

SECOND REPORT

DISCOVERY PROJECTS

**Our Heritage, Our Stories:
Linking and Searching Community-
Generated Digital Content to develop
the people's national collection**

JUNE 2023

University of Glasgow | The National Archives
The University of Manchester

Table of contents

Executive summary	1
Abstract	6
Aims and objectives	7
Partnership Structure	9
Staffing structure	12
Overall programme	14
Events and consultations.....	16
Research approach.....	20
Research results.....	27
Project outputs	29
Cross-project collaboration	32
Sustainability and infrastructure	33
Interim development towards recommendations.....	34
Contacts	36
Annex	37

Authors

*Lorna Hughes (University of Glasgow), Marc Alexander (University of Glasgow),
Hannah Barker (University of Manchester), Riza Batista-Navarro (University of
Manchester), Ewan D Hannaford (University of Glasgow), Goran Nenandic
(University of Manchester), Pip Willcox (The National Archives)*

Executive summary

Our Heritage, Our Stories (OHOS) is opening up the wealth of existing community-generated digital content (CGDC) across the UK using innovative automated approaches. This work is drawing on multidisciplinary academic expertise in digital humanities, archives, history, linguistics, and computer science at our HEI partners, the Universities of Glasgow and Manchester, with world-leading archive and digital infrastructure development at The National Archives (TNA), the project's lead Independent Research Organisation.

The project draws together cultural heritage, humanities and computer science to dissolve existing barriers and develop scalable linking and discoverability across CGDC and the collections of TNA. We are collaborating in this process with leading UK heritage organisations, including Tate, the National Libraries of Scotland and Wales, Manchester Histories (Archives+), the National Lottery Heritage Fund, the Public Record Office of Northern Ireland, and a network of smaller regional and local heritage organisations holding digital content created by and relating to communities.

The project will develop an automated pipeline for the processing and enrichment of CGDC and showcase this newly enhanced and connected CGDC in a public-facing Observatory with an enhanced and generous interface. The Observatory will include a Remix Suite of open-source tools for the creation of new stories, allowing individuals and communities to 'remix', visualise, and share original and remixed metadata as linked data/stories. We will also develop a post-custodial model of best practice for the creation and management of CGDC.

Achievements

Key Year One milestones were accomplished: construction of our prototype AI pipeline (Milestone 2) which was successfully tested with a collection of CGDC; and our Observatory server was established for internal use (M1). In our second year, a prototype of our Observatory interface was developed (M3) internal to TNA for further refinement.

Other key milestones were achieved in our second year. The History Lab completed our first set of history research demonstrator case studies (M5); and the History and Archives Labs scoped over 100 community archives that have potential CGDC for our AI pipeline. We received data and metadata from our partners, the National Library of Scotland, who have been generous with their time, expertise, and collections. Thanks to data they have provided, in April 2023 the AI Lab will process a new tranche of data and revise the pipeline (M4), leading to deliverable 1. The AI, History, Linguistics, and Archives Labs have collaborated on a data strategy that informs the journey of data through the AI pipeline and into the public facing deliverables that are the responsibility of TNA (the Observatory Interface, D2, and remix set of tools, D3). Test data is emerging from community heritage organisations, and it has been agreed that we will receive a significant set of data from People's Collection Wales that will enable refinement of the pipeline and generate a critical mass of content for testing our data strategy and AI pipeline.

Several core data-related outputs have already emerged from the project, with these also forming the basis for the iterative development of refined future versions and informing the development of key deliverables:

- The AI and Linguistics Labs have developed a prototype of the automated methods for knowledge extraction and linking from CGDC. The ingested and enriched data from this initial prototype represents a first iteration of our final enhanced CGDC dataset.
- The Archives Lab have worked with the History Lab to engage with a network of community archives to scope CGDC collections.
- The Observatory team have developed a foundation for further versions of the underlying graph database of the project.
- The Observatory team have developed the search across enriched CGDC (OHOS database) and The National Archives' Discovery data, with a temporary internal interface to demonstrate the functionality, although this has yet to be made available in an operational format for testing by the wider team.
- A project website, describing the project and providing updates on project activities has been developed, with an appropriate academic domain being secured for this (www.ohos.ac.uk). A project twitter has also been set up (@OHOS_NatColl).
- Our partner Manchester Histories (working with Archives+) have completed a major report with the project on community heritage in the North West of England, feeding into scoping of community archives information needs and research on the ecosystems of CGDC.
- Preliminary results have been developed from an experiment by the Archives, History, and AI Labs to use Natural Language Processing to search for CGDC funded by the National Lottery Heritage Fund.

We have also delivered encouraging early research results based on testing multilingual data in our pipeline; this was presented at an international conference resulting in offers of new partners and data.

Central to the delivery of these data outputs has been engagement with partners, especially National Library of Wales, National Library of Scotland, and an extensive network of community archives. We have also revisited Manage Your Collection materials scoped in year one, now reappraised as more foundational to project goals than originally recognised.

We have developed foundational project strategies. Central to these is our Data Strategy, a key document that encapsulates the project workflow and highlights key data interventions required. This is available in two formats – a narrative version (annex C), and a diagram (annex D). This was developed from September 2022 to January 2023, with final signoff and agreement received from TNA in February 2023. These will be living documents, refined as further tranches of data are run through our pipeline. Earlier iterations of our Data Strategy defined the four tranches of CGDC in scope for the project and explored important and ongoing connections with community archive networks, including Community Archives and Heritage Group (CAHG) Scotland and Manchester Histories (working with Archives+), and our expanding network of community archives.

We have also developed an Engagement Strategy, which is a framework for stakeholder engagement, the development of user personae, and the basis for user testing. Our extensive partner network will ensure that all stakeholder groups have been involved in key stages of the project.

Further, we have developed a publications and dissemination strategy that ensures equitable publication of project outputs, especially respecting community input. We are keen to ensure that presentation of project results respects the contribution of partners responsible for intellectual input, and that the quality of any outputs remains high. For this reason, we have instituted an internal peer review process for all outputs linked to the project (including conference posters and presentations).

Challenges

Our project has faced significant recruitment challenges since Summer 2022. TNA have also faced significant upskilling issues with their project staff particularly in relation to key technologies. The academic partners have sought to assist by running a variety of training sessions, including a three-day in-person workshop in York.

As with all the TaNC Discovery Projects, we have faced particular issues in filling posts at our Independent Research Organisation partner which – in common with other areas across the TaNC programme – has faced issues in making appointments in the current challenging heritage environment in the UK. TNA had three Research Software Engineers (RSE) in post from May 2022 - February 2023, at which point one of them left the project. TNA has also had an Agile Delivery Manager in post since February 2022 (post uplifted to Senior Agile Delivery Manager in Feb 2023).

A central challenge has been scoping data for the pipeline. The partners have lacked access to key support on metadata, data models, ontologies and authority data, which TNA have been unable to provide due to recruitment and capacity issues as well as the challenges faced by the organisation due to external pressures. This has meant that the project has not been able to draw on staff with the levels of relevant expertise that were outlined in the bid.

These challenges have caused delays in terms of key milestones. The Observatory Interface (M3) was launched internally to TNA, but is not yet accessible to the wider project and partners for testing (due M13). Development work on the remixer suite (M6) has been limited, which has meant pushing back the deadline for this to September 2023 from the originally anticipated June 2022. Shared and constructive approaches for addressing these issues are in development.

There have been difficulties in retaining PDRAs at the HEI partners, as our researchers in the Archives, History, and AI Labs all left the project, leaving vacancies in all labs of several months (Archives Lab: 5 months, AI Lab: 6 months; History Lab: 3 months). While there has been a flux in staffing, on the positive side all vacancies at the HEI partners attracted very good candidates, and a large number of applications for all posts were received, allowing us to hire PDRAS best suited to the second phase of the project.

As of March 2023, we have two vacancies at TNA to be advertised: a metadata support post, which was outlined in the bid and has been unfilled since the start of the project, and an RSE. We are also advertising for a new project manager post at the University of Glasgow.

We have faced challenges associated with the slow re-opening of the heritage sector post lockdown – a lot of work went into formalising our project agreement due to the pressures on IROs and heritage organisations. This has also meant a reconfiguration of our calls for the Community Fund. Initial plans for workshops raising awareness of the fund and co-developing proposals with

community groups had to be redeveloped in year one and year two. The Community Fund will be disbursed in full, to a revised schedule and with more targeted engagement. We have also developed a plan for a greater degree of engagement with our second IRO Partner, Tate, in the final phase of the project, and we still plan to develop activities with an IRO partner we were unable to officially contract with due to administrative issues at the start of the project, The British Museum.

We have been significantly affected by the global pandemic, and certain of our more optimistic predictions and schedules for the first phase of the project had to be revised. We also faced short periods of staff absence (with Covid affecting every investigator, with the PI affected in summer 2022). In particular, developing our project workshops has been challenging – two scheduled and fully developed workshops to be held during this period had to be reformulated and rescheduled, especially as it has not been possible to ensure or guarantee the required breadth and depth of community participation in events. Strike disruption in the IRO and university sector also required cancelling one fully-organised workshop at very short notice (with consequences for our first Community Fund call, as outlined above). These events will be held (in some cases in a reimagined format) later in the project and their rescheduling should have minimal effect on the progress of other work in the meantime. In a positive development that responds to these challenges, we are now working with other Discovery projects and the Programme Directorate to open up workshops across the programme, enabling a more joined up approach to workshops and avoiding duplication of effort.

Despite the challenges in managing a project of this size and complexity, we have continued to develop our partnership across the three organisations in the project, and regular and supportive systems have been implemented for meetings, reporting, and cross-team engagement. We also had a positive all team workshop in York with a professional facilitator to support cross project working and to support the development of a shared understanding of the bid, the project, and the key deliverables and outcomes.

Partnership structure

Our key project partners are: National Library of Scotland (NLS), National Library of Wales (NLW), Public Records Office of Northern Ireland (PRONI), Wikimedia, Manchester Histories, Tate, Software Sustainability Institute (SSI), Digital Preservation Coalition (DPC), Association for Learning Technology (ALT), National Lottery Heritage Fund (NLHF), Dictionaries of the Scots Language (DSL), and the Historical Thesaurus of English (HT). Each of these partners contributes different expertise and skillsets to the project, in addition to access to resources, data, and feedback for varying aspects of project development.

Project outputs

- *OHOS* has developed a first iteration of the project AI pipeline (Deliverable 1) to assess, harvest and integrate very large amounts of CGDC into TNA's Discovery; work on further iterations is ongoing.

- A first iteration of an interface for depositing collections has also been developed by the AI Lab, which will enable community groups to utilise the AI pipeline to improve their archival metadata and facilitate the contribution of data for further enhancement and use on *OHOS*.
- To highlight the value of linking CGDC to existing collections, we have begun the creation of compelling academic research studies using CGDC (Deliverable 5) These have primarily been produced by the History and Impact Lab so far, but will also be developed by other Labs to be published in the fields of history, linguistics, archive studies, and digital humanities as demonstrators of the value of this newly-discoverable content.
- We have developed several key project policy documents that represent our continually developing understanding of the process of gathering, integrating, and sustaining CGDC, including our project Data Strategy and Engagement Strategy.
- Each Lab has also produced various dissemination activities, with research on *OHOS* at several national and international conferences. Additionally, the Linguistics Lab has a journal article in editorial stages, scheduled for publication in May 2023.

Interim Recommendations: *OHOS* priorities with implications for a national collection

Archives lab: continue to scope the ecosystems of CGDC, feeding into the development of a post custodial model for CGDC through extensive engagement with community archives and collecting organisations.

History and Impact Lab and Archives Lab: continue to work with community groups to develop best practice recommendations for creating, managing, and using CGDC, with ethical pathways for making accessible and linking data produced by these groups.

History and Impact Lab: consult ‘clusters’ of producers and potential users of CGDC; addressing themes of modern British history well represented by GCDC, including the women’s peace movement; local and community history; pop music and pop cultural history; Second World War memory studies; disability activism; black history; migration history.

History and Impact Lab: continue to engage historians in a discussion about the use of CGDC through presentations at key historical conferences.

Linguistics Lab: continue research on potential significance of varied collections as a tool for promoting language equity and preserving linguistic diversity and variety in the UK.

The Observatory Lab: continue to explore technical- and human-centred recommendations for a national collection, using standard approaches, infrastructure, metadata, data models and authority data. Take a human-centred approach to designing solutions, involving broad range of users to create a community of invested users with ownership of the final product.

Across the project: nurture a shared understanding of terminology across disciplines to help communication between the disciplines in the project.

Abstract

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. *Our Heritage, Our Stories (OHOS)*, 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 TNA 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 re-used and reimagined, we will help it survive and be nourished, for the future and for our shared national collection.

Aims and objectives

- Situate community-held and community-generated digital content (CGDC) as part of a national collection: make it discoverable and linkable using existing and emerging approaches, including AI, and create a post-custodial approach to discovering and managing community-generated content. We will leverage the full range of our partnerships' strengths to do this, via a set of integrated Labs that utilise the project team's wide-ranging experience and expertise.
- Undertake an extensive and collaborative programme of research to build semantic mapping and representation of CGDC via AI-generated knowledge graphs. Through research into the information ecosystems of CGDC, including its creation, description and management, we will scope the full range and complexity of this content across the UK to understand its linguistic and cultural diversity and richness, including the use of community languages, dialects, and expressions to authentically describe entities, experiences, and content.
- Deliver a series of computational and infrastructural interventions that will make CGDC searchable and discoverable, and break down existing silos. These include co-designing and building an AI 'pipeline' to assess, harvest, and integrate large amounts of CGDC data into our Observatory at The National Archives (TNA); tools to semantically link this content to the collections of TNA and beyond, including complementary collections held by community-facing bodies, especially local museums, archives, and heritage organisations; and the production of 'generous interfaces' offering rich, browsable views that make clear context and relationships between collections, co-designed with CGDC creators and holders.
- Build and implement a sophisticated project framework that is at the cutting edge of linking digital collections and which leverages existing resources, platforms, and technologies, including TNA's Discovery catalogue system and Manage Your Collections platform, to visualise, remix, and share CGDC.
- Develop and disseminate research studies showcasing the value of remixing, enriching, and reusing CGDC for new cross-disciplinary and cross-collection research questions that would be impossible to formulate or address without a unified approach to the content.
- Use these studies to undertake an extensive process of vibrant and inclusive public engagement with stakeholders representing a broad and truly national cross-section of all aspects of community heritage. This includes collecting institutions; local archives and historical societies working with community-facing content; archival and collecting professional associations; specialist subject networks, including teaching and training networks; and individuals and community groups creating and using CGDC.

- Share innovative methods and exchange findings across the TaNC programme in order to co-develop best practice, discover potential new collaborations, and engage with the wider questions of a national collection.
- Exemplify new ways of understanding 'citizen history', while diversifying and extending audiences, centring and amplifying previously marginalised voices.
- Co-design and prototype post-custodial models of archive management with partner and collaborating organisations and the extensive community of practice working with this material. A new 'toolkit' will ensure future discovery and use, and create liminal and porous connections between communities and collecting organisations, situating community expertise as central to the recording, safeguarding, and communication of their artefacts.
- Produce research outputs on our advances in AI, our new understanding of the language and dialect of community content, the use of CGDC in historical research, and our models and toolkit of new post-custodial archival practice, and also release code and documentation for our tools and pipeline.

Partnership Structure

OHOS is an interdisciplinary University-IRO collaboration, with a network of collections partners and CGDC stakeholders, as well as organisations at the forefront of linguistic, technical, educational, and digital preservation development. The project lead is at the University of Glasgow, where the project PI, and Deputy PI and Co-I, and a fractional PM post are based, with a new Project Manager post in recruitment. There are two fractional PDRAs at (in archives, and linguistics/project management). Glasgow is responsible for overall project management and allocation of project roles and responsibilities, the Archives Lab, and the Linguistics workstream. The new project manager post at Glasgow, with overall responsibility for delivery across the whole project, has been created, funded through salary underspends. At Manchester, there are three Co-Is, two in AI and one in History, and two fractional PDRAs in each of these areas. Manchester is responsible for delivering the History Lab and the AI Lab. Our intersecting 'lab'-based structure feeds into the development of our AI pipeline and our CGDC Observatory, which is the responsibility of The National Archives (TNA). The project's five-person team at TNA comprises a Co-I (with responsibility for all aspects of Observatory delivery), a senior agile delivery manager, and three Research Software Engineers (RSEs) (see Appendix A).

National Library of Scotland (NLS)

NLS are providing the project with access to CGDC within their collections for the development of the project's automated approaches and for showcasing in the final project Observatory. This will also contribute multilingual data in the provision of Scots and Gaelic materials. They are also contributing expertise in the areas of community collections, cultural heritage, and digital content management to develop three case studies for the project, feeding into the development and testing of project outputs.

National Library of Wales (NLW)

NLW are providing access to community-generated materials within their collections for the development of the project's automated approaches and for showcasing in the final project Observatory. Additionally, their specific expertise and extensive experience is supporting the project in: scoping the methodologies of community content creation by People's Collection Wales staff at NLW; research input from collections staff working on the challenges of collections with community generated content and metadata, including multi-format collections that are hybrid and complex (such as the Brith Gof and Eisteddfod collections, input on bilingual authority data, and on the challenges of semantic mapping of Welsh language content).

Public Records Office of Northern Ireland (PRONI)

PRONI is supporting *OHOS* by providing access to unique community-generated collections held by PRONI, contributing to our development of automated approaches to processing community-generated digital content. PRONI are also providing expertise in community generated, and community held, archives and collections in digital and born digital form, as well as contributing to

the development and delivery of key case studies and workshops throughout the course of the project.

Wikimedia

Wikimedia are providing support to *OHOS* by: running a joint Language Data and Wikidata Workshop; contributing to the project's White Paper, 'Capturing and amplifying multilingual community heritage'; and providing specialist advice on our approach to data linking, including in linking multilingual data.

Manchester Histories (working on behalf of Archives+)

Manchester Histories works with a broad range of community groups and archives and are contributing to *OHOS* by enabling the project to incorporate data from these organisations into the project's AI pipeline development and the final project Observatory. They are also providing their extensive expertise in collaborating with community organisations, contributing to the development of the project's post-custodial model and facilitating greater collaboration with further community groups.

Tate

Tate are our second IRO partner. They are currently carrying out focussed research and development for *OHOS*, including: sharing specific findings from relevant research projects conducted by Tate, especially those engaging with communities working with Tate archives; contributing practice-led experience to the development of case studies; advice and input into community-facing activities; and organising public engagement activities with communities of practice.

Software Sustainability Institute (SSI)

SSI are contributing to *OHOS* through collaborating on key dissemination activities, including developing case studies and workshops that showcase project activities and situate these in the broader community of practice, including cross-disciplinary communities.

Digital Preservation Coalition (DPC)

DPC are providing *OHOS* with specific expertise in the information ecosystem for CGDC and metadata challenges in this area. Furthermore, DPC are providing access to their diverse global network of members; this contains extensive expertise in working with community generated content and challenges of community co-creation, including organisations already using and developing a post custodial approach. This expert network is being used to validate and amplify the research and outputs of the project, creating impacts in broader sectors and geographies than could typically be engaged.

Association for Learning Technology (ALT)

OHOS is being supported by ALT through their sharing of: data and case studies from their research on the digital heritage of community organisations; expertise on digital training and development, especially for library, archives, and museum education, and support in the development and dissemination of training materials; and support in the use of ALT's Accreditation Framework for Learning Technology, CMALT, to embed project research and outcomes in academic programmes in archives and community history.

National Lottery Heritage Fund (NLHF)

NLHF are contributing their expertise in collaborating and engaging with community heritage to *OHOS*, as well as their experience in managing and sustaining community generated digital content. This will support *OHOS* in community engagement activities and organisation of the project community fund.

Dictionaries of the Scots Language (DSL)

DSL are providing the project with licensed access to the full data of DSL, the national record of the Scots language from the twelfth century onwards, with over 75,000 entries (including definitions, quotations, and variant spellings). This data will be implemented to improve the automated approaches of the AI and Linguistics Labs in interpreting multilingual CGDC. As well as providing assistance in interpreting and integrating this data, DSL will also be contributing to: a case study about the use of its data on *OHOS*; a project White Paper ('Capturing and amplifying multilingual community heritage'), drawing on DSL's expertise with government and policymakers; and a one-day Language Data workshop.

Historical Thesaurus of English (HT)

HT are providing *OHOS* with licensed access to the full data of HT, the largest ontology in existence of concepts recorded in the English language over the last thousand years, including English and Scots dialect forms, with over 800,000 words across 250,000 categories. This data will be implemented to improve the automated approaches of the AI and Linguistics Labs in interpreting multilingual, multidialectal, and historical CGDC. As well contributing this dataset, HT are also providing their expertise in historical and multilingual data, natural language processing, and complex conceptual ontologies.

Staffing structure

Investigators

Prof Lorna Hughes (Principal Investigator, University of Glasgow) oversees and manages the project with responsibility for budget oversight, partner management and liaison, risks, and contingencies, in addition to leading the Archives Lab and drawing together the archival and infrastructural elements of the project.

Prof Marc Alexander (University of Glasgow) deputises and supports Hughes in project management, and leads work in the Linguistics Lab, particularly in the areas of multilingual and multidialectal inclusion and variation, computational linguistics, and language ontologies.

Pip Willcox (The National Archives) leads the Observatory Lab, overseeing its technical delivery, deputy-leads the History and Impact Lab, and coordinates TNA-based workshops.

Prof Hannah Barker (University of Manchester) leads the History and Impact Lab, overseeing and designing the scoping and use of CGDC to build research capacity.

Prof Goran Nenadic (University of Manchester) leads the AI Lab and is responsible for the overall design and implementation of the AI pipeline.

Dr Riza Batista-Navarro (University of Manchester) leads the design of NLP tools and methods within the AI Lab.

Managers

Dr Ewan Hannaford (University of Glasgow) is part-time Project Manager, responsible for overall project planning, project liaison with TaNC programme activities, Partner relationships, and PI support.

Hazel Jell (The National Archives) is Agile Delivery Manager, co-ordinating the delivery of the project Observatory. Since February 2023, her role was uplifted to Senior Agile Delivery Manager to include cross-Lab collaborative planning and data management.

To be appointed: Project Manager at Glasgow, responsible for overall programme delivery and workplan implementation as well as heritage partner management.

Research Assistants/Associates & Research Software Engineers

Dr Andrew Bewsey and **Harshad Gupta** (The National Archives) are Research Software Engineers, working on the construction of the Observatory, and the development of its analytical tools and user interface. One RSE is to be recruited following departure of Waltteri Nybom.

Archival Metadata Specialist post at TNA: a post originally intended to be one of the RSE posts, and vacant since the start of the project, is currently in recruitment at TNA.

Dr Stefan Ramsden (University of Manchester) is research assistant within the History and Impact Lab contributing to identifying and engaging community archives, developing the project's post-custodial model for community archival management.

Rhiannon Lewis (University of Glasgow) is research assistant for the Archives Lab, researching community archives and post-custodial frameworks, disseminating these frameworks, and identifying and liaising with existing and prospective community archive partners.

Dr Youcef Benkhedda (University of Manchester) is research assistant within the AI Lab, developing the automated approaches to ingesting, processing, and enriching community-generated digital content.

Dr Ewan Hannaford (University of Glasgow) is part-time research assistant working on linguistics-based work in the Linguistics Lab (including AI coordination work and multilingual/multidialectal data).

NB: PDRAs in the History Lab (**Stefan Ramsden**); Archives Lab (**Rhiannon Lewis**) and AI Lab (**Youcef Benkhedda**) were hired to replace departing PDRAs from these Labs, being recruited as PDRAs that were best-suited to the second phase of the project.

Diagrams of the OHOS staffing structure and Labs can be found in Annex A.

Staff training activities

We have undertaken an internal training programme in order to upskill project staff in key aspects of the project, the data in scope, and the tools and methods required for delivery. This was discussed and initiated at the project's Year 2 Kick-Off Workshop in October 2022, which included a team building and mediation session (Agenda attached in annex H). A further programme of activities was begun in February 2022, to address identified and reported gaps in knowledge and understanding.

Workshops:

- Oct. 12th-14th 2022: Year 2 Kick-Off Workshop
- Feb. 24th 2023: What is CGDC? Presenters: Hannah Barker and Lorna Hughes, audience: TNA and PDRAs.

Overall programme

OHOS workplan overview		
Activity	Start date	End date
1: Scope and map current community archive practices and barriers	Oct 2021	Sep 2023
2: Outreach and build community of practice with community archives	Oct 2021	Sep 2023
3: Develop processing and enrichment methods for CGDC	Oct 2021	Apr 2024
D0: Data tranche 0 collection and processing	Nov 2021	Jun 2022
4: AI pipeline V1 and NLP model development	Nov 2021	Jun 2022
5: Partner CGDC data gathering for ingest in D1	Dec 2021	August 2023
6: Establish platform & protocols for data sharing across project team	Jan 2022	June 2023
7: Initial research demonstrator case study development (History)	Mar 2022	Jan 2023
8: Refine CGDC processing/enrichment during AI pipeline development	June 2022	Feb 2024
9: Observatory V1 development	Jun 2022	Sep 2022
M1: Data tranche 0 processed and AI pipeline V1 developed	June 2022	
10: AI pipeline V2 development and NLP model revisions	Jul 2022	Dec 2023
11: Observatory and Manage Your Collections integration	Jul 2022	Feb 2024
D1: Data tranche 1 collection and processing (low complexity CGDC)	Aug 2022	Mar 2023
M2: Observatory server online (internal)	September 2022	
M3: Observatory interface developed (internal)	September 2022	
C1/C2: Community Fund Calls (Targeted)	May 2023	Nov 2023
12: Remixer Suite V1 development	Oct 2022	Sep 2023
13: Develop model of best practice and dissemination for CGDC	Oct 2022	Sep 2024
C3: Community Fund Call (Open)	Dec 2023	Mar 2024

14: Produce case studies of community-managed artifacts and collections	Feb 2023	Jul 2024
M4: AI pipeline V2 developed	March 2023	
M5: First research demonstrator case studies release	March 2023	
15: Refinement of CGDC processing methods, inc. linguistic resources	Apr 2023	Aug 2024
16: AI pipeline V3 development and NLP model revisions	Apr 2023	Dec 2023
D2: Data Tranche 2 collection and processing (medium complexity CGDC)	May 2023	Dec 2023
M6: Remixer Suite V1 developed	September 2023	
M7: Integration of multilingual resources into AI pipeline	September 2023	
17: Remixer Suite V2 development	Oct 2023	Sep 2024
M8: AI pipeline V3 developed	December 2023	
18: AI pipeline V4 (final) development and refinement	Jan 2024	Jun 2024
C4: Community Fund Call (Open)	Feb 2024	Jul 2024
D3: Data tranche 3 collection and processing (high complexity CGDC)	Feb 2024	Sep 2024
M9: Observatory interface finalisation and delivery	March 2024	
19: Launch of Observatory, sustainability work	Mar 2024	Sep 2024
M10: Final set of research demonstrator case studies release	April 2024	
M11: Release of post-custodial model/toolkit	June 2024	
M12: AI pipeline V4 (final) completion and release	June 2024	
M13: Remixer Suite V2 (final) developed and released	September 2024	
M14: Release of all project documentation	September 2024	

Events and consultations

OHOS events and consultations		
Completed		
Event/consultation	Dates & participants	Description
Project start-up workshop	10th December 2021 10 participants – all Labs.	Kick-off workshop to introduce project team in person and discuss project roadmap, including tasks, roles, deliverables, and milestones.
Partner information workshop	21st June 2022 15 participants – all Labs, partner representatives.	Workshop to update all primary partners (DPC, NLS, NLW, PRONI, SSI, Tate, Wikimedia) on project progress and discuss areas of further collaboration. Followed up with individual partner meetings to outline specific tasks.
Landscape scanning survey	Created – June 2022 Paused due to changes in staffing in Archives and History Labs.	First user survey created and ready for distribution (release TBC). To understand what is happening in community archives and other local or family history groups, particularly in respect of digital capabilities.
‘Our Stories, in Our Words: Exploring language varieties in the Our Heritage, Our Stories project’ Presentation at PALA 2023 Conference	6th-9th July 2022 Audience – international, approx. 30	Presentation by Linguistics Lab (Ewan Hannafor and Marc Alexander) at PALA 2023 conference at Aix-Marseille University, France. This talk focused on the linguistic benefits of the project in making accessible the wealth of currently undiscoverable and unsearchable CGDC in the UK, as well as showcasing the project to an international academic audience.
‘Our Heritage, Our Stories: Methods and Models for Working with Community Generated Digital Content’	8th-10th September Plenary session at DH 2022 conference the key UK DH conference, with an	Presentation by Lorna Hughes, Diane Scott, and Ewan Hannafor.

Presentation at Digital Humanities Congress 2022 conference	international audience of researchers and practitioners. Audience – national, approx. 70	
Year 2 Kick-Off workshop	12th-14th October 2022	This workshop was originally intended to focus on how CGDC is (re)used across users, researchers, and producers, and how this will feed into project outputs. However, clear differences in project visions meant this was replaced by a workshop that focused on aligning understanding of the project aims, objectives, and outputs between the HEIs and our lead IRO partner. This was conducted within a facilitated workshop kicking off our second year on the project, aligning project vision and shared understanding of project goals/outputs.
‘Integrating language varieties in the Our Heritage, Our Stories project’ Presentation at Data & Digital Humanities 2023 conference	8th-10th March 2023 Audience – international, approx. 15	Presentation by Linguistics Lab (Ewan Hannaford and Marc Alexander) at DDHUM 2023 conference at University of Minho, Portugal. This talk focused on the challenges and benefits of integrating diverse linguistic varieties into institutional collections, drawing on work on OHOS as a case study.
‘Linguistic variation and institutional collections’ Presentation at Language & Power 2023 (LAP2023) conference	22nd-24th March 2023 Audience – international, approx. 30	Presentation by Linguistics Lab (Ewan Hannaford and Marc Alexander) at LAP 2023 conference at University of Münster, Germany. This talk focused on the potential impacts of diverse linguistic varieties in institutional collections on linguistic prejudice and language equity, drawing on work on OHOS as a case study.

Presentation at 2023 History and Archives in Practice Conference	29th March 2023 Audience – national, approx. 30.	Presentation by History and Impact Lab (Hannah Barker and Stefan Ramsden) at HAP2023 at Institute of Historical Research, London. This talk focused on the role of CGDC in historical research.
Forthcoming		
Event/consultation	Dates & participants	Description
Archives workshop series (replacing CGDC producers workshop)	Delayed – This targeted workshop series was postponed due to staffing changes in the Archives and History Labs, and activities have now started (as of March 2023).	A series of targeted workshops organised with a range of groups and around specific issues including disabled and neurodiverse people, race and ethnicity, LGBTQI+, and family historians.
Remixing workshop	November 2022 Delayed – This workshop series has been delayed until the prototype remixer suite is ready for testing.	Exploratory workshop looking at varying approaches to the ‘remixing’ (i.e. the presentation, analysis, and comparison) of CGDC across diverse archives, materials, and audiences.
Historian engagement workshop	This work has begun working with Manchester Histories and local and family historians. Work with academic historians will commence after the summer, following a period of extending industrial action in the university sector.	Collaboration-centred events involving local and academic historians, developing project networks, furthering co-design approach, and informing design of historian-focused elements of project.
Post-custodial approaches workshop	September 2023 Will be later than planned due to PI time spent on cross-project project management and partnership working, and vacancy in archives lab.	Workshop developing project’s post-custodial model for the curation and management of CGDC, facilitating input from diverse stakeholders.
AI pipeline: Producers & users workshop	September 2023	Aimed at helping to refine automated approaches to ingest and enhancement of CGDC, through discussions with users and producers around how they create, manage, and use these materials.

Ethics and data sharing models workshop	October 2023 Tbc – will be developed in partnership with other Discovery projects.	Workshop developing the ethical approaches of the project in collaboration with holders of community materials, ensuring research and use of data is not exploitative and that the project contributes to communities.
Language data and wikidata workshop	November 2023	Feeding into the integration of further linguistic resources into the AI pipeline, this workshop will explore the use of language data in enhancing automated approaches to the enrichment of CGDC, and how this links to existing models, such as wikidata.
ALT: Post-custodial toolkit workshop	June 2024	Showcase of post-custodial toolkit in collaboration with ALT, enabling further refinements to this framework before final project release.
BYO CGDC datathon	July 2024	Bring Your Own datathon using final versions of AI pipeline and Observatory, allowing members of the CGDC community to test these on their own data to visualise, analyse, and compare their CGDC materials with other resources.
Closing symposium	September 2024	Closing symposium showcasing the final versions of the Observatory, AI pipeline, and post-custodial model, as well as reporting on further project outputs.

Research approach

Overview

Research and development on *Our Heritage, Our Stories* is distributed across five Labs: AI, Linguistics, Archives, History, and Observatory. These Labs operate semi-autonomously, with distinct teams and approaches, but with continual and overarching collaboration to deliver the project's key goals and milestones. We have an integrated approach to project design, with our guiding methodologies being iterative co-design, active research, co-production, and continuous evaluation and validation. Through this approach, *OHOS* will link a wide variety of CGDC content and metadata – including catalogue data, born digital content, and textual data, and image, moving image, and audio metadata – in order to enhance digital search and interoperability and make these resources available to a fresh range of audiences and users. To do this, we are using cutting-edge machine learning and AI approaches to discover and harness CGDC data and metadata currently invisible to other communities, researchers, institutions, and the general public. Data is being collected and processed in a series of tranches according to our content and data strategy, collecting and processing data of increasing complexity. Meanwhile, our work on language and dialect will enable us to engage with the complexity of CGDC data and metadata on their terms, refining generic methods so as to more appropriately and comprehensively capture the context-dependent richness of CGDC. Our Observatory will offer these materials to the public, as well as other researchers, through an innovative interface that incorporates a “remixer suite” of tools for the analysis, comparison, and reuse of CGDC. The design of this Observatory is being iterated through collaborative discussions with academic researchers, community groups, and institutional archives. This work is being conducted across our Observatory, History and Archives Labs, enabling us to holistically scope key ethical, structural, and logistical considerations and incorporate these factors into the design of our outputs for the most effective use and reuse of CGDC. Data flow through the project is organised in accordance with the project's Data Strategy (Annexes C and D), which is organised to ensure optimal Lab input into each stage of data ingest and processing – this procedure is being iteratively refined as we work through project data, to optimise and develop a tested model of best practice for the integration of CGDC into institutional settings as a project output.

Project Labs are organised centrally by the project management team at the University of Glasgow (consisting of PI Hughes, Deputy PI Alexander, and PM Hannaford and a Project Manager to be appointed) and the Senior Agile Delivery Manager at TNA (Hazel Jell), to facilitate this cross-Lab collaboration. Regular project meetings should occur between RAs and RSEs each Monday, to enable cross-project oversight and knowledge sharing between Labs. Investigators' meetings are organised on an ad-hoc basis around investigator availability, occurring approximately once per month, to steer overall project progress. The project team has a shared Slack workspace which enables quick communication across Labs. RA's and RSE's use Slack to share their weekly plans, improving awareness of activities across Labs. Our project Advisory Board is available to advise on progress and overall strategy, provide operational oversight, act as a positive feedback loop, and advise on broader applicability of outputs, ensuring value/impact beyond the project. This board has two sub-panels: the Observatory Reference Panel, featuring key national and international GLAM

organisations and other stakeholders, contributors, and beneficiaries providing advice on matters relating to digital heritage, digital methods, and the Observatory's presentation to the digital heritage community; and the Community Content Reference Panel, the forum through which diverse community heritage organisation voices can be heard throughout the project, sharing their priorities and perspectives on the usability and effectiveness of the Observatory.

AI Lab

Natural Language Processing (NLP) is a sub-area of Computational Linguistics or, more generally, Artificial Intelligence, that concerns the automated and computerised processing and analysis of unstructured textual data at scale. As such, NLP offers a suite of methodologies for the extraction, analysis, linking, preservation, and discoverability of the knowledge contained in CGDC, often in purely textual form, that is currently disconnected from mainstream collections, and so NLP forms the principal approach of the project's AI work (complemented by the work of the Linguistics Lab). In adopting this approach, *OHOS* aims to overcome a barrier that most community content contributors are currently faced with when trying to integrate their content into wider datasets: the need to invest time and effort into transforming their collections into formats that conform to the parameters of existing data capturing mechanisms and models. This approach not only erodes the complexity and heterogeneity of CGDC, thereby undermining its value, but it is also unfeasible in practical terms for many community organisations because they have limited funds, resources, and time available to expend on wrangling data. Instead, *OHOS* will enable community organisations to submit their collections for connecting to similar resources as they are, relying on the employed AI pipeline to extract and link meaningful and relevant meta-data.

To achieve these aims, the NLP pipeline of the *OHOS* project is working on performing multiple tasks in relation to two main technical directions. In the first technical direction, information extraction and semantic enrichment will transform purely textual data into Knowledge Graphs (KG), by extracting and disambiguating entities of interest and relations between them. In the second technical direction, these KGs will then be instrumental to enabling advanced similarity-based search possibilities on this enriched content, extending the discoverability and explorability of CGDC and facilitating more advanced analysis of this data. Transforming purely textual collections into linked Knowledge Graphs requires several stages of processing: first, central entities in the content are identified (Named Entity Recognition); next, these entities are linked to public knowledge bases, thus effectively disambiguating them from other possible entities (Entity Linking); finally, any relations between these extracted entities are inferred (Relation Extraction). This combination of extracted, disambiguated entities, their canonical identifiers, and extracted normalised relations, allows the underlying data - CGDC, in the case of *OHOS* - to be transformed into corresponding, interlinked Knowledge Graphs. In this way, Knowledge Graphs offer a way of finding and discovering explicitly similar items, which will be complemented by further automated methods that enable materials with similar semantic content to be identified. In *OHOS*, these Knowledge Graphs will then facilitate the discoverability and analysis of contributed CGDC, by allowing complex querying of its content. For example, historians might be interested in all texts that mention specific people, events, places or date ranges, and will be able to utilise the extracted entities to investigate these items of interest. Furthermore though, the identification of common entities and relations across different sources will automatically link together previously disconnected collections, which will facilitate their

exploration and discovery; for example, a user initially interested in one specific collection might be recommended similar items from different collections that are also of interest, due to entities identified in their descriptions being related in the underlying Knowledge Graph.

Data-driven methods, such as state-of-the-art NLP, also require labelled data to learn to perform tasks. For example, to learn to extract named entities, AI models first need example sentences with manually annotated entities that demonstrate desired results. To this end, *OHOS* takes an iterative approach to developing its AI pipeline, including data of varying complexity and from varying members of the contributing communities at different stages. In addition to providing vital annotations to refine the developed models, this also constitutes a human-centric approach to the development of AI methods, which will ensure that the requirements of the communities are reflected in the underlying methodologies of the project. Furthermore, all information extraction tasks, such as the proposed semantic enrichment of CGDC, mandate a careful balance between precision (among the produced extractions, how many are correct?) and recall (among all true extractions, how many were produced?). Different considerations regarding this trade-off apply, depending on the eventual applications of this extracted information. For example, a historian might only be interested in items linked to specific date ranges that were recovered with high precision. Conversely, an enthusiast might want to recover all possible connections to their item of interest, such as their hometown, even if some of these turn out to be irrelevant. Balancing this intricate interplay, the *OHOS* AI pipeline and Observatory interface will facilitate different bespoke use cases, for example, allowing users to filter the constructed Knowledge Graphs of entities and relations by different metrics/criteria, or selecting different subsets of source materials, enabling the underlying CGDC to be used effectively by as wide a range of audiences as possible.

Linguistics Lab

While the state-of-the-art NLP approaches described above have been applied to various subject areas, they have not been widely employed in the general context of digital archives or, more specifically, to the preservation of CGDC. As a result, our AI research is also in collaboration with the Linguistics Lab to develop and refine these approaches for their application to such materials. With NLP tools typically trained on extremely large datasets, the smaller datasets of CGDC collections and the diversity of language contained within them (often incorporating regional and social language varieties) poses a challenge. To ensure these voices are appropriately represented, and to avoid perpetuating the exclusion of non-standard materials from mainstream collections, the automated processing methods used on *OHOS* will be refined by using models trained on multilingual data and incorporating further specialist linguistic resources, such as the Dictionaries of the Scots Language.

The Linguistics Lab is also conducting research on the most effective means of integrating and promoting linguistic varieties alongside data in institutional frameworks, such as in TNA's Discovery platform. Linguistics research into this area will contribute to approaches to incorporating and opening up materials contained in community collections that are written in, or represent, non-standard, diverse linguistic varieties.

Archives Lab

The focus of the Archives Lab is to ensure that there is a balanced, representative and comprehensive range of CGDC available to the project. This is integral to all aspects of project development (including AI, linguistics, and the Observatory design), and to ensure that CGDC is useable in new and innovative ways as a final project outcome. To do this, the Archives Lab is working actively with the key heritage and collecting organisations, especially those identified as the project's key "collaborating organisations" that are creating, managing, or disseminating CGDC, as well as the extensive stakeholder community working with this material. In our first year, consultation was carried out with CAHG Scotland, Manchester Histories, and Leeds City of Culture (LCC) 2023, in addition to existing project partners. In summer 2022, the Archives lab scoped over 100 CGDC collections. Now that a new RA is in post, the Archives lab is collaborating with the History Lab in working with a selection of these collections to extract metadata for processing by the AI pipeline.

The Archives Lab is deploying a mixed methods approach to develop a deep understanding of the ecosystem of community content – including its creation, description, and management, in close collaboration with heritage organisations and community groups. Understanding this ecosystem is a building block for the Observatory and AI Lab development and outputs, working in an iterative process with our project partners and our wider network of community archives and archival practitioners. The main goals of the Archives Lab are: scoping of CGDC and its ecosystems to an institutional/organisational taxonomy; scoping a semantics-aware, FAIR-compliant and sustainable post-custodial model for use across the sector; development of a data and collections strategy for the project; and scoping, development and documentation of a post-custodial model for CGDC.

As soon as our Observatory interface and remixer suite are delivered (including prototypes to enable co-creation), we will be able to demonstrate to archives and the community the value and impact of making CGDC accessible through OHOS, driving development of a greater number of hybrid and diverse stories. When our project has the required metadata and data support, we will also be able to map the 100+ collections of CGDC we have scoped to community-wide TNA's data models. In addition, when the required infrastructure that needs to be built for further engagement with communities and to showcase the use of their data has been delivered by TNA, we will be able to demonstrate how it can be repurposed in the ways intended in the bid, and further engage the community

The Archives Lab is also developing the project's content and data strategy. Content and data in scope is structured in four tranches, in order to be accessible for iterative development stages by the other Labs:

- Tranche 0: content already available in TNA's Discovery, through Manage Your Collections, to develop awareness of existing good practice in data structures, metadata, and content description.
- Tranche 1: data with which we have close connections internally or via partners, to ensure we can perform manual checks and provide input to the AI process intensively, quickly, and with high levels of confidence.
- Tranche 2: data gathered through our wide network of partners and interested organisations, including multilingual data

- Tranche 3: highly-unstructured data from new sources discovered and scoped via our community engagement.

In addition, the Archives Lab is undertaking ethical and meaningful co-curation with community archives to generate best practice frameworks and effective training resources which can be embedded into archival education and professional development for working with CGDC. We have reviewed existing community digital practice, including guidelines for CGDC development, with the Digital Preservation Coalition and Manchester Histories for further updating, and we plan to also have them reviewed by TNA's Archives Sector Development team.

History and Impact Lab

This Lab is scoping the content, tools and methods required for the project, and framing key research questions for a series of research demonstrator case studies. These studies are already showing the value of CGDC for research across multiple disciplines, engaging academic, community, and family researchers in the use and potential of discoverable and linked CGDC, including generating crucial links between the creators and potential end-users of newly opened up CGDC. Once the TNA has shared the Observatory interface and remixer suite of tools, these demonstrators will use the Observatory to remix and reuse CGDC, working with an emerging community of practice (scoped as part of our Engagement Strategy). The Lab is producing studies of focused periods and times. Examples include research on areas such as the records, artefacts and oral histories of post 1950 migration into the UK to reveal compelling new narratives about the histories and development of contemporary society, and telling new and richly detailed stories about the community histories of disability.

The History and Impact Lab is also actively working with the National Lottery Heritage Fund to gather examples of existing CGDC funded in the past 15 years, and working with the AI Lab and the Archives lab to use NLP methods to uncover the data.

Collaboration between Archives Lab and History and Impact Lab

The History and Impact Lab is collaborating with the Archives Lab to develop the project's community engagement approaches, allowing us to work with diverse stakeholder communities who have different priorities, interests and needs. The Archives Lab is identifying and collaborating with a number of independent community-produced digital archives that record, safeguard, and make known records, artefacts and oral histories of migration, religious and ethnic identity, and the social history of post-war UK at a local level. Our approach is flexible and responsive, centring the requirements of each community: rather than imposing a one size fits all model, we are working with each community to understand their needs before developing an offering of training, resources, and support. This work has two key aims:

- To better understand the wealth of CGDC that is currently held within communities and how communities use, or would like to use, this material. Through working with communities, we will provide training and resources to help communities connect and enhance their CGDC, with their informed consent.

- To investigate barriers and accessibility issues, developing recommendations for addressing these. This will include disability and neurodiversity, as well as any barriers due to class, race, gender and sexuality.

We are working with communities separately, in order to serve them most effectively, with bespoke and tailored workshops. We are also carefully considering hierarchies and power dynamics, developing a post-custodial framework for curating and engaging with community-generated digital content in a collaborative, sustainable, and ethical manner. Travelling to a community venue rather than expecting groups to meet us in the large institutions where we work (universities and national libraries) can help to give some power back, and it demonstrates that we value each contribution. We have identified key groups to work with in order to provide best-practice case-studies for working with community digital archiving groups. The next step is to think through the ethical framework for this work: how do we ensure that all participants feel that their contribution is valued, that no avoidable harm is caused, and how do we approach informed consent (i.e. making sure participants understand how their data will be used)? In this, we will draw on the expertise and experience of our partner organisations.

Within this framework, we have developed an important collaboration with Manchester Histories, part of Archives+, who have supported our scoping of community organisations developing GCDC in the North West of England, and developed a report on this work. This will be the basis of our next phase of collaboration with Manchester Histories, which includes community engagement (supported by our community fund) with key organisations identified, providing content for our pipeline and remixer tools (when they are completed), adding to the sources that will enable new stories to emerge.

Observatory Lab

The Observatory Lab's overarching approach is to iteratively deliver prototypes for the Observatory interface, cross-searching functionality, and remixer tools, which will build towards the final project Observatory for the public to discover, visualise, and compare CGDC with The National Archives' Discovery catalogue.

During development, TNA chose to change their approach and shift from creating three separate prototypes (detailed in our first report), to progressively developing one prototype. Initially, this prototype is based on requirements gathering within the academic partners on the project but will also address the needs of casual and community users, with search, filter and refine across CGDC and The National Archives' Discovery catalogue. Other prototypes may emerge during the project, if time and resourcing allows.

The Observatory is following an agile approach to development. This involves taking an iterative approach to exploring and developing solutions. The project's Research Software Engineers (RSEs) at TNA work in time-boxed iterations (sprints), of three weeks, to deliver a set amount of work (user stories). For *OHOS*, we have story types to distinguish between development and research, to break work into manageable chunks or specific questions to answer. The agile approach enables frequent delivery and feedback opportunities within the team, and the ability to shift activities depending on the outcome of research and user engagement activities.

We are also scoping relevant linked data approaches, analysing interfaces based on linked data both within and outside the cultural heritage sector. We have also been looking at existing projects that use timelines, maps, and annotation tools, and begun our user research to understand our potential users. This work has allowed us to draft personas, user journeys, and interface designs. Research activities are also following the agile approach, defining research questions to be answered during a sprint, using story point estimation (an estimate of the complexity of a piece of work) to guide the required depth of the research. Engagement surveys and co-design workshops will be undertaken with potential user groups (e.g. community archive groups and academic researchers) to gather requirements and design tools to meet their needs.

During the design and development process, we are undertaking technical research to identify the most suitable technical solutions. We are scoping technical solutions used in the cultural heritage sector and beyond, based on principles of using open-source software where possible while following software engineering best practice.

Research results

The automated approaches to interpreting and integrating CGDC into a unified collection are in ongoing development, though early results from our prototype pipeline suggest that the current approaches of entity recognition, linking, and relation extraction can be successfully applied to CGDC. Working on a small sample dataset, our pipeline is currently able to identify key entities within free-text CGDC, link these with canonical resources (where appropriate), extract relations between these entities, and construct these entities and relations into a Knowledge Graph for further exploration. While the utility and accuracy of these results is limited at this early stage, our next stages of the project will iteratively finetune and improve the pipeline's underlying methods, including via integrating further data and resources into the pipeline, so that the extracted entities and relations are increasingly relevant and salient to the CGDC being uncovered.

As our methods become more precise, and results more meaningful, we will also be expanding upon our existing tools for the presentation and analysis of findings through the *OHOS* Observatory. This will provide users with intuitive ways to explore CGDC and tailor these explorations to their specific inquiries, such as through the construction of custom knowledge graphs based on key entities, metrics, or data sources.

The Observatory Lab have changed the preferred database from Blazegraph, instead opting to use Amazon Web Services Neptune graph database. This option is more secure and integrates more easily with other cloud infrastructure in use. RDF-star is no longer used as the data format for the enriched CGDC, unblocking the ability to use the Neptune option. We have evaluated various JavaScript frameworks for UI development, and chosen to use React for the temporary interface. A Content Management System (CMS) will be implemented for the longer-term interface and access to other Observatory features and Remixer suite. The use of a CMS will allow the creation of an intuitive Observatory interface more quickly than development from scratch.

Research into the existing landscape and interfaces based on linked data has highlighted engaging interactive elements to allow users to interrogate the data, which we expect will need to be a key feature of the Observatory. We carried out user interface (UI) research amongst the project team to identify features and use cases envisioned. Our next challenge will be moving from tightly scoped subjects and from aggregated data to understanding how we could discover and work with data 'in the wild'. Over the course of the project, we will address issues of data aggregation, distributed data approaches, dynamic data models, and creating/applying specialised tools to generalised data, in order to develop a clear, cohesive interface for diverse CGDC that enhances the utility and usability of these resources, across the widest possible range of audiences and stakeholders.

The Observatory Lab have investigated data model design options for *OHOS* enriched CGDC data. [Project Omega](#) is developing a pan-archival linked data catalogue at The National Archives, part of The National Archives' commitment to reimagining archival practice by pioneering new approaches to description and access. It has created a sustainable, flexible data model that supports complex born-digital and digitised records. The TNA *OHOS* team have reviewed the existing *Omega* data model to identify if any adaptations may be needed to represent *OHOS* data, and a proposed data

model is being tested with samples of CGDC. This may require further adaptations during the project.

Research into the scale of linguistic varieties within CGDC is ongoing, with challenges identified in the automated interpretation of dialectal terms by existing NLP approaches. Collaboration between the AI Lab and Linguistics Lab aims to examine whether the integration of language-specific resources may improve performance of NLP techniques, with current work focusing on attempts to integrate the knowledge contained in the Dictionaries of the Scots Language into the AI pipeline. This has also been recognised as an opportunity for the promotion of linguistic variety within institutional settings, with the research from the Linguistics Lab into the value of such promotion currently in editorial stages as a journal article (scheduled for publication in May 2023). The article focuses on how institutional collections can democratise the production and curation of materials containing regional/social language varieties and so facilitate their integration into collections more representative of diverse linguistic and cultural landscapes.

The Archives and History and Impact Labs have carried out a complex scoping of 100 collections of CGDC. This, and a major report developed by and with Manchester Histories, is feeding into documentation of the ecosystems of CGDC. Our initial explorations into this area have already demonstrated the complexity of ethical, logistical, and sociological concerns involved in the production, curation, and (re)use of CGDC. Working with communities for their invaluable input into this model, *OHOS* has already established several significant links with key community networks. Strengthening these over the course of the project will enable our research to establish a core network of CGDC producers and users. Results from this, and related research by our Archives and History and Impact Labs will be codified into the project's post-custodial model for management of CGDC, a key project output.

The History and Impact Lab has built on this scoping of a large number of community-held collections to develop a series of researcher use cases/history stories that synthesise community-generated content with officially held archives and records. These have been developed as a series of blogs, which are scheduled to be published on the project website. These stories exemplify the type of reuse and remixing that will be enabled through the Observatory website when our remixer tools are launched.

Project outputs

Current outputs

Several data-related outputs have already emerged from the project, with these forming the basis for the iterative development of refined future versions:

- The AI Lab have currently developed V1 of the AI pipeline, a prototype of the automated methods for knowledge extraction and linking from CGDC (Milestone 2).
- Data that has been ingested and enriched using the AI pipeline V1 forms the first iteration of our enhanced CGDC dataset.
- The Observatory team have built the infrastructure with CI/CD pipeline(s) to build and tear down the infrastructure on demand.
- The Observatory team have developed the search across enriched CGDC (OHOS database) and The National Archives' Discovery data, with a temporary internal interface to demonstrate the functionality.
- The Observatory team have swapped the graph database from Blazegraph to using Amazon's Neptune for hosting enriched CGDC, to align with technologies in use at The National Archives.
- A project website, describing the project and providing updates on project activities has been developed, with an appropriate academic domain being secured for this (www.ohos.ac.uk). A project twitter has also been set up (@OHOS_NatColl).
- Our project Data Strategy (Annexes C & D) has been established to describe the process of collecting, processing, and integrating CGDC into TNA's Discovery platform.
- A report on community heritage has been completed by and with Manchester Histories
- Researcher use cases/history stories have been completed by the History and Impact Lab.
- Preliminary results have been generated from a project developed by the AI, History and Impact, and Archives Labs to use NLP to search for CGDC among projects funded by the NHLF.
- The Linguistics lab has produced research exploring the value of linguistic diversity in institutional collections, as facilitated by the inclusion of community materials.
- The AI Lab, History and Impact Lab, Archives Lab, and Linguistics Lab have scoped metadata and documentation practices for unstructured data used internationally to inform the next iteration of the AI pipeline.

Existing project outputs will be extended over the course of the project to include:

- Further and final versions of Observatory (Deliverable 2 – milestones for which have been delayed at TNA) will be iteratively produced as the project progresses, with initial prototypes being developed into progressively more refined interfaces and tools that fulfil the requirements and desires of CGDC users and producers and the general public. Integral to this will be the delivery of the first and subsequent iterations of the project Remixer suite (Deliverable 3 – milestones for which have been delayed at TNA), containing a variety of innovative tools for visualising, aggregating, and comparing CGDC. These public-facing outputs will facilitate public

and researcher exploration of diverse CGDC through a unified platform, empowering users to tell new stories, in new ways, from this rich but currently underutilised resource.

- Candidate set of remixer tools which are being actively researched and experimented with: comparison tool, SPARQL query builder, search text associated with images, visualisation, mapping tools, and use of Jupyter notebooks.
- Fundamental to the development and delivery of our Observatory and Remixer suite will be the increasingly sophisticated versions of our AI pipeline and the underlying datasets of enriched CGDC produced by this automated model, including new and emerging methods for semantic linking and data ‘crosswalks’ (Deliverable 4). Final versions of the project pipeline and CGDC datasets will be made publicly-available upon completion, allowing users to build upon the work of the project to produce custom tools and processing methods for more niche applications, whilst also fully opening up the underlying data and resources to the public.
- The post-custodial model developed by the project (Deliverable 6) will be disseminated across our networks and partners as teaching materials, describing a model of best practice for the creation and curation of CGDC.
- Refinements to the project Data Strategy will be iterative as data is processed, with this strategy complementing the post-custodial model in describing a model of best practice for the integration of CGDC into institutional collections.
- A series of project White Papers will be produced (Deliverable 7), discussing the complex theoretical, practical, and systemic issues addressed by the project. These policy and practice shaping White Papers – aimed at Independent Research Organisations and other heritage organisations, practitioners, funders, and government, and those delivering training to the heritage professions – will be entitled: ‘Saving UK CGDC at Risk’, ‘Metadata Crosswalks Beyond Discovery’, ‘Emerging Approaches to Data Aggregation and Rights’, ‘Post-Custodial Archival Management’, ‘Language as a Heritage Object’, and ‘Ethical Use of AI in Community Contexts’.
- A series of research demonstrator case studies will continue to be produced for the project, building on an existing set of case studies for history, in collaboration with our project partners, to show the value of CGDC for research. These will engage academic, community, and family historians on the use and potential of discoverable and linked CGDC, including generating crucial links between the creators and potential end-users of newly opened up CGDC. These studies will inform our outreach work to explore with CGDC creators and holders the transformational potential of such collections through linked and remixed stories, and make the studies available for future users.

Dissemination and publications

The project has developed a publications strategy and process for approving and peer reviewing publications, outputs, and presentations by project team members (RSEs, PDRAs, Co-Investigators, and the PI). This is to ensure that all outputs are consistent with planned project deliverables, and reflect the aims and objectives of the project.

Project team members have engaged in several dissemination activities, producing the following outputs:

- Presentation at PALA Conference, July 2022: Style and Sense(s) in Aix-en-Provence, France, entitled: 'Our Stories, in Our Words: Exploring language varieties in the Our Heritage, Our Stories project'. This talk focused on the linguistic benefits of the project in making accessible the wealth of currently undiscoverable and unsearchable CGDC in the UK, as well as showcasing the project to an international academic audience. [Conference programme](#). (Ewan Hannaford, Marc Alexander)
- Journal article for a special issue of the Journal of Documentation (focusing on uses of Artificial Intelligence in the provisioning and use of digital cultural heritage collections with restricted or difficult access), entitled 'Our Heritage, Our Stories: Developing AI Tools to Link and Support Community-Generated Digital Cultural Heritage'.
- Poster presentation on technical approaches at the Digital Humanities at Oxford Summer School 2022.
- Presentation at Sheffield DH Conference, September 2022, [Our Heritage, Our Stories: Methods and Models for Working with Community Generated Digital Content](#) (Lorna Hughes, Diane Scott, Ewan Hannaford).
- History and Impact Lab has written exploratory case studies showing how historians might use CGDC to shed light on aspects of modern British history (including historiographical context, potential questions to be addressed, review of CDGC sources 'in the wild' and in institutional repositories). The case-studies use CGDC to explore the 'postmemory' of the Second World War, the history of disability activism, and online musical communities. These case-studies are written in a blog format ready for publication on the project website.
- History and Impact Lab has written use-cases exploring how historians might interact with tools for searching, linking and remixing data.
- Presentation at Data & Digital Humanities 2023 (DDHUM2023) international conference, University of Minho, Braga, Portugal. March 2023 – 'Integrating language varieties in the Our Heritage, Our Stories project" [Conference programme](#) (Ewan Hannaford, Marc Alexander)
- Presentation at Language & Power 2023 (LAP2023) international conference, University of Münster, Münster, Germany. March 2023 – 'Linguistic variation and institutional collections'. [Conference book of abstracts](#) (Ewan Hannaford, Marc Alexander)
- Presentation on history and CDGC at 2023 History and Archives in Practice Conference, Institute of Historical Research, London, March 2023 (Hannah Barker and Stefan Ramsden)

Future dissemination and publications

- The Linguistics Lab has a presentation on *OHOS* at the PALA 2023 international conference, and a journal article in editorial stages for a special issue of the *International Journal of Language Studies* (publication scheduled for May 2023). This journal article will also feed into the upcoming project white paper, 'Language as a Heritage Object'.
- Hughes will present a Keynote at the Oxford 2023 DH Summer School, "One step up: using digital humanities to bring community generated digital content into a national collection".
- The Sheffield DH conference paper will be published in the proceedings of the conference.

Cross-project collaboration

The project has many synergies with other Discovery Projects, and we are keen to engage as broadly as possible around core challenges: there is potential for common debate and development around the ways that participation can define our shared stories, the challenges of language as heritage, the nature of digital heterogeneity, and how we can remix and reuse previously unseen parts of our national collection while empowering communities to share their heritage. We have attended an extensive range of cross project workshops, including the communicating colonial legacies event. We have had a series of meetings with PIs and core staff of the other Discovery projects, identifying key ways to work together and avoid the duplication of effort, and to find shared solutions to challenges faced by the projects. Collaboration informs all our research: our focus is building on existing structures and strengths to dissolve barriers, build research capacity, and foster public engagement. Key areas for future discussion include bringing persistent identifiers to use of unstructured content, as well as the use of IIIF approaches to delivery of CGDC. In particular, we could fruitfully work with *Transforming Collections* to identify common approaches to surfacing suppressed stories about CGDC.

Sustainability and infrastructure

The technical infrastructure of *OHOS* can be seen in Annex B outlining the key data responsibilities of each Lab and the data flow between these. Fundamental technologies supporting the underlying infrastructure of the Observatory are hosted on Amazon Web Services: Neptune, an API gateway, a basic UI app, and a CMS.

Short-term data storage

Data used for processing in the University of Manchester will be stored in their Research Data Storage facilities, which conform with the UKRI Research Data Management guidelines. At TNA, datasets will be stored in a secure cloud hosted graph database with automated backup, and code for the prototype toolkit/interface will be stored in a code repository. TNA uses Amazon Web Services as its cloud compute environment and follows a Cloud First policy in addition to using open-source solutions and open standards, in accordance with the UK Government Technology Code of Practice. Data collected by the University of Glasgow will be stored on centrally managed servers, with nightly backups and multi-location backup storage. Dedicated and secured network drives will be used in line with Glasgow's data management policy.

Long-term data storage:

The final linked and enriched dataset created by the project will be made congruent with TNA's Discovery data store, in accordance with an agreed trust and ownership model and with respect to FAIR Data Principles. Data in the Discovery data store is licensed under the Open Government Licence (OGL) and is accessible via a public API. Data contributed to Discovery which describes material held outside TNA can be edited and/or withdrawn by its holders through the MYC tool. Sustainability of data deposited in Discovery is founded on TNA's commitment to such data (or Discovery's successor portals), which will be hosted by TNA indefinitely. Our databases, documentation, and research datasets will all be stored in Glasgow's or Manchester's dedicated research data and e-print repositories and other relevant stores, such as Figshare. Code, data models, and our public-facing static website will be archived by our university IT systems and repositories in accordance with AHRC requirements, for a minimum of at least three years.

Ensuring continued access and use of digital outputs

Our output data within Manage Your Collections will rely on TNA's core institutional commitment to archiving the data stored within MYC (including its future updates or its successor portals) over the long term. AI and tool development is an area of rapid iteration and improvement, and so for maximum sustained usability and accessibility all code will be made available through GitHub under an MIT licence (which is compatible with the OGL), and documentation under the OGL. The availability of our tools and other relevant outputs on this platform is the most effective means of ensuring sustainability, and that outputs are available for integration into other tools and resources for a national collection.

Interim development towards recommendations

Archives lab will continue to scope the ecosystems of CGDC, which will feed into the development of a post custodial approach for CGDC.

History and Impact Lab and Archives Lab will continue to work with the community groups they have identified to develop best practice recommendations to guide working with CGDC; the recommendations will make special reference to ethical pathways for making accessible and linking data produced by these groups.

History and Impact Lab will undertake consultation based on 'clusters' of producers and potential users of CGDC; these clusters will centre on themes of modern British history that are well represented in the CGDC material and are of interest to historians, including the women's peace movement; local and community history; pop music and pop cultural history; Second World War memory studies; disability activism; black history; migration history.

History and Impact Lab to engage historians in a discussion about the use of CGDC through presentations at key historical conferences (Oral History Society 2023; Family Archives conference 2023; History and Archives in Practice 2023) and preparing publications for key journals (Social and Cultural History; History Workshop).

Linguistic Lab research has identified the potential significance of institutional, and by extension, national, collections as a tool for promoting language equity and preserving linguistic diversity in the UK. Though this research is ongoing, and will be more fully expounded in the 'Language as a Heritage Object' white paper, the Lab is recommending the diversification of linguistic varieties within institutional settings to be an important consideration for attempts to build a national collection. Research on *OHOS* has determined community archives to be an invaluable resource in providing access to linguistic varieties that might otherwise be inaccessible to institutions and researchers.

The Observatory Lab research has identified a number of technical- and human-centred recommendations for a national collection. The Lab recommends using standard cloud infrastructure where possible; the Observatory is being built using Amazon Web Services technologies, including the Neptune graph database. These are stable, secure, and consistent with technologies in use across The National Archives. In keeping with consistency, standards such as W3C should be used where available, such as use of terms for API functionality, to enable easier reuse of code and understanding amongst the technical audience.

To ensure easy interoperability with other Linked Data, it is recommended to use an RDF (Resource Description Framework) graph database, rather than an LPG (Labelled Property Graph) database, as it is flexible, reliable, and the most common choice in Digital Humanities projects. Equally, it is recommended to reuse existing ontologies, such as Omega, Dublin Core, schema.org, and the Web Annotation Ontology.

Lessons learnt through the iterations so far are to begin with data samples and model prior to any UI design and development. Understanding the expected data and data model early will ensure there is

a shared understanding, and avoid future re-work on the UI, leading to more efficient development and delivery of value to users.

In terms of the human-centred approach, it is recommended to involve a broad range of users from a variety of professional, personal and interdisciplinary backgrounds – representing multiple ‘persona’ types – in the design of the data model, interface and tools. This process can create a community of invested users, with a variety of expectations and needs from the end product(s), who can be involved in testing and reviews, and feel ownership of the final product. Inclusion in the design of the data model, focusing on intended uses and questions that would be asked of the data, can directly feed into the technical infrastructure and UI design.

Due to the interdisciplinary nature of the project teams, and for consistency in communications and publications, it is important to gain a shared understanding of terminology across disciplines early in the project. Creating a glossary of key terms that is kept up to date is important to help communication between the disciplines in the project, ensuring these are understood, and understanding where terms may have different meanings to different disciplines.

Contacts

General addresses

Project email – contact@ohos.ac.uk

Project website – www.ohos.ac.uk

Project twitter – @OHOS_NatColl

Principal investigator

Lorna Hughes, lorna.hughes@glasgow.ac.uk

Co-investigators

Marc Alexander, marc.alexander@glasgow.ac.uk

Hannah Barker, hannah.barker@manchester.ac.uk

Riza Batista-Navarro, riza.batista@manchester.ac.uk

Goran Nenadic, gnenandic@manchester.ac.uk

Pip Willcox, pip.willcox@nationalarchives.gov.uk

Managers

Ewan Hannaford, ewan.hannaford@glasgow.ac.uk

Hazel Jell, hazel.jell@nationalarchives.gov.uk

RAs and RSEs

Andrew Bewsey, andrew.bewsey@nationalarchives.gov.uk

Harshad Gupta, harshad.gupta@nationalarchives.gov.uk

Ewan Hannaford, ewan.hannaford@glasgow.ac.uk

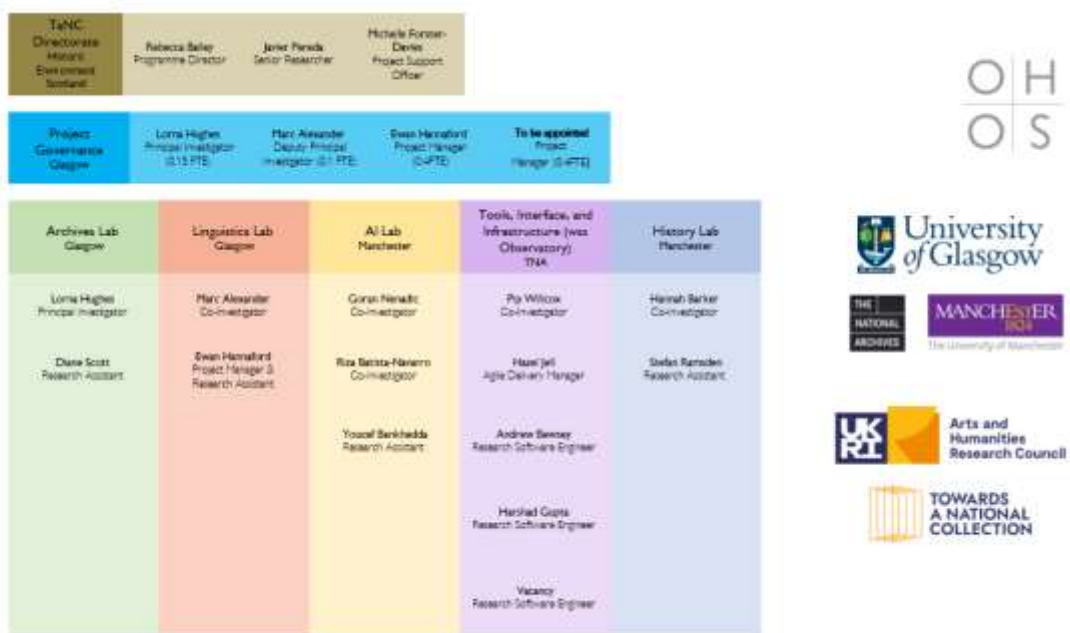
Stefan Ramsden, stefan.ramsden@manchester.ac.uk

Rhiannon Lewis, rhiannon.lewis@glasgow.ac.uk

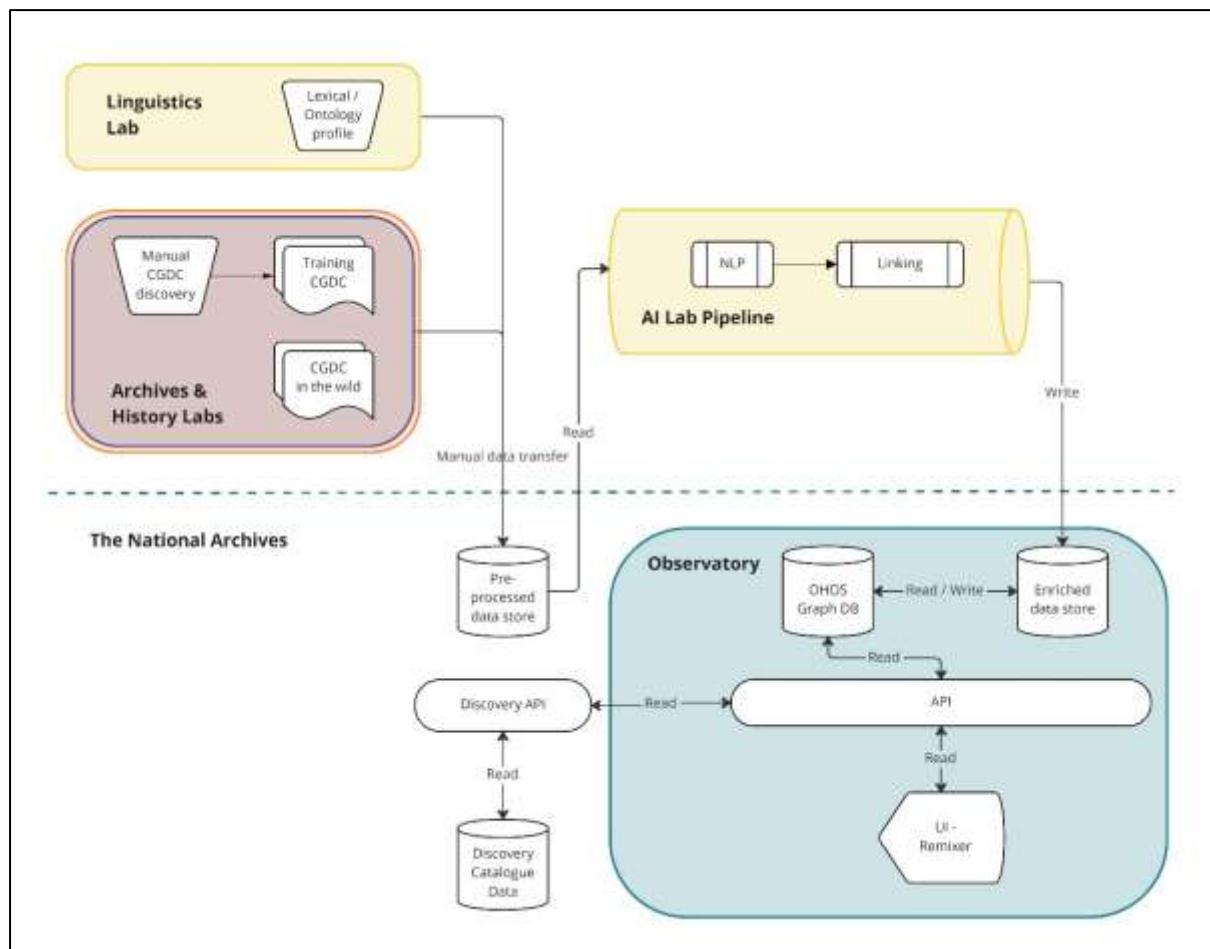
Youcef Benkhedda, youcef.benkhedda@manchester.ac.uk

Annex

Annex A: Staffing structure diagrams



Annex B: *OHOS* technical infrastructure



Annex C: Working project data strategy narrative

1. Selection

Led by Archives and History and Impact Labs, working around case studies and highly targeted content selection, and in collaboration with partners and communities. Barker and Hughes, and archives/history PDRAs, working with partner organisations (NLS, NLW, PRONI, Manchester Histories, and also reviewing stuff in TNA Discovery).

Content is initially identified from established partners and known data sources, with wider community CGDC being integrated in later tranches. Any data in scope for scraping is identified by Archives and History and Impact Labs at this point, to be scraped using methods established at TNA (these methods may also be applicable in any subsequent automated identification and reviewing/mapping further CGDC in the Observatory).

1. Checking and data review/mapping

Led by Archives Lab, with Data Guardian (NB: “Data Guardian” was the description of this post used in the bid, but now configured as a “metadata support post” at TNA which has not yet been appointed).

Requires Data Guardian’s support on metadata and data structures, and mapping metadata to existing standards, e.g. Dublin core? EAD? Mets). History and Impact lab input provided where appropriate. **Archives and History** and Impact **PDRAs** work through a data checklist to identify all descriptive aspects of data in scope for each tranche, including:

- a. Data model?
- b. Metadata scheme used?
- c. Metadata available?
- d. In MYC?
- e. Any hope of getting this accessible via Discovery?
- f. Possible collaboration with data owners – look at collection ownership, ethics of exposure, use as basis for ethics framework
- g. Relevant authority data check (against AFs shared by TNA)

This will be documented for reference across the project.

Data guardian and Archives Lab/History and Impact Lab map data to identify entities that can be linked to existing authority data (names, places, people, events) and ontologies. This process will

also involve discussion (relating to understanding metadata) with partners responsible for that data and with Data Guardian to best understand it.

CGDC is **also analysed by the Linguistics and Archives Labs** to look for patterns, key words, phrases/structures relevant to help text mining, relevant ontologies and closed vocabularies, specialist terms, key characteristics of the language and structure of various data types etc, and anything which will help AI Lab's work. This will produce a list of authority data that will be maintained and expanded by the Data Guardian.

2. Categorise CGDC as tranche 1, 2, or 3

Data is assigned to a data tranche by Archives Lab with input from History and Impact Lab.

Tranches made up of data from known partners (for tranches 0-1) and incorporating more complex and diverse materials in later tranches (tranches 2-3). Unstructured and undocumented community materials in final tranche. Call initial datasets from each tranche Raw Data 1, 2, 3.

3. Import/transfer to AI Lab

Led by Data Guardian. Raw Data 1/2/3 sent to Manchester for importing into processing/enrichment pipeline (this will include both data and metadata files). CGDC now in AI Lab's 'processing' data store (AKA experimental data store).

4. Creation of Knowledge Graphs

Led by AI Lab. Extracting entities in CGDC, linking entities to canonical authority data (wikidata, dbpedia, etc), relation extraction between entities, semantic similarity metrics, storing entities and relations in KG. Supported by Data Guardian on authority data, ontologies, etc based on TNA cataloguing practice and expertise. **Supported by Linguistics Lab** in input into refining methods and incorporating language resources, where relevant. Greater input on reviewing language data for tranches 2 and 3.

5. Transfer of data as KGs to TNA

Led by AI Lab. Co-ordinated with Data Guardian. Processed knowledge graphs of data from each CGDC tranche are signed-off for Stage 7 from AI Lab as meeting quality checks and ready for TNA to use. We call these processed datasets from each tranche Enriched Data 1, 2, 3.

Samples of Enriched Data 1/2/3 undergo internal validation checks by relevant labs and with relevant partners/data holders, providing annotations to AI Lab for use as further enhancing training datasets.

6. Data ingest, integration, and access at TNA

TNA ingest the AI Lab output (Enriched Data 1/2/3) into the project's staging server at TNA (called "OHOS data store" in first stage bid, here called "Observatory staging data store").

Data Guardian works on mapping Enriched Data to TNA metadata standards (e.g. as used for Discovery/MYC data) and works with RSEs to build tools/solutions to automate this mapping process.

RSEs and Data Guardian use these solutions to integrate Enriched Data to make it work with Discovery and existing TNA infrastructure, such as MYC. The output of this process is ingested and integrated datasets for each tranche: we call these datasets Integrated Data 1, 2, 3.

In parallel, **RSEs** at TNA scope and develop 'remixer' tools to work with Integrated Data (e.g. advanced search, comparison search, entity-based search, filtering, Keyword-In-Context, etc). UX/Interface RSE makes it viewable and linkable via an Observatory interface (see Observatory workplan for further information).

7. Review and feedback

Once data is integrated and usable through the Observatory interface, **Data Guardian** checks output and feeds back to Archives Lab and AI Lab metadata recommendations for subsequent Enriched Data, identifies missing metadata for linkage via Archives lab, continues to map metadata. Samples of Integrated Data 1/2/3 are selected and sent to AI, Archives, History and Impact, and Linguistics Labs for feedback.

Samples of Integrated Data 1/2/3 are reviewed by relevant labs and with relevant partners/data holders, providing feedback on the way the data is discovered, formatted, and linked.

Archives Lab coordinate external collections partners to look at their data via the Observatory interface and feedback comments/improvements. This is an opportunity to discuss ways they deal with their own data and suggest enhanced approaches to making CGDC available online, so that it is more discoverable.

AI Lab work on annotated sample data, and this will be iterated with Glasgow colleagues and data partners to improve understanding and/or revise sample/training dataset. AI Lab iterate process after working with sample data, working with Glasgow (Archives/linguistics) and TNA (data guardian).

8. Research on data by History and Impact Lab

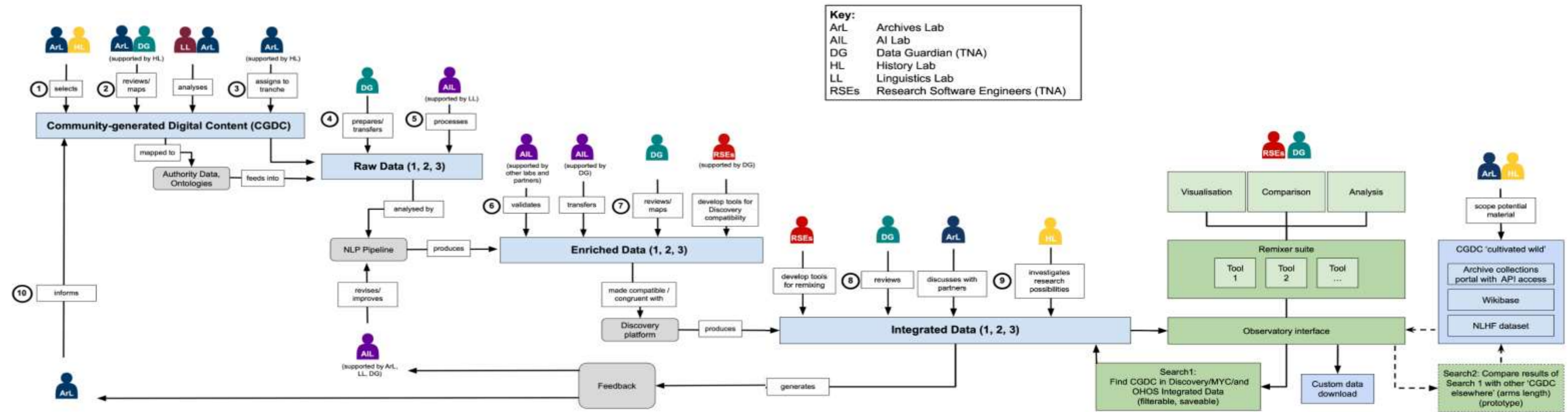
History and Impact Lab investigate research possibilities uncovered within Integrated Data 1/2/3, using tools made available in Observatory.

9. Research on post-custodial by Archives Lab

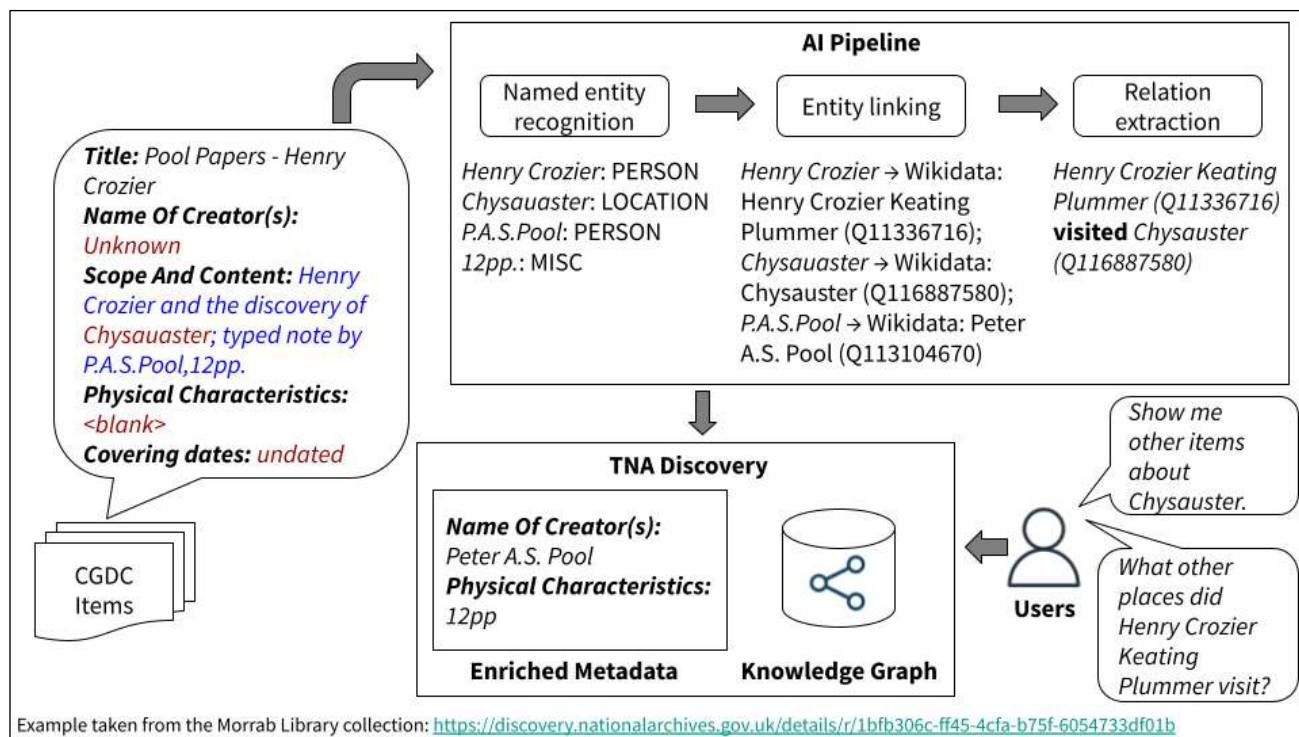
Archives Lab investigates the ecosystem of CGDC in each tranche to inform subsequent development of new models for post-custodial development and management of existing and new CGDC. Archives Lab also engage with a wider network of collections holders to explore further tranches of data, feeding into the selection of data for the next tranche.

10. Repeat process for each tranche

Annex D: Working project data strategy diagram

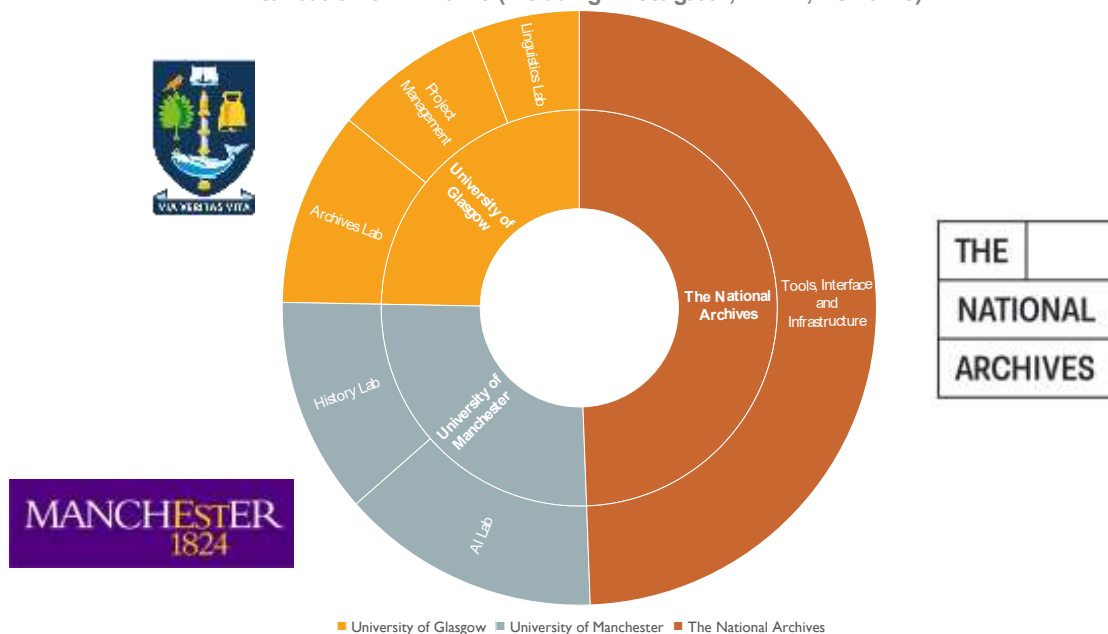


Annex E: Visual Depiction of the AI Pipeline



Annex F: Resource allocation

Distribution of FTE time (including investigator, PDRA, RSE time)



Annex G: CGDC Story example

CGDC and Historical Use-Case 2

OHOS History and Impact Lab, with Archives Lab

CGDC stories have been developed by the History Lab working with the Archives Lab, to understand historical data gathering and research journeys, and how these might be mapped and replicated in project outputs, especially the remixer and observatory interface features. The following is an example of these stories.

Mary Elizabeth Eppynt Phillips – researching an individual

Historians, whether amateur family historians, biographers, or academic historians will often wish to locate a wide range of sources relating to an individual, whether well-known or obscure.

The following Use-Case describes steps that a historian might go through in searching for source material on an individual, using Mary Elizabeth Eppynt Phillips as an example.

Mary, the first woman doctor to graduate in Wales, is relatively well-known – she has a Wikipedia entry, and a number of websites include profiles of her (see for example, Llangammarch Wells Local History Website <https://www.llangammarchhistory.co.uk/local-people-of-interest/dr-mary-eppynt-phillips.html> and Breconshire Local and Family History Society have written a short book about her: *Dr Mary Elizabeth 'Eppynt' Phillips 1874 - 1956: A Pioneering Woman from Brecon* (2018).

Nonetheless, there is more to be researched about her life, and her story touches on a wide range of themes that interest academic historians, including women in medicine, the women's suffrage movement, medicine in wartime, the Voluntary Aid Societies. A historian wishing to write an authoritative entry for Mary in the Oxford Dictionary of National Biography (she is not currently included) would need to refer to primary sources.

The following steps would not necessarily be conducted in this order – much would depend on the level of existing knowledge, and in practice there would be some overlapping of the steps – while it's a good idea to read around the context of a research theme early on, there's not much point in putting a huge amount of time into that if there aren't many primary sources, so scoping out the primary sources and reading around the topic will often go hand-in-hand.

Imagining that I do not have the time or resources to visit physical archives, in order to research Mary's life for the ODNB, I would start by reading the existing secondary material online. This would allow me to construct a rough chronology of her life and career that would enable more targeted searching (for example, I would learn when and where she was born - Merthyr Cynog, 1874 - where she studied - the University College of South Wales and Monmouthshire, the Royal Free School of Medicine). The secondary material would also alert me to the organisations she was part of (British Gynaecological Society, National Union of Women's Suffrage Societies). I would construct a timeline of Mary's life with as much detail as possible using this secondary material.

I would then turn to the Observatory to look for primary sources to help me confirm and flesh out this chronology. I would conduct searches using various possible iterations of Mary's name (Dr M.E. Phillips, Mary Phillips, Mary Eppynt Phillips etc), as well as conducting some searches relating to her educational institutions and places of work.

Hopefully, the Observatory would bring me the excellent repository of CGDC on Mary that her niece had digitised and uploaded to the People's Story Wales (figure 1)

(<https://www.peoplescollection.wales/discover/query/mary%20eppynt>)

This material comprises 18 documents, including letters, newspaper clippings, photos, a very useful CV document with testimonials, and a scrapbook. I would relatively quickly scan through this material, getting a sense of what it included, and download the documents (Figure 2). I would keep them all in a file on the computer. The Observatory search would hopefully also bring me the medal cards for her, held in the TNA collection, which can be downloaded (WO 372/23/32919 and WO 372/23/32920) (figure 3). These documents list her as being in France and Malta, give her 'Qualifying Date' as 1914, her corps and rank as the 'French Red Cross Rank: Doctor'. The document shows that she received both Victory and British medals. At present, TNA Discovery does not bring me any relevant digitised collections from elsewhere, via Manage Your Collections; a quick search of other relevant repositories does bring some material. The Wellcome Library, for example, has a medical book written by a Mary Elizabeth Phillips in 1931, *Elementary biology for matriculation and allied examinations*. We know Mary was writing medical books at this time, and it seems likely that this is her work, but further research would be needed to ascertain if this book were by her or a namesake. A very useful source, which has been used by most online authors writing about Mary, is Welsh Newspapers online (<https://newspapers.library.wales/home>) which can be searched through the National Library of Wales. A text search in this database quickly finds a number of articles on Mary, from which we can obtain biographical information and details on her interests and activities. Some of the articles also quote her directly. I would, as far as possible, download and save all of this material into files.

Next, I would search for academic literature to plug gaps in my knowledge in relation to Mary and the themes her life touches on – women's work and education around the turn of the 19th/20th centuries, women in medicine, Voluntary Aid Societies in the First World War, the Suffragette movement. This would also help me to find interesting angles and debates that researching Mary's life might illuminate. I would search for this literature using Google, Google Scholar, the Royal Historical Society's Bibliography of British and Irish History, academic library catalogues, and search through back issues of relevant academic journals (eg *Social and Cultural History*, *Twentieth Century British History*, *Social History of Medicine*). This search would also reveal whether there was any existing academic work referencing Mary.

I would then turn to read the documents I have collected through the Observatory search in more detail, adding to the chronology I have already started to construct. I would annotate this with questions about gaps in the chronology, and observations about how different parts of her life might be further explored in dialogue with the wider historiography. I would also make notes on my reading in a word document.

In order to give wider biographical context, I would use Ancestry.co.uk, to search for Mary in the Censuses, finding out where she lived and a little more about who she lived with, and about her parents.

My primary and secondary research would expand into examining the contexts Mary was part of, the educational institutions, organisations such as the British Gynaecological Society, the National Union of Women's Suffrage Societies (Figure 4 and Figure 5), the Scottish Women's Hospital for Foreign Service, the French Red Cross. I would be interested in Mary's social networks, and would use all of the above techniques to find out more about Mary's family, her work associates and the people she had contact with through her work and voluntary activities. Mary's geographical movements have some bearing on this – she moved a lot for work (she lived in Nottingham, Leeds and Stockport, at least, in a period of few years), and may have come into contact with new influences as a result. Attempting to reconstruct Mary's social milieux would help relate Mary's story to wider social history.

In terms of processing and remixing the material, for this particular project the ability to link the material to a timeline would be good (as in the Queensland Architecture website). The facility simply to download and to save material, or to save it in an online location, would be also useful. As the research expanded to encompass Mary's various social networks, the ability to visualise Mary's social contacts might also be useful (as in the 'Beyond Notability' network visualisation).



Figure 1

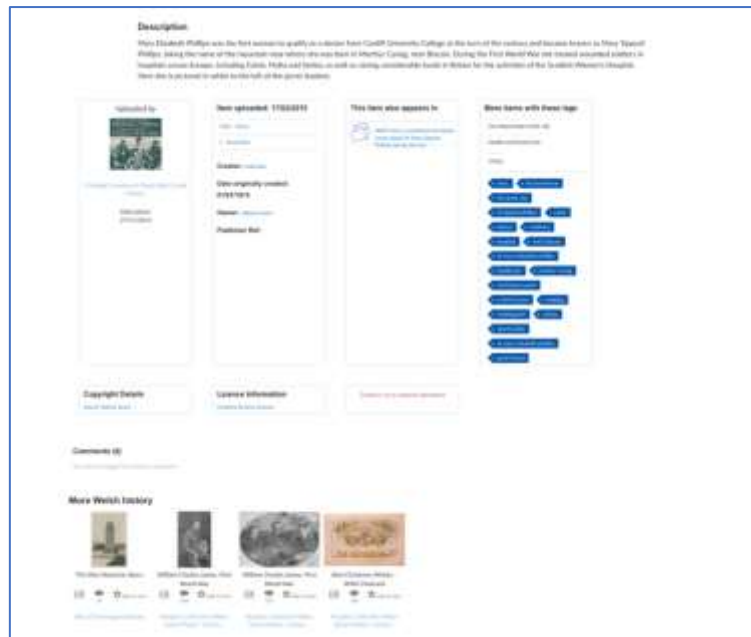


Figure 2.

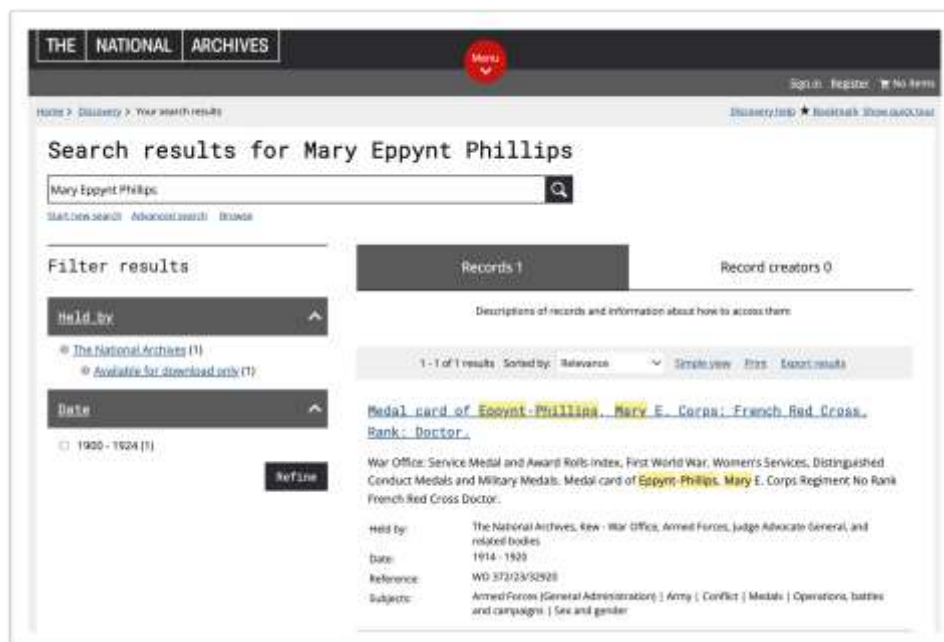


Figure 3.

Annex H: Y2 Kick-off and facilitated session, October 2023

Our Heritage, Our Stories

Year 2 Kick-Off Workshop Agenda



Hilton York, 1 Tower St, York YO1 9WD

12th-14th October

Description

This project workshop is intended to provide the team with a full and comprehensive understanding of Our Heritage, Our Stories, and the project's structure, objectives and deliverables. It is also intended to ensure all team members are certain as to the roles and responsibilities which exist across the project, and to determine a clear understanding of the expertise, knowledge, and relevant experience of each team member. Following the workshop, it should be clear what tasks are being completed by each lab, and by each team member within these labs, and everyone should know who is responsible, and is the main contact, for each element of the project.

Our clear and open shared discussion of roles and responsibilities will also establish a comprehensive understanding of the dependencies and requirements of project work across labs, so that we collectively know what is needed by each lab at different stages of the project and what effects delays in one area may have on others. This will allow us to figure out how we will deliver the project goals in the remaining time available, as well as who to speak to if there are challenges/issues in a particular area. We won't be discussing the details of particular labs or technical elements of the project in this workshop - this workshop is focusing on providing a higher level, shared understanding of project aims, roles, and timelines.

We will also collectively establish a shared framework for communicating and reporting across the project, so that we have a standardised, well-understood, and transparent system for sharing information and resources across the project. Going forward, this will enable us to see our progress on each required deliverable and address challenges hindering progress more efficiently. The outcome of the meeting is a roadmap for all future project activities and communications, with a clear indication of the people responsible for each task.

Intended Outcomes

- Clear understanding of the tasks to be carried out in each workstream, and the person responsible for each.
- Clear understanding of dependencies and requirements across workstreams.
- Agreed methods for communications and reportings, and "closing the circle" on assigned tasks.
- Roadmap for all future project activities and communications, with a clear indication of the people responsible for each task.

Schedule

Wednesday 12th October

Half day, afternoon only. Lunch not provided.

12:30	Arrivals, for 13:00 start
13:00-13:20	Team-wide introductions and icebreaker
13:20-13:50	Project Vision: overview of the project, its background (funding and TaNC programme, including cross-programme collaborations and communications), and success and failure metrics (LH)
13:50-14:10	Research case study: example of project vision in practice, demonstrating why this matters and what it all looks like in use (HB)
14:10-14:20	Deliverables: Overview of key project deliverables (MA)
14:20-15:00	Break (refreshments provided)
15:00-15:45	Roles and responsibilities: Discussion of roles and responsibilities document (pre-circulated and printed), detailing core roles and responsibilities of project labs and individuals within them (LH)
15:45-16:45	To-do lists: Discussion of to-do lists from each lab (EH)
16:45-17:00	End of day 1

Thursday 13th October

Facilitated by CreativeHuddle, customised version of Team Tune Up workshop (<https://www.creativehuddle.co.uk/outlines/team-tune-up>). Lunch provided.

08:30	Arrivals, for 09:00 start
09:00-09:30	Recap of project vision (LH)
09:30-10:30	Facilitated session 1

10:30-10:45	Break (refreshments provided)
10:45-12:00	Facilitated session 2
12:00-13:00	Lunch (provided by Hilton York)
13:00-15:15	Facilitated session 3
15:15-15:30	Break (refreshments provided)
15:30-16:30	Facilitated session 4
16:30-17:00	End of day 2
18:45	Dinner at Pizza Express York, paid for by project. River House, 17 Museum St, York YO1 7DJ.

Friday 14th October

Half day, morning only. Lunch not provided.

08:30	Arrivals, for 09:00 start	
09:00-11:00	Split simultaneous debriefs	
	Co-Investigators (Clifford room) Shared debrief discussion. Outcomes of workshop, next steps, roadmap for project going forward.	RAs & RSEs (Rose Within the Walls) Shared debrief discussion. Outcomes of workshop, feedback on activities, start answering questions raised, finalise communications strategy from RSE/RA perspective, discussion on accomodating research interests
11:00-11:15	Break (refreshments provided)	
11:15-12:00	Full team debrief: Reconvene in Rose Within the Walls for shared discussion of simultaneous debriefs and joint roadmap of next steps.	
12:00	End of day 3 and workshop	

Attendees

Marc Alexander, marc.alexander@glasgow.ac.uk, +447740333561

Hannah Barker, hannah.barker@manchester.ac.uk

Riza Batista-Navarro, riza.batista@manchester.ac.uk

Youssef Benkhedda, youcef.benkhed@gmail.com

Andrew Bewsey, andrew.bewsey@nationalarchives.gov.uk

Jenny Bunn, jenny.bunn@nationalarchives.gov.uk

Harshad Gupta, harshad.gupta@nationalarchives.gov.uk

Ewan Hannaford, ewan.hannaford@glasgow.ac.uk, +447901505732

Lorna Hughes, lorna.hughes@glasgow.ac.uk, +447969311930

Hazel Jell, hazel.jell@nationalarchives.gov.uk

John Moore, john.moore@nationalarchives.gov.uk

Goran Nenandic, gnandic@manchester.ac.uk

Waltteri Nybom, waltteri.nybom@nationalarchives.gov.uk

Stefan Ramsden, drstefanramsdend@gmail.com

Valentina Vavassori, valentina.vavassori@nationalarchives.gov.uk

Pip Willcox, pip.willcox@nationalarchives.gov.uk