
Co-Is: University of Surrey, Royal Botanic Garden Edinburgh, V&A Museum
Partner organisations: Museum of Domestic Design and Architecture, Gainsborough Silk Weaving Company, Sanderson Design Archive
Final report: https://doi.org/10.5281/zenodo.5710412

The Deep Discoveries project explored the potential of computer vision (CV) search for content discovery within and between digitised image repositories. The research led to the design of a prototype search platform enabling cross-collection image linking by harnessing the ability of CV methods to identify and recognise visual patterns without the need for preliminary integrated descriptive metadata. Searching in this manner allows for content linking based on attributes such as pattern, colour and motif, and creates the opportunity for users to discover unforeseen connections between image collections across the country. The research also introduced explainable AI methods, which allow users to enter into a visual dialogue with the AI so as to refine their search tasks. During the 18-month project, research carried out by a user experience (UX) research team from two GLAM Independent Research Organisations (The National Archives and the V&A Museum) informed the work of computer vision scientists at the University of Surrey. Using an agile working methodology and design sprints, the technological advances and UX findings were integrated into a prototype design by consulting interaction design (ID) partners from Northumbria University. The project worked with four partner organisations representing different owners and creators of visual collections to open up participation in funded research to smaller organisations, to glean a better understanding of their needs, and to assess the opportunities and challenges involved in gathering visual collections for the purpose of employing CV-based search and discovery tools.

Heritage Connector (February 2020 – December 2021)

PI: John Stack, Science Museum Group
Co-I: University of London
Partner organisations: V&A Museum, Natural History Museum
Final report: https://doi.org/10.5281/zenodo.6022678

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. This is largely a legacy of the development of these catalogues from handwritten paper records used primarily for managing collections rather than public access. The form of the catalogues means that the potential for new kinds of digital research, access and scholarly enquiry remain dormant. Searching across collections is currently possible only through aggregation, which is labour-intensive to implement, or by third-party search engines where results are variable and unreliable. This project applied artificial intelligence techniques to connect similar, identical and related items within and across collections. The primary research question was 'How can existing digital tools and methods be used to build relationships at scale between poorly and inconsistently catalogued digitised collection objects and other content sources?'

The Heritage Connector created a linked data Knowledge Graph that enables new forms of research and exploration. Furthermore, it explored the opportunity for computer-generated links with Wikidata to provide new levels of structure and machine-readable data that can form the foundation of new types of discovery and access. The Heritage Connector used a range of technologies including Named Entity Recognition, entity linking, open data and Knowledge Graphs. These methods created a large-scale data source of links. Computational inquiry to generate links via an application programming interface (API) enabled the creation of a range of proof-of-concept demonstrator research and discovery tools.

Persistent Identifiers as IRO Infrastructure (February 2020 – January 2022)

PI: Rachael Kotarski, British Library
Co-Is: University of Glasgow, Royal Botanic Garden Edinburgh, The National Gallery
Final report: https://doi.org/10.5281/zenodo.6359926

Heritage organisations in the UK house at least 200 million physical and digital objects. Being able to uniquely identify these objects supports their discovery, use and curation – you cannot provide persistent or even consistent access to an item if you don't know what it is. Accession numbers are a key component in all collection and library management systems, but these only cover selected objects within an individual collection. To fully realise the potential of our national collections, we need to link together collections across institutional boundaries.

Persistent Identifiers (PIDs) provide a long-lasting, clickable link to a digital object, recognised by UKRI as a tool for making content findable, accessible, interoperable and reusable (FAIR), and enabling citation and metrics. Supporting wider use of PIDs for collection objects, environments, specimens and related items will allow long-term, unambiguous linking that will create a digital national collection. However, the challenges, utility and wider benefits of PIDs are not well understood across the heritage sector.

The project brought together best practices in the use of PIDs, building on existing work and projects. Through a mixture of workshops, surveys, desk research and case studies, the project gathered evidence to develop an effective toolkit for the sector to make wider use of PIDs and provided recommendations on an approach to PIDs for colleagues and institutions across UK heritage.

Provisional Semantics: Addressing the challenges of representing multiple perspectives within an evolving digitised national collection (February 2020 – February 2022)

PI: Emily Pringle, Tate
Co-Is: University of the Arts London, Imperial War Museum, National Trust
Partner organisation: Royal Museums Greenwich
Final report: https://doi.org/10.5281/zenodo.7081347

Provisional Semantics originally set out to 'examine how museums and heritage organisations (and hence the digitised national collection) can develop ethical, equitable and transparent readings of artworks and artefacts that include the historically under-represented perspectives of people of African and Asian descent'. As the project progressed, it became apparent that this framing was too narrow and the terminology inadequate.

At the interim stage, the aim of the project was reframed as seeking 'to examine how museums and heritage organisations can develop interpretations and presentations of artworks and artefacts that acknowledge: the spiritual, cultural and historical value of artefacts and artworks; the context of their production, use and display in regions of the world that were part of the British Empire; where relevant, the nature of their subsequent transfer to the UK; and the historically underrepresented perspectives of British people with African, Caribbean and Asian heritage'.

The project tested and critically appraised different approaches by which cultural institutions can work with internal and external stakeholders to review existing cataloguing and interpretation practices and outputs, through three case studies at Tate, Imperial War Museum and the National Trust. The project sought to locate the case studies within a broader analysis of the sector's work in this area through a literature and practice review focused on the guidance and research available to support galleries, libraries, archives and museums, and the wider UK heritage sector, to engage critically in recording, writing and producing knowledge about art and artefacts.

The research team interrogated their own positionality alongside examining the effectiveness of project-based working in addressing the need for museums and heritage organisations to reinterpret and re-present objects in ways that can accommodate multiple, shifting interpretations.

As across the cultural and heritage sector, and society more widely, Covid-19 impacted the case study organisations and the members of the Provisional Semantics research team profoundly. Consequently, the project's processes, timescales and stakeholder interactions had to adapt to continuously changing circumstances. Despite, and perhaps because of, this, the research surfaced important findings and offers several recommendations regarding the opportunities for and barriers to ethical collaborative cataloguing and interpretation practice. Provisional Semantics offers key insights into the scale of the challenge in terms of the values, processes, practices, and resources needed for museums and heritage organisations to genuinely represent UK heritage.

Preserving and Sharing Born-Digital and Hybrid Objects From and Across the National Collection (February 2020 – March 2022)

PI: Natalie Kane, V&A Museum
Co-Is: Birkbeck College (University of London), British Film Institute
Partner organisations: British Library, Tate
Final report: https://doi.org/10.5281/zenodo.7144616

Contemporary culture is increasingly digital. However, this prevalence of digital culture poses a significant challenge to collecting organisations which are responsible for acquiring, preserving and making culture available to the public, now and in the future. In considering how to make our national collections accessible, we must consider born-digital and hybrid material as an increasingly important and uniquely challenging part of those collections.

This project focused on three challenges:

1. **Collections management** – the policies, governance, systems and standards needed to support born-digital collections.

2. **Digital preservation and conservation** – the skills, software and hardware needed to preserve them for the future.

3. **Meaningful access and experience** – the development of modes of access that do not merely represent digital culture as static but facilitate 'live' engagement with it, evocative of the complex and multivalent experiences it entails.

The project brought together an interdisciplinary team of academic and collections-based researchers and museum professionals, along with museum and heritage sector and industry expertise. By harnessing the collective skills, knowledge and challenges of individuals and institutions involved with different types of born-digital and hybrid cultural heritage, the project called for a move towards a greater understanding of the needs, challenges and affordances of born-digital and hybrid objects and their place within collections by setting the direction for further research. It also provided recommendations for the sector that embrace experimental collecting and proposed new models of stewardship, suggested new models for acquisition and provided potential contributions to standards.

Engaging Crowds: Citizen research and cultural heritage data at scale (February 2020 – April 2022)

PI: Pip Willcox, The National Archives
Co-Is: Zooniverse, University of Oxford, Royal Botanic Garden Edinburgh, Royal Museums Greenwich
Final report: https://doi.org/10.5281/zenodo.7152031

The project explored the current and potential practice of engaging diverse audiences with cultural heritage collections through the creation, use and reuse of heritage data. The last two decades have seen a revolution in volunteering programmes, as cultural heritage organisations have adopted digitally enabled approaches to crowdsourcing, and this project was part of that wider landscape.

The project had three focuses: community consultation on citizen research in cultural heritage organisations, including through workshops; prototype tool development for online crowdsourcing; and evaluating the tool through three citizen research projects and survey analysis. The project engaged with the wider community through seeking volunteers for the three citizen research projects and working with them once the projects launched, through our workshops, through conferences and through an open call for information about previous cultural heritage projects that used digitally enabled citizen participation. Taken together, the results of this work informed recommendations for best practice in encouraging and supporting meaningful public involvement with heritage collections.

Practical Applications of IIIF as a Building Block Towards a National Collection (February 2020 – April 2022)

PI: Joseph Padfield, The National Gallery
Co-Is: University of Edinburgh, British Library, National Portrait Gallery
Partner organisations: Royal Botanic Garden Edinburgh, Stanford University Libraries, Science Museum Group, Digirati, V&A Museum, IIIF Consortium
Final report: https://doi.org/10.5281/zenodo.6884885

Although a vast volume of digitised content has been created relating to our world-leading cultural heritage collections, digital resources often languish in institutional silos without the ability to combine or cross-reference them. The project highlighted and demonstrated the opportunities and benefits the International Image Interoperability Framework (IIIF) open standard can offer to institutions, researchers, scholars and students to open up these silos and to describe, present and reuse digital resources at scale. The project explored how IIIF can virtually connect digitised collections from different organisations and how these digital resources can be used as the foundations for further research. It experimented with presenting IIIF resources for utilisation by diverse audiences and considered ways in which IIIF can be used to lower barriers to uptake as well as create new opportunities for digital reinterpretation and presentation.

Efficient methods of using IIIF to build collaborative online resources were also explored to begin to demonstrate the potential for dynamic, cross-collection searching. As the organisation, presentation and reuse of digital resources would be a fundamental element of a digital national collection, the project sought to highlight how the consistent advocacy for, use and ongoing support of this type of established and mature set of interoperable open standards is the only realistic way of connecting, at a national level, hundreds if not thousands of collections and sources of digital resources together in an economically feasible and sustainable manner.

Locating a National Collection (February 2020 – July 2022)

PI: Gethin Rees, British Library
Co-Is: University of Exeter, National Trust, Historic Royal Palaces
Partner organisations: Historic England, English Heritage, Historic Environment Scotland, Museum of London Archaeology, Portable Antiquities Scheme
Final report: https://doi.org/10.5281/zenodo.7071654

The project helped cultural heritage organisations to use geographical information – such as where objects were made and used or the locations they depict and describe – to connect collections and engage the public. Through workshops, audience research and software development, the project developed a set of recommendations for using location to enhance the discovery of digital records across diverse collections. A set of thematic and technological case studies connected site records from historic environment organisations with objects from galleries, libraries, archives and museums virtually.

The Pelagios Network of researchers, scientists and curators developed a methodology that uses geographical information to connect research data with considerable success. The project built on their methodology by exploring methods of accessible and meaningful presentation to the public. The engagement work of the project encompassed survey and focus groups to understand the attitudes, behaviour and motivations of audiences such as community groups, heritage visitors and schools towards cultural heritage and location-based technologies. The infrastructure work created two web apps: a curation tool, Locolligo, and a web-map interface that can be embedded in organisational websites, Peripleo. The project encourages cultural heritage organisations to take up a common approach to creating and presenting geographical information with the ultimate aim of spearheading a movement beyond text-based searches in cultural heritage.

Covid-19 Projects

Making It FAIR: Understanding the lockdown 'digital divide' and the implications for the development of UK digital infrastructures (January–November 2021)

PI: Julian Richards, York University
Co-Is: Museum of London Archaeology
Partner organisations: The Collections Trust, Culture24, The Audience Agency, Intelligent Heritage, Knowledge Integration
Final report: https://doi.org/10.5281/zenodo.5846220

The project responded to challenges faced by smaller museums struggling to engage online with audiences during lockdown and beyond. The difficulties faced by these smaller museums (and many larger ones too) mattered to AHRC's aspirations for the digital humanities because they would leave a huge volume of potential source material simply unavailable to researchers. In the team's experience, too much museum activity relating to digitised collections was resulting in outputs that did not meet the FAIR data principles (data should be findable, accessible, interoperable and reusable).

The project team drew on academic researchers, museum sector support organisations and commercial IT practitioners, each bringing different skills and perspectives to bear on both the action and research sides of the work. Between January and September 2021, the project team worked with a cohort of eight small museums as they navigated the challenges of staying connected with existing audiences, and reaching new audiences, through collections-focused digital content.

The cohort received training, mentoring and technical support to plan and carry out digital storytelling experiments.

The methodology was built around the Let's Get Real collaborative action research approach developed by Culture24 over a number of previous projects, but adapted for delivery online in a time of home working and social distancing. In addition was a core collaborative action research study which included a socio-technical challenge: as the participants encountered difficulties along the way, the project team responded where possible and prototyped simple tools that demonstrate how a fully developed infrastructure might support the smallest and least resourced museums.

By considering a fully rounded picture of the digital problems faced by small museums, the project revealed insights into the scope and nature of the national infrastructure challenge.

Visitor Interaction and Machine Curation in the Virtual Liverpool Biennial (January–August 2021)

PI: Leonardo Impett, Durham University
Co-I: Liverpool John Moores University
Partner organisation: The Liverpool Biennial
Final report: https://doi.org/10.5281/zenodo.5770448

The project started from the observation that most machine learning and artificial intelligence systems are deployed in a GLAM context as either search engines, or as ways to automate cataloguing. In addition, the machine learning systems used in GLAM settings (e.g. textual or visual search engines) are almost exclusively 'unimodal': they work with one modality of information at a time. Instead, we proposed to use machine learning systems in a more tightly interactive setting, as a mixed-initiative system (an important paradigm for computer—human interaction in the context of artificial intelligence research). Furthermore, we used machine learning systems that translate between modalities – that turn images into texts, and vice versa. Beyond developing and launching our mixed-initiative co-curation system with the Liverpool Biennial, we also spent time investigating the implicit bias in the most widely used multimodal neural network, OpenAI's 2021 CLIP. Understanding bias in such networks will be an important part of using them in GLAM settings, both for mixed-initiative interactive systems like ours, and for more traditional search-engine or cataloguing-oriented systems. This led to new data (in terms of how audiences interact differently with active human—machine co-curation systems), and new research directions for digital curation, digital exhibition design and machine learning for visual art.

Our computer—human co-curation prototype, which makes extensive use of multimodal deep learning, is online at ai.biennial.com. A special issue of the Liverpool Biennial's Stages journal on 'Curating, Biennials and Artificial Intelligence' was published in Open Access to coincide with the prototype's release. Code to reproduce the main co-curation prototype on other datasets is available on GitHub, alongside a separate code repository containing experiments to investigate the bias embedded in CLIP, the main multimodal deep learning network used in the project.

Digital Footprints and Search Pathways: Working with national collections in Scotland during Covid-19 lockdown to design future online provision (January 2021 – March 2022)

PI: Gobinda Chowdhury, Strathclyde University
Co-Is: University of Edinburgh, National Museums Scotland
Partner organisation: National Galleries of Scotland
Final report: https://doi.org/10.5281/zenodo.6624800

With the aim of understanding how people accessed cultural heritage content during the Covid-19 lockdown period, the project conducted log analysis of access to the digital collections of the National Museums Scotland (NMS) and National Galleries of Scotland (NGS) over a period of 12 months from April 2020 to March 2021 and compared those with the equivalent months for the previous three years.

Key insights from this analysis include:

- Patterns of access during the year saw greater engagement with collections during lockdown compared to the previous years but with a very significant prevalence of new over returning visitors.

- Most access to the NMS and NGS sites came from computers (desktops or laptops), reflecting the particular conditions of the pandemic but also the continued existence of a broad device ecosystem of access to collections.

- Despite the relevance of many NGS and NMS objects and collections to developing a richer understanding of contemporary social, political and other issues, users do not often search for such issues, and user studies demonstrate that users might not expect to find 'topical or trending' content on cultural heritage collections; and although few online collection items are 'tagged' with language associated with contemporary issues, inconsistencies and what is tagged make exploring online collections challenging for users.

- The return on investment for the use of external platforms is mixed and institution specific. Different collection items gain traction in different places, and the number of items available on the partner platforms does not necessarily translate to more views on the cultural institutions' website.

The project findings revealed the need for:

- Better understanding of user (and non-user) needs and contexts.

- Strategies to prioritise the digitisation and sharing of collection items online, to identify which partner platforms to use and with what benefits, and to reach out to different audiences to achieve different levels of engagement.

- Creation of minimum data standards – metadata, vocabulary – which are user focused.

- Future use of emerging technologies like AI and machine learning for linking collections and improving search experience for cultural heritage collections.

Discovery Projects

The Sloane Lab: Looking back to build future shared collections (October 2021 – September 2024)

PI: Professor Julianne Nyhan, University College London and TU Darmstadt
Co-Is: University College London, the British Museum, and the Natural History Museum
Partner organisations: British Library, Historic Environment Scotland, Royal Botanic Garden Edinburgh, National Museums Scotland, Archives and Records Association, Collecting the West project funded by the Australian Research Council, Down County Museum, National Galleries of Scotland, Oxford University Herbaria and metaphacts
Project website: https://sloanelab.org

The founding collection of the British Museum is a rich area to explore how we can reconnect dispersed heritage connections using state-of-the-art technologies. This is because the British Museum's original 1753 founding collection of Sir Hans Sloane is now split across three different institutions (the British Museum, Natural History Museum and the British Library) and the digital information that describes this founding collection sits in the different institutions in a range of different systems that are not currently set up to talk to one another. By focusing on catalogue records, and the vast, remaining collections of Sir Hans Sloane, the Sloane Lab project is researching how we can work with interested communities and heritage organisations to link the present with the past so as to allow the currently broken links between Sloane's collections and catalogues to be re-established across the Natural History Museum, British Library and British Museum (plus others that have relevant material).

The main outcome of our project will be a freely available, online digital lab (the Sloane Lab) that will offer researchers, curators and the interested public new opportunities to search, explore, and critically and creatively use and reuse digital cultural heritage.

Unpath'd Waters: Marine and maritime collections in the UK (October 2021 – November 2024)

PI: Mr Barney Sloane, Historic England
Co-Is: Universities of Ulster, York, Southampton, Portsmouth, Bangor, Bradford and St Andrews, Glasgow School of Art, Historic Buildings and Monuments Commission for England, National Oceanography Centre, Historic Environment Scotland, Museum of London Archaeology, Royal Museums Greenwich
Partner organisations: Royal Commission on the Ancient and Historical Monuments (Wales), Maritime Archaeology Trust, Nautical Archaeology Society, Mary Rose Trust, Wessex Archaeology, Manx National Heritage, Marine Management Organisation, Northern Ireland Department of Communities, Cadw, Lloyd's Register and Lloyd's Register Foundation, Protected Wreck Association, British Geological Survey, UK Hydrographic Office
Project website:
https://unpathdwaters.org.uk

The UK Marine Area covers 867,400 km^2, 3.5 times the UK terrestrial extent.
Our marine heritage is extraordinary.

Shipwrecks from the Bronze Age to the world wars bear testimony to Britain as an island nation, a destination for trade and conquest and, in the past, the heart of a global empire. Coastal communities have been shaped by their maritime heritage with stories of loss and heroism. Deeper in time, what is now the North Sea was dry land, peopled by prehistoric communities. Our current land would have been distant uplands above hills and plains and rivers now lost and forgotten.

Numerous collections represent this heritage, covering 23,000 years, including charts, documents, images, film, oral histories, sonar surveys, seismic data, bathymetry, archaeological investigations, artefacts, artworks and palaeoenvironmental cores. These are unconnected and inaccessible. This matters because the story of our seas is of huge interest to the UK public, with millions visiting maritime museums annually, and marine exploitation increasing dramatically for energy, minerals, trade, food and leisure.

To unlock new stories and effect sustainable management, we must join up our marine collections. Unpath'd Waters brings together universities, agencies, museums, trusts and experts to confront this challenge. AI is being applied to innovate searching across collections, simulations to visualise landscapes, and science to identify wrecks and research their artefacts. Unpath'd Waters will deliver management tools to protect our most significant heritage and invite the public to co-design new ways of interacting with the collections. The methods, code and resources created will be published openly so they can be used to shape the future of UK marine heritage.

Transforming Collections: Reimagining Art, Nation and Heritage (November 2021 – January 2025)

PI: Professor susan pui san lok, University of the Arts London
Co-Is: University of the Arts London, Tate, National Museums Scotland
Partner organisations: Arts Council Collection, Art Fund, Art UK, Birmingham Museums Trust, British Council Collection, Contemporary Art Society, iniva (Institute of International Visual Art), Jisc Archives Hub, Manchester Art Gallery, Middlesbrough Institute of Modern Art, National Museums Liverpool, Van Abbemuseum and Wellcome Collection
Project webpage: https://www.arts.ac.uk/ual-decolonising-arts-institute/projects/transforming-collections

Whose voices, bodies and experiences are centred and privileged in collections? Transforming Collections is underpinned by the belief that a 'national collection' cannot be imagined without addressing structural inequalities, contested heritage and contentious histories embedded in objects. The project aims to uncover patterns of bias in collections systems and narratives, support digital search across collections and reveal hidden connections that open up new interpretative frames and 'potential histories' of art, nation and heritage.

Led by UAL in close partnership with Tate among 15 UK partners and one international partner, the project seeks to surface suppressed histories, amplify marginalised voices, and re-evaluate artists and artworks long ignored or side-lined by dominant narratives and institutional practices.

Our approach brings together academic and artistic research into collections and museum practices with participatory machine learning (ML) development and design. Working with smaller as well as larger collections and archives, and assuming that all data is biased, 'messy' and incomplete, the interactive ML development is critically shaped and driven by the research questions. The resultant case studies and lightweight adaptable ML are envisaged as research resources and tools to prompt critical reflexive analyses. Patterns generated through the bespoke creation of dynamic categorisations or tags refined by the user (that would not otherwise be visible through standard search functions within collections' databases) will encourage the rethinking of habitual formulations, hierarchies and values expressed in collections' digital records, while surfacing new connections between disparate objects and makers, to shape new research. A series of artistic residencies will also lead to new works that critically and creatively activate the research and ML tools. A public programme with Tate Learning will generate insights and understandings of the ways in which the project's research into and with museums and machine learning can enable new stories to be told with caution.

The Congruence Engine: Digital tools for new collections-based industrial histories (November 2021 – January 2025)

PI: Dr Tim Boon, Science Museum Group
Co-Is: Science Museum Group, British Film Institute, Historic Buildings and Monuments Commission for England, National Museums Scotland, Universities of London, Leeds and Liverpool, University College London
Partner organisations: BBC History, Birmingham Museums Trust, Bradford Museums and Galleries, BT Heritage and Archives, Grace's Guide to British Industrial History, Isis Bibliography of the History of Science, MadLab, The National Archives, National Museum Wales, National Museums of Northern Ireland, National Trust, Saltaire World Heritage Education Association, Society for the History of Technology, Tyne & Wear Archives & Museums (Discovery), V&A Museum, Whipple Museum of the History of Science, Wikimedia UK
Project webpage: https://www.sciencemuseumgroup.org.uk/project/the-congruence-engine

The capacity to connect historical objects and sources lies at the heart of this project as it does with the everyday museum and historical practices it is designed to support. Curators creating displays often combine artefacts, images and audio-visual materials from different heritage organisations. Amateur historians connect records from diverse sources to understand their ancestors and locales. Academic historians critically connect archive sources with existing literature to create new histories. All rely on connecting different fragments as they sew the quilts of our local and national histories. But this project seeks to make such linkages at the national scale of all collections, enabling access together to significant numbers of relevant items from many GLAM organisations holding heritage in many media. We aim to model a world where users can explore data neighbourhoods where information about heritage items from many sources will be readily to hand – museum objects, archive documents, pictures, films, buildings, and the records of previous investigations and relevant activity.

The project may be characterised as a kind of 'social machine' (Shadbolt et al 2022), a form of socio-technical practice requiring human propulsion and intervention in conjunction with the affordances of computing (including machine learning) techniques. The project is modelling the hybrid technical-historical practices that will be necessary whenever any kind of user wishes to undertake cross-collections work for curatorial and/or historical purposes. Our outputs address three kinds of audience: for the TaNC directorate, we will provide a 'design specification' of our 'social machine' replete with worked examples and overarching meta-considerations. For the museum-visiting public, we are creating a digital exhibit that shows how a national collection for industrial history might be created. For the workers in our home disciplines, we will communicate via conferences and two books that advocate for the virtues of working digitally across collections.

Our Heritage, Our Stories: Linking and searching community-generated digital content to develop the people's national collection (October 2021 – January 2025)

PI: Professor Lorna Hughes, University of Glasgow
Co-Is: Universities of Glasgow and Manchester, The National Archives
Partner organisations: Tate, the British Museum, Association for Learning Technology, Digital Preservation Coalition, Software Sustainability Institute, Archives+, Dictionaries of the Scots Language, National Lottery Heritage Fund, National Library of Scotland, National Library of Wales, Public Record Office of Northern Ireland and Wikimedia UK
Project website: https://ohos.ac.uk

Community-generated digital content (CGDC) is one of the UK's prime cultural assets. However, CGDC is currently 'critically endangered' due to technological and organisational barriers and has proven resistant to traditional methods of linking and integration. The challenge of integrating CGDC into larger archives has effectively silenced diverse community voices within our national collection. The project responds to these urgent challenges by bringing together cutting-edge approaches from cultural heritage, humanities and computer science.

Existing solutions to CGDC integration, involving bespoke interventionist activities, are expensive, time-consuming and unsustainable at scale, while unsophisticated computational integration erases the meaning and purpose of both CGDC and its creators. Our approach is fundamentally different: our project is using innovative multidisciplinary methods, AI tools, and a co-design process to make previously unfindable and unlinkable CGDC discoverable in our virtual national collection.

Our project is developing approaches to dissolve barriers to create meaningful new links across CGDC collections. We are also developing new methods of engagement, and making this content accessible to new and diverse audiences through a major new public-facing Observatory at The National Archives where people can access, reuse and remix this newly integrated content. This will facilitate a wealth of fresh research, while also embedding new strategies for future management of CGDC into heritage practice and training and fostering newly enriching, robust connections between communities and archival institutions. By enabling CGDC to be reused and reimagined, we will help it survive and be nourished, for the future and for our shared national collection.

Commissioned research

In addition to the three phases of grant funding awarded directly by AHRC, the Towards a National Collection programme directorate commissioned additional and interconnecting research of benefit to the overall programme and the wider cultural heritage sector. All of the reports are available in the TaNC community on Zenodo.

Online User Research Literature Review (December 2021)

Author: Dr Claire Bailey-Ross, Portsmouth University
Report: https://doi.org/10.5281/zenodo.5779826

The objective of the consultancy was to prepare a literature review and analysis of existing user research relating to the UK's gallery, library, archive and museum digital collections (including those related to the historic and natural environment). This included both published sources and unpublished internal reports, where these were made available by the institutions involved, with a concentration on material less than around six years old.

A Culture of Copyright (February 2022)

Author: Dr Andrea Wallace
Report: https://doi.org/10.5281/zenodo.6242611

This research provided a better understanding of the ways in which Open Access shapes how the UK's digital cultural heritage collections can be accessed and reused. The research comprised a scoping study and empirical analysis on Open Access policy and practice across the UK. The research was informed by interviews with UK cultural institutions.

Art UK: Opening up access to the nation's art (March 2022)

Author: Aidan McNeill, Art UK
Report: https://doi.org/10.5281/zenodo.6334193

Based on the work of Art UK, this research identified the major issues regarding image copyright standards across the GLAM sector and provided recommendations on how to manage these issues when linking collections and reusing digital images at a national scale.

Digital Collections Audit (March 2022)

Author: Collections Trust
Report: https://doi.org/10.5281/zenodo.6379581

The purpose of the consultancy was to undertake an audit to uncover the number, scale and attributes of the digitally accessible collections across the UK GLAM sector that might form part of a future UK digital collections research infrastructure.

User Research (June 2022)

Author: The Audience Agency in collaboration with Culture24
Report: https://doi.org/10.5281/zenodo.6684165

The objective of the consultancy was to carry out audience consultation in order to understand, at this fairly early stage in the programme, what users want and need from a future national collection digital infrastructure.

Foundation Projects Consolidation Report (February 2023)

Author: Dr Carlotta Paltrinieri
Report: https://doi.org/10.5281/zenodo.7674815

This report was commissioned to consolidate the findings of the eight Foundation Projects funded through the TaNC programme. The main aims were to digest the research done by each Foundation Project and to examine their contributions to TaNC, the scholarly community and the wider cultural heritage sector.

Impact Evaluation (first phase August 2023, second phase due March 2025)

Author: Diffley Partnership
First phase report: https://doi.org/10.5281/zenodo.8269842

The objective of the consultancy is to evaluate and report on the impact of TaNC in a focused number of areas of the cultural heritage and academic sectors, linked to the programme's key performance indicators.

Research User Consultation (July 2024)

Author: Claire Bailey-Ross, University of Portsmouth
Report: https://doi.org/10.5281/zenodo.12751226

The objective of the consultancy was to consult a wide range of research users in higher education institutions and Independent Research Organisations on their priorities and needs from a future UK digital collections infrastructure.

Total Economic Value Implementation (July 2024)

Author: Alma Economics
Report: https://doi.org/10.5281/zenodo.12755041

The objective of the consultancy was to establish the Total Economic Value of a digital cultural heritage collection infrastructure, in a way that is compliant with the UK Government Treasury Green Book.

Digital Collections Training Materials (November 2024)

Author: Ex Nihilo (Oxford University)

The series of training materials for small, medium-sized and community-based cultural organisations are designed to help those organisations in the generation and management of digital collections.

Also published was the programme directorate's **International Benchmarking Review** (December 2021). https://doi.org/10.5281/zenodo.5793173

In addition, consultants Human Economics undertook scoping work for the Total Economic Value study and worked with the programme directorate on an audit of open technologies, ongoing analysis of completed research, and the development of these policy recommendations. These all contributed to the work of the programme directorate but as internal documents were not published and thus are not listed above.

Glossary

API

API is the acronym for application programming interface – a software intermediary that allows two applications to talk to each other. APIs are an accessible way to extract and share data within and across organisations. They connect to the Internet and send data to a server. The server then retrieves that data, interprets it, performs the necessary actions and sends it back to the application. The application then interprets that data and presents you with the information you want in a readable way.

The term 'API' has been used generically to describe connectivity interfaces to an application. However, over the years, the modern API has taken on some unique characteristics that have truly transformed the technology space. First, modern APIs adhere to specific standards (typically HTTP and REST), which enable APIs to be developer-friendly, self-described, easily accessible and understood broadly.

Additionally, today, APIs are treated more like products than code. They are designed for consumption for specific audiences (e.g. mobile developers), and they are documented and versioned in a way that enables users to have clear expectations of their maintenance and lifecycle.

CIDOC CRM

CIDOC CRM (Conceptual Reference Model) is an international standard (ISO 21127) for the representation of cultural heritage information. It provides a formal structure to describe the implicit and explicit concepts and relationships used in cultural heritage documentation.

Computer Vision

Computer vision is a field of computer science that focuses on enabling computers to identify and understand objects and people in images and videos. Like other types of AI, computer vision seeks to perform and automate tasks that replicate human capabilities. In this case, computer vision seeks to replicate both the way humans see and the way humans make sense of what they see.

The range of practical applications for computer vision technology makes it a central component of many modern innovations and solutions.

Conceptual Reference Model (CRM)

A general framework or data model for representing information about the world. It offers a standardised way to describe entities (things) and their relationships in various domains.

Data as a Service (DaaS)

Data as a service is the sourcing, management and provision of data delivered in an immediately consumable format to an organisation's users as a service. It allows non-IT expert users to produce actionable insights without needing to understand the underlying technology, focusing on obtaining and utilising the right data for specific business tasks. The main objective is to enable users to access and utilise data without needing to handle the complexities of data management and infrastructure.

Data Model (Semantic)

A semantic data model is a way to organise data in a computer system. It captures not just the data itself but its meaning. This makes the data easier for both people and computers to understand.

Data Visualisation Tools

Data visualisation is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualisation tools provide an accessible way to see and understand trends, outliers and patterns in data.

Digitalisation

Digitalisation is the process of using and implementing digitised content in a broader and richer context within and beyond its original setting. It involves the transformation of the socioeconomic environment through the adoption, application and utilisation of digital artefacts. This encompasses various sectors and enhances how digital and even physical content is integrated into further environments, influencing how organisations monitor and evaluate impact outcomes through digital strategies.

Digitisation

Digitisation is the process of converting analogue cultural objects and goods to digital formats. This generally includes the selection and stabilisation of materials (which can often involve the preparing and preserving of the physical condition of the materials before they are digitised), prioritisation of collections, capture and quality assurance of captured materials, management and administration of digital resources, collection and creation of metadata, and the use and distribution of digital objects. In summary, digitisation is the production of digital items through technologies that convert, represent and enhance a previous digital or analogue object.

Digitally Born (Collections)

Digitally born refers to materials originally created and stored in a digital format rather than converted from a physical medium. These collections encompass a wide range of content, such as digital manuscripts, electronic records, online publications and digital art. Unlike traditional physical records, digitally born collections are intrinsically linked to their digital environment, relying on technology for their creation, preservation and access.

International Image Interoperability Framework (IIIF)

IIIF is a set of open standards for delivering high-quality, attributed digital objects online at scale. It is also an international community that is developing and implementing the IIIF APIs. IIIF is backed by a consortium of leading cultural institutions.

Knowledge Base

A system for storing and sharing knowledge using RDF (Resource Description Framework) and linked data. It enables the interoperable storage of information in RDF triples, facilitating efficient and standardised data exchange and retrieval.

Linked Open Data

Linked data is a set of design principles for sharing machine-readable interlinked data on the Web. When combined with open data (data that can be freely used and distributed), it is called linked open data (LOD). It is able to handle large datasets coming from disparate sources and link them to open data, which boosts knowledge discovery and efficient data-driven analytics.

Metadata

Metadata is data that provides information about other data (data about data), summarising basic details to make data easier to find, manage and use. It includes various types such as descriptive metadata (title, author, keywords), structural metadata (how data is organised) and administrative metadata (creation date, permissions).

Natural Language Processing

Natural Language Processing (NLP) refers to the branch of computer science – and more specifically, the branch of artificial intelligence or AI – concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics – rule-based modelling of human language – with statistical, machine learning and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

NLP drives computer programs that translate text from one language to another, respond to spoken commands and summarise large volumes of text rapidly – even in real time.

Open Access

Open Access (OA) is a publishing model that provides free, immediate and unrestricted online access to scholarly literature. The key principles of OA include the removal of price barriers (e.g. subscription fees) and permission barriers (e.g. copyright and licensing restrictions), allowing users to read, download, copy, distribute, print, search or link to the full texts of these resources.

Open Knowledge

Open knowledge can be understood as information that is freely accessible and reusable by anyone. Through open knowledge, anyone can access, use, modify and share the information with no legal, technological or social restrictions on how to use it. Open knowledge builds upon open data, which means the raw data is available and usable.

Open Licensing

Open licensing is a way to give people permission to access, use and share data. An open licence clarifies to others that they can access, use and share the data.

Open Science

Open science is a set of principles and practices aimed at making scientific research accessible to everyone.
It emphasises inclusivity, equity, and sustainability in the production and dissemination of scientific knowledge.

Open Source

Open source refers to software whose source code is freely available for use, modification and distribution by anyone. It is released under a software licence that outlines permissions and conditions for its use.

Open Standards

An open standard is a standard that's available for anyone to access, use or share. For example, open standards for data facilitate the publication, access, sharing, and utilisation of high-quality data for individuals and organisations. In addition, technology interoperability standards are specifications that define the boundaries between two objects that have been put through a recognised consensus process. The consensus process may be a formal de jure process supported by national standards organisations (e.g. ISO, BSI), an industry or trade organisation with broad interest (e.g. IEEE, ECMA) or a consortia with a narrower focus (e.g. W3C). The standards process is not about finding the best technical solution and codifying it but rather finding the best consensus-driven solution with which all the participants can live.

Persistent Identifier

A Persistent Identifier (PI or PID) is a long-lasting reference to a document, file, web page or other digital object. Most PIDs have a unique identifier which is linked to the current address of the metadata or content. Unlike URLs, PIDs are often provided by services that allow you to update the object's location so that the identifier consistently points to the right place without breaking.

RDF Database (Triplestore)

An RDF triplestore is a type of database that stores semantic facts. RDF, which stands for Resource Description Framework, is a model for data publishing and interchange on the Web that works on open standards.

Being a graph database, triplestores store data as a network of objects with materialised links between them. This makes RDF triplestores the preferred choice for managing highly interconnected data. Triplestores are more flexible and less costly than a traditional relational database, for example.

The RDF database, often called a semantic graph database, is also capable of handling powerful semantic queries and of using inference to uncover new information out of the existing relations.

Repository

A repository is a centralised place where data, metadata digital objects or other kinds of information are stored and maintained. Repositories are the core to storing data, ensuring their preservation and facilitating their search and retrieval. The vast majority of GLAM organisations use repositories, where they provide tools for metadata management (e.g. CMS), access control (e.g. DAMS) and data sharing (e.g. websites, exhibitions).

Site Search

Site search is the functionality that enables users to search a given website's content or product catalogues with speed and relevance. A good site search function is tailored to the specific website. Not only does a good site search constantly index the site to ensure the latest content is easily accessible, but it also guides users as they explore a website's content, helping them discover content they might not have even known they were interested in.

Shared Vocabularies

A shared vocabulary helps people and organisations communicate the concepts, people, places, events or things that are important to meet their needs or solve their problems.

A good shared vocabulary focuses on a specific area and uses clear, unambiguous definitions of the words and concepts they contain.

Shared vocabularies range from simple lists of words and their meaning to more complex products. The complexity of a vocabulary depends on the complexity of the problem being solved.

Wikibase

Wikibase is an open-source software for creating and managing collaborative knowledge bases, enabling the integration and sharing of linked open data. It provides a flexible, customisable infrastructure for organising structured data, facilitating participation from both humans and machines.

TOWARDS
A NATIONAL
COLLECTION

UKRI

Arts and
Humanities
Research Council