TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

FIRST REPORT
DISCOVERY PROJECTS

Our Heritage, Our Stories:
Linking and searching community-generated digital content to develop the people's national collection

OCTOBER 2022

University of Glasgow| The National Archives
University of Manchester

Table of Contents

Executive summary 1
Abstract 4
Aims and objectives 5
Overall project structure 7
Staffing and project structure 8
Project partners 9
Overall programme 12
Events and consultations 14
Research approach 17
Research results 23
Project outputs 24
Cross-project collaboration 26
Sustainability and infrastructure 27
Contacts 28
Annexes 29

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* is opening up the wealth of existing community-generated digital content (CGDC) across the UK using innovative automated approaches. This work is drawing on the multidisciplinary expertise of researchers in archives, computer science, history, and linguistics from the University of Glasgow, University of Manchester, and The National Archives (TNA). The project will develop an automated pipeline for the processing and enrichment of CGDC, showcase this newly enhanced and connected CGDC in a public-facing Observatory, and develop a post-custodial model of best practice for the creation and management of CGDC.

Achievements

Our project is currently at full staffing capacity, following successful recruitment at all project partners (although upcoming departures mean we will be re-recruiting for one new role, detailed below). We have built a flourishing partnership across the three organisations in the project, and regular and supportive systems have been implemented for meetings, reporting, and cross-team engagement that works well.

We have developed foundational project strategies, including our Engagement Strategy which is building stakeholder engagement, the development of user personae, and the basis for user testing; and our Content and Data Strategy, which has defined the four tranches of CGDC in scope for the project, and explored important connections with community archive networks, including Community Archives and Heritage Group (CAHG) Scotland and Manchester Histories (working with Archives+). This has expanded our connections with community archives. We have built our partner network, ensuring that all partners have been involved in the foundational stages of the project and the development of the project plan, making clear the input required from each of them at key stages of our workplan.

All three Milestones for year one of the project have been accomplished: construction of our prototype AI pipeline (M1) which has been successfully tested in harvesting CGDC from TNA's Manage Your Collections; our Observatory server has been established (M2); and an internal prototype of our Observatory has been developed (M3) for further refinement.

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 & Linguistics Lab have developed a prototype of the automated methods for knowledge extraction and linking from CGDC.
- Ingested and enriched data from this initial prototype represents the first iteration of our final enhanced CGDC dataset.
- The Observatory team have developed the foundation for further versions of the underlying graph database of the project.
- The Observatory team have also produced an early developmental model of the second Observatory prototype.
- 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).

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.

Challenges

As with all Discovery Projects, we have faced issues in filling all posts, especially posts at our Independent Research Organisation (IRO) 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. We currently have a vacancy following the departure of one PDRA, however we now have a good opportunity to hire someone who is best suited to the second phase of the project. This is true of the replacement PDRA in the History Lab, who starts on 1 October 2022 (Stefan Ramsden).

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 first call for the Community Fund, which was to be launched at a workshop which we had to cancel (see below), meaning a wider call will be issued in coming months. The Community Fund will be disbursed in full, to a revised schedule.

We remain in a global pandemic, and certain of our more optimistic predictions and schedules for this period have had to be revised, as well as short periods of staff absence (with Covid affecting every investigator). In particular, our project workshops have been challenging – while we have held our first workshop, the two following workshops to be held during this period have 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.

Partnership structure

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 Historical Thesaurus (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 is developing an AI pipeline (Deliverable 1) to assess, harvest and integrate very large amounts of CGDC into TNA's Discovery; for this to be used, we will create a major research-driven public-facing output in the form of an enhanced interface (Deliverable 2) to search and semantically link this newly enriched data to make it widely discoverable as part of a national collection.

To allow individuals and communities to ‘remix’, visualise, and share original and remixed metadata as linked data/stories, we will create a Remixer suite of open-source tools (Deliverable 3). We will also prototype approaches to new and emerging methods for semantic linking and data ‘crosswalks’ (Deliverable 4). To highlight the value of linking CGDC to existing collections, we will create compelling academic research studies using CGDC (Deliverable 5), to be published in the fields of history, linguistics, archive studies, and digital humanities as demonstrators of the value of this newly-discoverable content. As our AI work facilitates CGDC moving towards a post-custodial model, we will engage the community in co-production of methods, training, and tools (Deliverable 6) for archives and collecting organisations to effectively and sustainably harness the benefits of this new approach to managing collections.

Finally, based on the experience of the project and the need for all parts of the TaNC programme to engage with the wider questions around a national collection, we will produce a series of White Papers (Deliverable 7) which share our evidence-led framework to address policy and practice around CGDC, including a CGDC risk register, emerging post-custodial approaches to archives, language as a heritage object, scoping future metadata standards, and assessment of ethical uses of AI for the community context.

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 processes 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.

Overall project 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 are based. There are two fractional PDRAs at Glasgow (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. 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). At TNA, the project’s five-person team at TNA, includes a Co-I (with responsibility for all aspects of Observatory delivery), a delivery manager, and three Research Software Engineers (RSEs).

Staffing and project 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 text mining and in bridging AI and humanities archive data, particularly in the area of multilingual content and language ontologies.
- Pip Willcox (The National Archives) leads the Observatory Lab, overseeing its technical delivery, deputy-leads the History & Impact Lab, and coordinates TNA-based workshops.
- Prof Hannah Barker (University of Manchester) leads the History & Impact Lab, overseeing and designing the scoping and use of CGDC to build research capacity.
- Prof Goran Nenandic (University of Manchester) leads the AI & Linguistics 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 & Linguistics Lab.

Managers

- Dr Ewan Hannaford (University of Glasgow) is project manager, in charge of overall project co-ordination and planning, and research assistant working on linguistics-based work in the AI & Linguistics Lab (including the AI coordination work and multilingual/multidialectal data).
- Hazel Jell (The National Archives) is agile delivery manager, co-ordinating delivery of the project Observatory, including its design and user interface.

Research Assistants/Associates & Research Software Engineers

- Dr Andrew Bewsey (The National Archives), Waltteri Nybom (The National Archives), 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.
- Dr Jessica Hammett (University of Manchester) was research assistant within the History & Impact Lab for the first year of the project, contributing to identifying and engaging community archives, developing the project's post-custodial model for community archival management. She will be replaced from 1 October 2022 by Dr Stefan Ramsden.
- Dr Diane Scott (University of Glasgow) is lead researcher 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 Viktor Schlegel (University of Manchester) is research assistant within the AI & Linguistics Lab, developing the automated approaches to ingesting, processing, and enriching community-generated digital content.

Project partners

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 is 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 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 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)

NHLF 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 & Linguistics Lab 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 & Linguistics Lab 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.

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

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 2022
D0: Data tranche 0 collection and processing	Nov 2021	Jun 2022
4: AI pipeline V1 and NLP model development	Nov 2021	Jun 2022
C1: Community Fund Call 1 (targeted call)	Merged with C2	
5: Partner CGDC data gathering for ingest in D1	Dec 2021	Mar 2023
6: Establish platform & protocols for data sharing across project team	Jan 2022	Sep 2022
7: Initial research demonstrator case study development	Mar 2022	Jan 2023
8: Refine CGDC processing/enrichment during AI pipeline development	Apr 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**	
C2: Community Fund Call 2 (targeted call)	Sep 2022	Nov 2022
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 3	Jan 2023	Mar 2023

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**	
C4: Community Fund Call 4	Mar 2023	Jun 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
C5: Community Fund Call 5	Aug 2023	Nov 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**	
C6: Community Fund Call 6	Jan 2024	Apr 2024
18: AI pipeline V4 (final) development and refinement	Jan 2024	Jun 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 – 9th June 2022 Release – Forthcoming (Autumn 2022)	First user survey created and ready for distribution (release TBC, autumn 2022). To understand what is happening in community archives and other local or family history groups, particularly in respect of digital capabilities.
Forthcoming		
Event/consultation	Dates & participants	Description
Archives workshop series (replacing CGDC producers workshop)	From Autumn 2022 – Spring 2023 (with Community Archives Heritage Group Scotland & Manchester Histories)	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.

ARCadia	September 2022	ARCadia is a festival that will celebrate the public opening of the Advanced Research Centre (ARC), a flagship new building for the University and the people of Glasgow. We are running a community day event that is engaging, activity-led and family friendly, exploring the role of archives in the past and present.
Evaluating reuse workshop	October 2022	Workshop focusing on how CGDC is (re)used across users, researchers, and producers, and how this will feed into project outputs.
Remixing workshop	November 2022	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	March 2023	Collaboration-centred event involving local and institutional historians, developing project networks, furthering co-design approach, and informing design of historian-focused elements of project.
Post-custodial approaches workshop	June 2023	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	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 on *Our Heritage, Our Stories* is distributed across four key Labs: AI & Linguistics, Archives, History & Impact, 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 co-design, active research, iterative design, participatory and agile design, 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, text, image, moving image, and audio data – 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 will be collected and processed in multiple tranches according to our content and data strategy, with progressive tranches collecting and processing data of increasingly 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 brand new tools for the analysis, comparison, and reuse of CGDC. The design of this Observatory will be informed through collaborative discussions with community groups, institutional archives, and expert researchers. This work will be conducted across our Observatory, History & Impact 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.

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), to facilitate this cross-Lab collaboration. Regular project meetings occur each Monday, with the theme of these meetings operating on a rotating basis: the first Monday of each month is a meeting of all the investigators for core project planning, discussions, and decision-making; the second Monday of each month is an impact-focused meeting relating to engagement, outreach, and dissemination (principally relating to the Archives and History & Impact Labs); the third Monday is an all-hands roundtable for cross-Lab discussions; and the fourth Monday of each month is a technical delivery meeting for discussing key technical questions (principally relating to the AI & Linguistics and Observatory Labs). Our project advisory board advises on progress and overall strategy, provides operational oversight, acts as a positive feedback loop, and advises on broader applicability of outputs, ensuring value/impact beyond the project. This board has two sub-panels, meeting twice a year: 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 & Linguistics 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 & Linguistics work. 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 complimented 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.

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 & linguistics research is also developing and refining 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. 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 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.

Archives Lab

The focus of the Archives Lab is to ensure that that there is a balanced, representative and comprehensive range of CGDC available to the project. This will inform all aspects of project development (including AI, linguistics, and the Observatory design), and also be available in new and innovative ways as a final project outcome. To do this, the Archives Lab is working actively with all the heritage and collecting organisations in the project that are creating, managing, or disseminating CGDC, as well as the extensive stakeholder community working with this material. For example, so far, consultation has been carried out with CAHG Scotland, Manchester Histories, and Leeds City of Culture (LCC) 2023, in addition to existing project partners.

The Archives Lab is deploying a mixed methods approach to develop a deep understanding of the ecosystem of community content – including its creation and management, in close collaboration with heritage organisations and community groups. This preparatory work 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; 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 development and documentation of a post-custodial model for CGDC.

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, 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 TNA Archives Sector Development.

History & 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 will show the value of CDGC 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. 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 will produce studies of focused periods and times – indicative examples of these 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, or telling new and richly detailed stories about community responses to war and its aftermath and impact. The Lab is also actively working with the National Lottery Heritage Fund to gather examples of existing CGDC funded in the past 15 years, including using AI methods to uncover the data.

Collaboration between Archives Lab and History & Impact Lab

The History & 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 will identify and collaborate 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 CDGC 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 will work 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. 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.

Observatory Lab

The Observatory Lab's overarching approach is to iteratively and concurrently deliver three prototypes, and web scraping capability to feed into the NLP pipeline, which will lead to the development of the final project Observatory for the public to view, analyse, and compare CGDC. The prototypes are:

- Existing content management system based (Prototype 1)
- Bespoke (two-dimensional) (Prototype 2)
- Experimental (three-dimensional) (Prototype 3)

Initially, these prototypes are addressing casual and community users, with browse, search, and annotation tools. Other prototypes may emerge during the project. Web scraping capabilities will also be built, generating knowledge graphs from partner and community sites to train the NLP pipeline and link with other CGDC.

The Observatory is following an agile approach to development. This involves taking an iterative approach to research and development, as Research Software Engineers (RSEs) work in time-boxed iterations (sprints), in our case three weeks, to deliver a set amount of work (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, and the ability to shift activities depending on the outcome of research and user engagement activities.

We are also undertaking research into the existing landscape surrounding CGDC and linked data approaches, analysing interfaces based on linked data both within and outside the cultural heritage sector. We have also been analysing existing use of timelines, maps, and annotations 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.

Collaboration between Observatory and Archives Labs

The Observatory and Archives Labs are working together on the development and implementation of our Engagement Strategy. The key components of this are:

- scoping and design of engagement methods, including surveys, interviews, workshops, and general outreach, including project outputs such as our website, social media accounts, etc.
- carrying out stakeholder engagement, in order to understand the user and creator communities that have an active interest in CGDC. This includes establishing a network of experts to fill the roles of our stakeholder panel, our Observatory reference panel and our community content reference panels (panels to be confirmed following the meeting of our first Advisory Board meeting), as well as users and communities
- ensuring broadest input into all aspects of project design and development, and user testing

Core to this strategy is ensuring that the input from our community partners (now and in the future) is not over-utilised, resulting in burnout, and that where appropriate we ensure we have a plan for funding detailed input in a way which respects the limited resources of many stakeholders.

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. Early research in the Observatory Lab suggests Blazegraph as the preferred graph database to host the enriched CGDC, as it is open source, scalable, and supports RDF-star. RDF-star is currently being used as the preferred data format, as it provides flexibility with the ability to nest labelling, and is more easily transportable between databases. Research is ongoing, as we continue to discuss and test RDF vs RDF-star, and TTL vs JSON-LD. For the bespoke interface, the Vue3 JavaScript framework has been identified as it is fast, extensible, and new, providing longevity to the prototype.

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. The engagement surveys and co-design workshops will be used to gather requirements of our users and identify the features needed. However, the research has raised further questions, as our investigated examples worked with tightly scoped subjects and from aggregated data rather than data 'in the wild', as *OHOS* intends to. 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.

Results from the research of our Archives and History & Impact Labs will be codified into the project's post-custodial model for management of CGDC, a key project output. 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.

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 & Linguistics 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 developed a Blazegraph graph database to host enriched CGDC, forming the foundation for further versions of the underlying graph database of the project to be developed (Milestone 1).
- The Observatory team have also produced an early developmental model of the second Observatory prototype (Milestone 3).
- 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).

The project team have also been engaged in several dissemination activities, producing the following outputs:

- Thirty-minute 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.
- Paper presentation, IIIF conference, June 2022: 'Miiify: distributed crowdsourced annotation'.
- 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.

Future outputs

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

- Further and final versions of Observatory 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, containing a variety of innovative tools for visualising, analysing, and reimagining 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.

- 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. 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 will be disseminated across our networks and partners as teaching materials, describing a model of best practice for the creation and curation of CGDC.
- A series of project White Papers will be produced, 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 also be produced for the project, 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.

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. 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 discussion include bringing persistent identifiers to use of unstructured content, as well as IIIF approaches. In particular, we could fruitfully work with *Transforming Collections* to identify common approaches to surfacing supressed stories about CGDC. However, while we have attended an extensive range of cross project workshops, including the communicating colonial legacies event, we are yet to be able to commit the time to fully explore active collaboration, in this busy first year of project activity.

Sustainability and infrastructure

The technical infrastructure of *OHOS* can be seen in Annex 0, outlining the key data responsibilities of each Lab and the data flow between these. Fundamental technologies supporting the underlying infrastructure of the Observatory include: Blazegraph; miiify; Vue3 app UI; and Kong API gateway.

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 through 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 integrated into 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 Discovery will rely on TNA's core institutional commitment to archiving the data stored within Discovery (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.

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

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

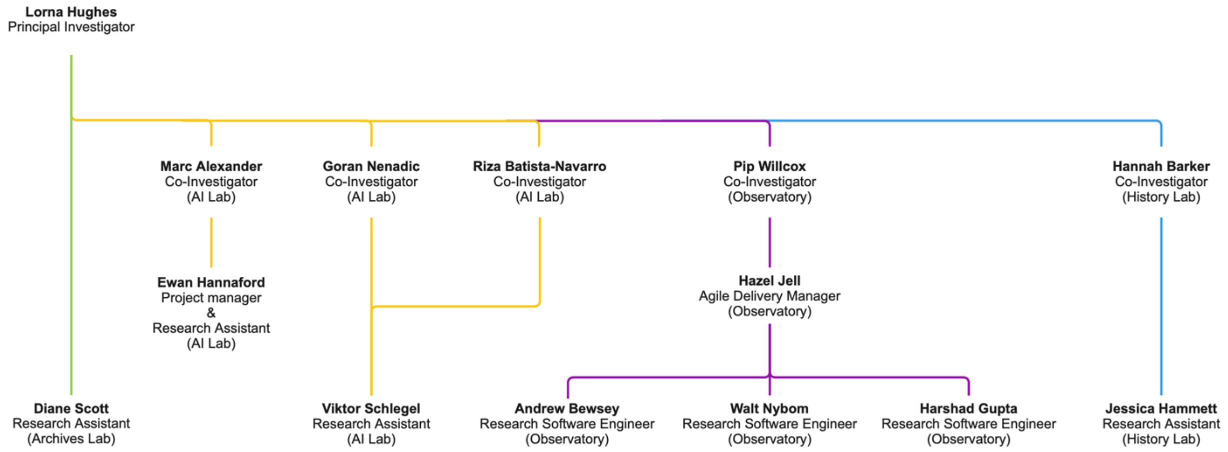
Diane Scott, diane.scott@glasgow.ac.uk

Viktor Schlegel, viktor.schlegel@manchester.ac.uk

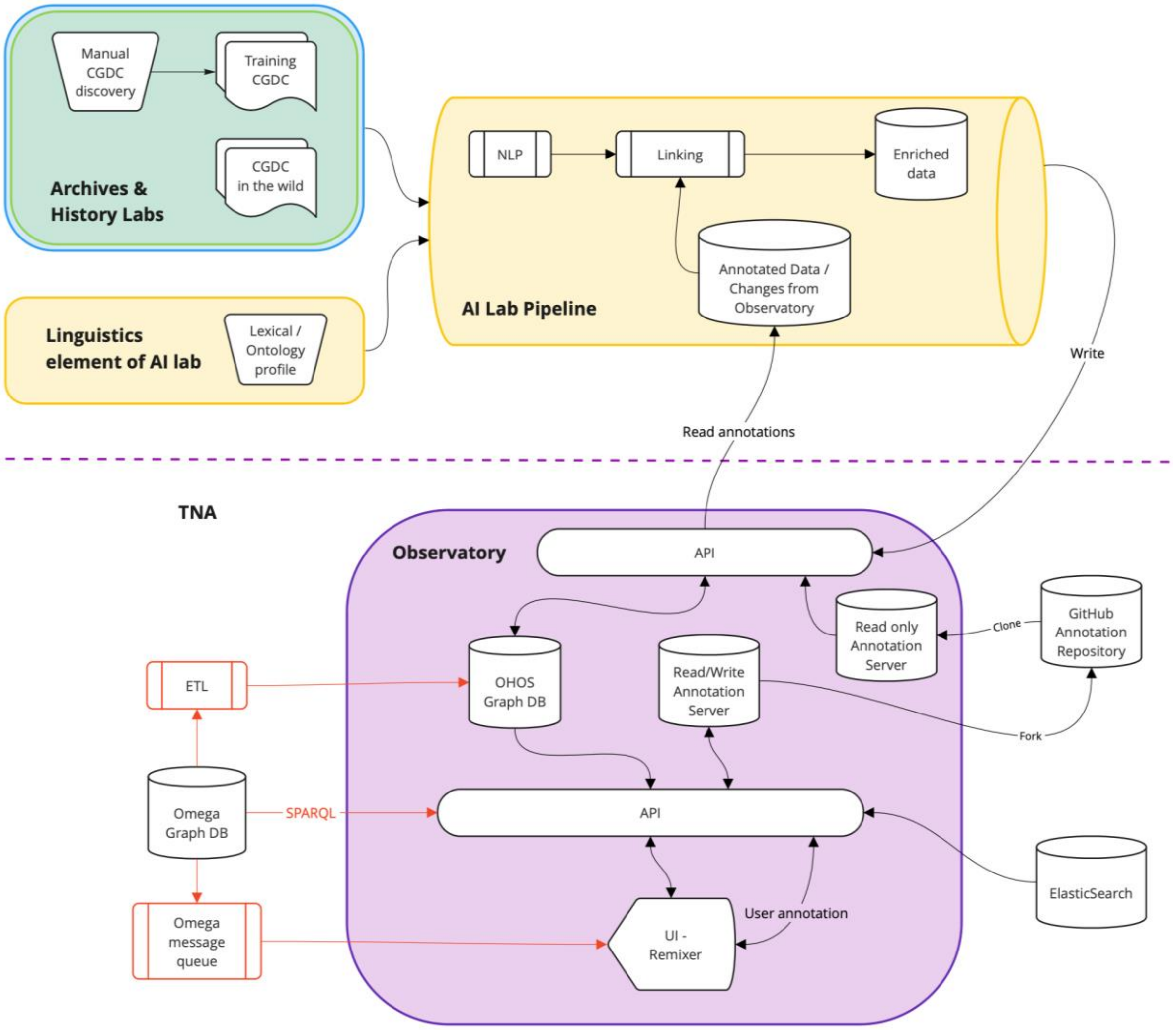
Annexes

Annex A: Staffing structure diagrams

Archives Lab	AI & Linguistics Lab	Observatory Lab	History & Impact Lab
Lorna Hughes Principal Investigator	Goran Nenadic Co-Investigator	Pip Willcox Co-Investigator	Hannah Barker Co-Investigator
Diane Scott Research Assistant	Riza Batista-Navarro Co-Investigator	Hazel Jell Agile Delivery Manager	Jessica Hammett Research Assistant
	Marc Alexander Co-Investigator	Andrew Bewsey Research Software Engineer	
	Ewan Hannaford Project Manager & Research Assistant	Harshad Gupta Research Software Engineer	
	Viktor Schlegel Research Assistant	Waltteri Nybom Research Software Engineer	



Annex B: *OHOS* technical infrastructure


Manual CGDC discovery
Training CGDC
CGDC in the wild
Archives & History Labs
Linguistics element of AI lab
Lexical / Ontology profile
NLP
Linking
Enriched data
Annotated Data / Changes from Observatory
AI Lab Pipeline
Write
Read annotations
TNA
Observatory
API
GitHub Annotation Repository
Clone
Read only Annotation Server
ETL
OHOS Graph DB
Read/Write Annotation Server
Fork
Omega Graph DB
SPARQL
API
ElasticSearch
User annotation
UI - Remixer
Omega message queue