



Archives & Records  
Association  
UK & Ireland

# arc magazine

[www.archives.org.uk](http://www.archives.org.uk)

No: 373

September 2020

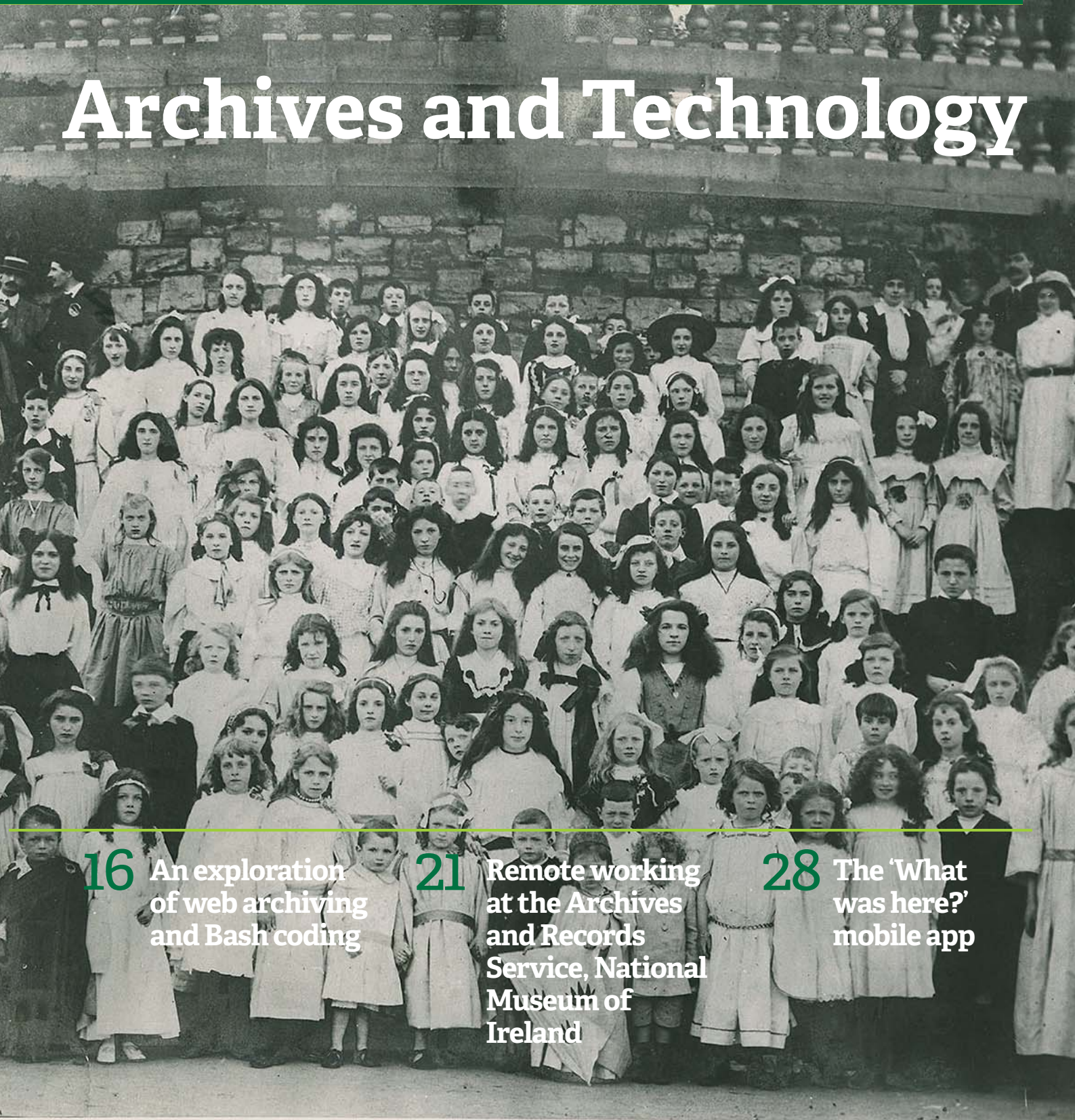
ISSN 2632-7171

## Archives and Technology

**16** An exploration  
of web archiving  
and Bash coding

**21** Remote working  
at the Archives  
and Records  
Service, National  
Museum of  
Ireland

**28** The 'What  
was here?'  
mobile app



EXPLORED  
DISCOVERED  
FOUND  
DETECTED  
CONNECTED  
ENGAGED  
UNEARTHED  
LEARNED



CONNECTED  
ENGAGED  
UNEARTHED

[www.archivescard.com](http://www.archivescard.com)



# Welcome...



## *Welcome to the Archives and Technology issue of ARC Magazine*

The events of the last 6 months have demanded the most resourceful, adaptable and creative responses from us, and this issue of the *ARC Magazine* is an excellent demonstration of how integral technology has been to this effort, and to the work of the recordkeeping sector in general. We were overwhelmed with high quality submissions this month, enough to fill the issue twice over!

In the articles that appear here we hope you'll find a balanced selection, including accounts of ambitious projects which take advantage of the opportunities brought by technology, advice for organisations to improve their digital preservation efforts, and more personal, reflective accounts of adjusting to remote working and how modern technologies keep us connected to our colleagues and our collections.

I hope, like me, you'll find the accounts of innovation and resilience presented here encouraging and inspiring.

A big thanks to Ashleigh Hawkins, Elisabeth Thurlow and Jenny Bunn from the ARA Archives and Technology Section for gathering and vetting submissions. And, of course, a huge thanks to all of our contributors!

**Adele Clarke**  
*ARC Editor*

ARC Magazine is the monthly publication of the Archives & Records Association (UK and Ireland)  
Prioryfield House  
20 Canon Street  
Taunton  
Somerset  
TA1 1SW  
Tel: 01823 327030

Fleadh na Leanbh [Children's Fete, part of annual Oireachtas competition] 1909, complete with children in costume. Courtesy of James Hardiman Library Archive, NUI Galway: Conradh na Gaeilge collection: G60

ARC Magazine advertising enquiries to:  
stewart@centuryonepublishing.uk  
Tel: 01727 893 894

Send articles/comments to:  
arceditors@archives.org.uk

ARC Magazine design by Glyder [www.glyder.org](http://www.glyder.org)

The Archives & Records Association (UK and Ireland) cannot accept responsibility for views expressed by individual contributors to ARC Magazine. It is a medium for informing members of news, information and ideas relevant to the profession, including archive conservation. It is not an official guide to procedures, concepts, materials or products.

# Contents

No: 373  
ISSN: 2632-7171  
September 2020

**arc**  
magazine

## Association news

- 3 Editorial
- 5 Opening lines
- 6 Professional development news
- 7 Collecting matters

## Archives and Technology

- 8 Chair Update

9

## From Magdalen's medieval Muniment Tower to Web 2.0

- 13 A new way to organise and search the archives: automated topic detection in Archives Portal Europe
- 15 User experience and access to born-digital data produced by publishers: the case of Carcanet Press
- 16 An exploration of web archiving and Bash coding
- 19 Trying to work out your digital preservation risks? Now you can DiAGRAM them
- 21 Remote working at the Archives and Records Service, National Museum of Ireland
- 23 From self-isolation to self-development
- 26 Conradh na Gaeilge collection
- 28 The 'What was here?' mobile app
- 31 Who was Frederick Lanchester?: a car manufacturer, engineer, scientist or inventor?



# Opening lines

**Sharon McMeekin**, Head of Workforce Development at the Digital Preservation Coalition, discusses the unique value of digital records in times of crisis and the successful From Novice to Know-How online training programme



Stating that these are times of monumental change seems almost clichéd now. From the uncertain financial and political future of a post-Brexit UK, through a global pandemic, to the social injustices and inequalities in our society finally being truly laid bare. We can only hope to emerge as a healthier society in all possible ways.

To do this we will need reliable, authentic, and accessible information: to aid good decision making, to facilitate vaccine research, to help inform and empower us as we work towards a fair and equal society, and to hold those in positions of power accountable.

It almost goes without saying, but I will anyway, that this information will be almost exclusively created, used, and shared in digital forms: be it government documents, research

data, or the messages of activists on social media. As keepers of our societies' records, it is our duty to ensure this digital information survives and is useable.

But herein lies a problem, years of underfunding and cuts have left us ill-prepared for the, often overwhelming, responsibility to preserve the exponentially increasing amounts of digital information. While there are several issues that could be discussed, as Head of Workforce Development for the Digital Preservation Coalition (DPC), my focus is the digital skills gap.

The 2019 survey on digital capacity in the UK archives sector from The National Archives UK (TNA UK) and Jisc reported that only 15% of archives in the UK felt they had full digital preservation capabilities.

*“As keepers of our societies' records, it is our duty to ensure this digital information survives and is useable”*

This was a substantial skills gap that needed to be addressed.

I stand in awe of TNA UK's resulting Plugged In, Powered Up strategy for building digital capacity in the sector. The strategy includes twenty strands of coordinated activities aimed at addressing issues around digital

Landing page of 'Novice to Know-How' online training programme. Reproduced by permission of the Digital Preservation Coalition (DPC)

**Novice to Know-How: Digital Preservation Skills for Beginners**

The Novice to Know-How learning pathway aims to provide beginners with the skills required to develop and implement simple digital preservation workflows within their organisation. It starts with a broad introduction to digital preservation issues and describes the steps we can take to address these. It then delves into potential workflows in more detail, examining the issues to consider, steps to take, and technological solutions that can be deployed. An emphasis is placed on free, easy-to-use solutions, and includes detailed demos of how to use the key tool, DROID.

**Learners will be able to:**

1. Explain what digital preservation is
2. Prioritise digital preservation actions at their organization
3. Develop and implement digital preservation workflows

**Time to Complete:**  
c. 11 hours

The development of this learning pathway was funded by The National Archives (UK) as part of the "Plugged In, Powered Up" digital capacity building strategy.

[Continue this learning path](#)

0%

**“***Novice to Know-How (N2KH), has surpassed the goals set for it. In the first 5 months, over 1000 learners have had access and, while most have been in the UK, learners have come from 33 countries***”**

engagement, access, preservation, and my favourite, skills. In this final area, I have had the honour of being directly involved.

At the DPC we have wanted to develop online training for several years, but lacked the capacity to make it happen. In late 2019, however, the invitation to tender for the development of online training as part of Plugged In, Powered Up offered the perfect opportunity.

The key aim of the project has been to create training that provides learners with the practical skills needed to establish simple digital preservation workflows. The structure of the training was developed collaboratively with colleagues at TNA UK, taking into consideration sector needs and good practice, coordinating with other strands of work.

I am happy to report the resulting learning pathway, Novice to Know-How (N2KH), has surpassed the goals set for it. In the first 5 months, over 1000 learners have had access and, while most have been in the UK, learners have come from 33 countries. Feedback has been excellent, frequently mentioning the clarity and accessibility of the content, even for absolute beginners, and the confidence found to start implementing digital preservation. We hope, in time, to capture and share information on the impact of N2KH.

The current learning pathway is not the end. We are already working with TNA to expand the number of tool demos in N2KH, and we will be developing complementary content from the DPC. Writing digital preservation policies, web archiving, metadata, and risk management are topics towards the top of the list.

As record keepers we have many hurdles to overcome. We hope to, at the least, help remove some of those hurdles relating to digital preservation skills development.

To find out more about the course, and other resources available from the DPC, visit our website: [www.dpconline.org](http://www.dpconline.org). Also, check out the Plugged in, Powered Up strategy and accompanying resources on the TNA website: [www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/](http://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/)

# Professional development news

In this month's professional development news we catch up with **Jim Costin** FMARA, a former Bridging the Digital Gap Trainee who recently qualified as a Foundation Member of the ARA.

## What is your current role?

I currently work in the Digital Content Unit at Cambridge University Library (the UL). I'm digitising a range of material including books from special collections, photographs and other material. I've also been helping out in the filming of a few keynote talks and conducting some research into multispectral imaging. Prior to working at the UL, I worked for The National Archives UK (TNA UK) as part of their Bridging the Digital Gap programme, seconded to The Borthwick Institute for Archives at The University of York. It was in that role that I was introduced to the Archives and Records Association (ARA) competency framework and, after reviewing the criteria, decided to apply for it.

## Why did you apply for Foundation Membership?

I applied for the membership as part of the Bridging the Digital Gap initiative run through TNA UK and supported by ARA. The fact that the membership was part of the scheme was one of the deciding factors in my application for the role. Having the ability to gain membership and a firm foundation to build a career in archives and special collections is, in my opinion, pretty invaluable.

**“***The fact that the membership was part of the scheme was one of the deciding factors in my application for the role***”**



# Collecting matters

Archives Sector Development has continued to operate and has adapted to changed circumstances, as **Mike Rogers**, Sector Development Manager: Midlands, Transport & Schools, explains.

For those of you unfamiliar with the work of Archives Sector Development, we are responsible for delivering much of The National Archives UK's (TNA UK) role as lead for the wider archive sector. We provide support to organisations ranging from national institutions to community archives, promoting best practice and collaboration, collating information on collections, and acting as a first point of contact.

A key part of the work of the Regional and Networks Team in particular is finding new contacts and travelling to the four corners of England to meet archivists, record creators and key stakeholders in person, and to view their collections and premises. Six of the team of ten live far from London, based within their allocated regions, and were already set up for home-working when not onsite at TNA, or on visits. However, many of our colleagues were very much based in our office in Kew.

Suddenly, in the middle of March, all this changed. Meetings and events were cancelled at short notice as the COVID-19 situation rapidly escalated. Within little more than a week, repositories throughout the UK had closed their doors. This left their staff suddenly working from home, isolated from colleagues, with few resources at hand, and often with limited access to corporate IT. This applied to us as well – three of us had left our laptops locked in drawers at Kew while on site visits.

For the first few weeks we monitored the situation nationally, developed advice on lockdown procedures and clarified our expectations in terms of public records and Archive Service Accreditation. Throughout this, we re-assured repositories that we were still functioning and available. We conducted surveys of the sector, arranged business continuity webinars and gathered information on when sites closed. We also sought to find out what level of remote service they were hoping to offer during lockdown, and what effects they anticipated the crisis would have on finances and staffing (including furloughing). Internally, we were getting to grips with indefinite home-working and new ways of communicating and collaborating.

As the weeks passed, we began to re-engage with individuals in the sector and began working with colleagues and external stakeholders in collating information that the sector would find useful when planning to re-open. After the initial shock, the routine work of the department has resumed: sales catalogues are being monitored as



Jim Costin. Image courtesy of Scott Maloney

## How does it feel to have qualified as a Foundation Member?

It feels really good actually! To know that all of the hard work I put in to both complete the competency framework and learning all the new skills has led to something which will help in my future career is a great feeling.

## What advice would you offer to others thinking of enrolling on the programme and develop their own application?

Keeping a diary of your activities will help when it comes to write up the application and provide evidence. By referencing those activities and providing documentation to support those, you can more clearly see what level you are working at and what needs to be done to get to the next level. And, whilst it may seem like quite a lot of work, the scheme is well worth doing in my opinion. It gives you the opportunity to self-assess and reflect on where your skills are at and how to continue to develop them.

auction houses have begun to function again, the four strands of the Archives Revealed and Collaborate and Innovate funding programmes have resumed, and both fundraising and digital training has taken place. Accreditation assessments and awards have been made, to the credit of repositories and assessors alike. Work has continued on developing a Level 7 Apprenticeship Scheme; New Burdens Funding allocations have been calculated and we are even gathering ideas for the latest instalment of *A Year in Archives*. We are currently preparing for a 'Reopening Inclusively' webinar, building on the survey results.

As I write, the wider sector, in particular local government archives, is gradually reopening to the public. We are monitoring this process, looking for similarities and differences of approach, innovative ideas that can be shared more widely, and indications of usage and demand. On a less positive but necessary note, as participants in the Crisis Management Team for Business Archives, we are actively seeking information about the increasing number of companies and independent schools entering administration, and alerting relevant local repositories in the hope of preserving key records.

So how have we adapted to the new circumstances? IT has been absolutely essential and, despite the occasional issue, has allowed us access to most of the resources we need, in a way that would have been impossible even a few years ago. Video conferencing has been central to our activities. Initially, this was largely used for internal communications, exchanging information and, crucially, allowing us to see familiar faces in a time of huge upheaval and uncertainty. Once things began to settle down, we started to use a range of online platforms for maintaining contact with individual repositories, introducing ourselves virtually to new contacts, attending meetings of external bodies, and delivering virtual training and support events. Some of these were adapted from planned physical events, while others were very much a response to current circumstances.

We have now entered a phase resembling 'business as usual', but with the key difference that we are still all home-based and operating virtually. Few of our department have visited Kew, even fleetingly, for months now. Site visits, in-person meetings and physical events are simply not taking place, but Archives Sector Development is very much in operation. We're available, as always, to offer advice and support to the sector, so do get in touch and check out our website.

Twitter: @UkNatArcSector

Website: [www.nationalarchives.gov.uk/archives-sector/](http://www.nationalarchives.gov.uk/archives-sector/)

Email: [ASD@nationalarchives.gov.uk](mailto:ASD@nationalarchives.gov.uk)

# Chair's Update

**Jenny Bunn**, section chair, introduces this unique issue from the Section for Archives and Technology, composed amidst the upheaval of a global pandemic.

Welcome to this Section for Archives and Technology special issue. Our work in putting this issue together took place whilst we were all experiencing the new sensation of lockdown in response to the COVID-19 crisis. The initial call for papers made the point that 'under the current exceptional circumstances, we are having to learn how to harness the power of the possibilities [brought by digital technologies] faster and more urgently than ever before'. A number of our contributors reflect on their experiences of remote working, with new laptops arriving like Christmas presents from IT colleagues, and one has even been prompted to reflect on the vulnerabilities and fragilities of the technological systems on which we have all been relying so heavily.

Nonetheless, at the point of writing this, the internet has not yet fallen over and the personal testimonies contained in this issue attest to the way in which we, as a profession, have successfully adapted projects and continued to carry out many aspects of our business (almost) as normal. A thought I have heard aired on a number of occasions is that of how on earth we would have coped if this had happened 10 or even 5 years ago. Personally I find this thought oddly heartening as it reminds me of quite how far we, as a profession, have come in such a relatively short time in terms of learning how to use the evolving technologies to further our work and its reach.

Within this special issue are reports of archive services

involved in; app development, experimenting with augmented and virtual reality, developing online exhibitions, building digital preservation workflows from scratch, web archiving, coding, and engaging in advanced data-wrangling and retro-conversion projects. Being old enough to remember when a word-processed finding aid was seen as cutting edge, the distance the profession has travelled constantly amazes me and I think we should all feel real pride in our growing technological-savviness and in everything that it has helped us to accomplish.

Technology of course, marches on, and it is also heartening to see in this issue that the profession continues to push itself forward. Alison Harvey reports on some of the continuing professional development she has undertaken during lockdown, reminding us that we all have the capacity to keep learning. And, at the other end of the scale, it is also good to see in this issue, reports that show how those larger institutions and organisations (The National Archives and Archives Portal Europe) are using their greater capacity and resources to continue to push research forward into that cutting edge of technology.

On behalf of the entire committee, I would like to thank all our contributors and to wish everyone well in what have been (and will continue to be) very difficult times. We hope this special issue will provide you with both some light relief and a bit of inspiration!



# From Magdalen's medieval Muniment Tower to Web 2.0

**Charlotte Berry**, of Magdalen College Oxford, takes us through the history of the muniment records' catalogues, and how lockdown has created the long-awaited opportunity to get them online.

**W**hat has been the main consequence of lockdown? No readers! Working from home with fewer interruptions! The result? Starting to put Magdalen's 48 volume calendar of our medieval collection into our online catalogue Epexio! This collection is

internationally significant and well used, but not the easiest to access for readers or staff trying to locate relevant material. This article looks at the practical challenges of converting its very old style catalogue from pdf to Epexio.

Magdalen's Muniment Room with original 15th century presses, deed boxes, chests and tiled floor. Courtesy of the President and Fellow, Magdalen College, Oxford







*The Original Entrance to the Cloisters at Magdalen College.*

Engraving of Magdalen College, showing the Muniment Tower on the left hand side by the scaffolding, 1789 (FA1/7/1P/4). Courtesy of the President and Fellow, Magdalen College, Oxford

Search or enter website name

Magdalen College Archive

Grant from Henry III to the Hospital of St John the Baptist of the mill called Kyngesmeine [Kingsmill] at Edendon, with the appurtenances and also the garden which was the Jews' garden in the suburbs of Oxford as a site for rebuilding the Hospital, reserving to the Jews space for a burial ground

Metadata Children Objects Digitisation Use Conservation History

Edit Edit reference number Finding aid

Identity

Reference number M/CRC/8  
Calendar number CRC 2  
Alternative reference number(s) Chartae Concessae 8  
Title Grant from Henry III to the Hospital of St John the Baptist of the mill called Kyngesmeine [Kingsmill] at Edendon, with the appurtenances and also the garden which was the Jews' garden in the suburbs of Oxford as a site for rebuilding the Hospital, reserving to the Jews space for a burial ground  
Date(s)  
Date 12 Jul 1231  
Date range 1231  
Level of description Item  
Repository Magdalen College Oxford

Content and structure

Scope and content

8 Grant from Henry III to the Hospital of St John the Baptist of the mill called Kyngesmeine [Kingsmill] at Edendon, with the appurtenances and also the garden which was the Jews' garden in the suburbs of Oxford as a site for rebuilding the Hospital, reserving to the Jews space for a burial ground, 12 Jul 1231

8a General pardon from Edward IV to Waynflete for all offences before 30 Sep 1470, with confirmation of possessions in mortmain, 10 Jan 1472

9 Confirmation by Henry III, at the instance of Walter master of the Hospital of St John the Baptist, of the charter granted by him on 12 Jul 1231, "quia sigillum nostrum quotic utebamur mutatum est", 18 Oct 1267

9a General pardon granted by Edward IV to Waynflete, specially for lands obtained in mortmain, 14 Jul 1481

10 Confirmation by Edward II of the four charters granted by Henry III to the Hospital of St John the Baptist, 7 Aug 1317

10a General pardon from Henry VI to the prior (John

New catalogue entry in Epexio for Chartae Concessae 8, ready to publish online. Courtesy of Magdalen College Archive



Magdalen College was founded in 1458 by the Bishop of Winchester, William Waynflete, who lived to a ripe old age and built up an impressive property portfolio across 20 English counties. Waynflete built the Muniment Tower especially to house the 13,000 or so title deeds to the lands which he had bought to endow the college. These so-called 'muniments' are still housed in original oak cupboards and made-to-measure deed boxes which are arranged by county/property. The collection was catalogued in manuscript form in the 19th century by college fellow William Dunn Macray, and then typed up into bound catalogues in the 1930s. These volumes were recently scanned into PDF format. While it has been useful to send these to readers as a short-term measure, the end goal has long been to get the Macray catalogue online.

Enter onto the scene - Epexio! This is the collection management system produced by the innovative and excellent Metadatis team. The college acquired Epexio in spring 2019 and work started on migrating existing finding aids across and customising the online catalogue. Lockdown has now happily sped up the process of migrating the Macray catalogue, enabling intensive editing to take place over a short time period.

In the procurement stage, Rachel Care at Metadatis had already experimented with using Adobe and OCR to extract the data from the extant PDF volumes. This resulted in an Excel spreadsheet being generated for each PDF volume. We ended up using FineReader instead, at a small annual subscription cost; the quality of the data extraction was far superior, as there were fewer mistakes in interpreting the letter-shapes used in the sometimes faint 1930s typescript. I ran all 48 or so PDF volumes through FineReader, which generated an Excel spreadsheet for each volume. This was Phase One of the retro-conversion. I then sent these files off to Metadatis for Phase Two, where the team wrote code to sort the data into a small number of ISAD(G)-compliant columns, matching the fields in Epexio's cataloguing template. These Excel files (a very rough version of the typescript catalogue) were then returned to Magdalen for Phase Three.

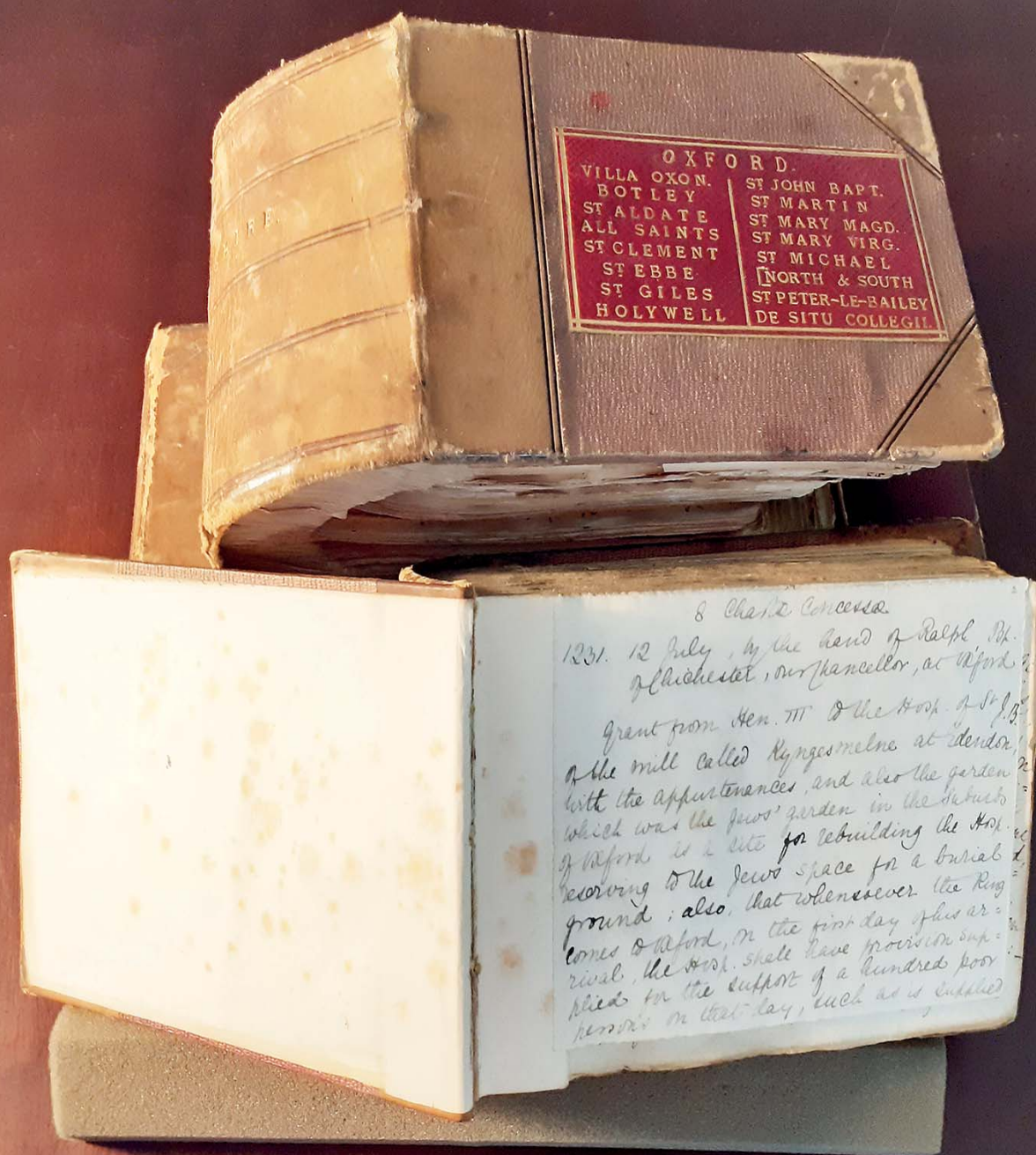
Phase Three was by far the longest phase of the process – undertaking detailed editing to iron out spelling mistakes, issues with



dates, problems with a complex numbering system, adding in extra fields not used in the original 19th century manuscript catalogue, etc. The aim was to create a spreadsheet for each county which was fully ISAD(G)-compliant and populated with clean data. Phase Four was the easiest of the project, simply sending the completed spreadsheets to Metadatis, who loaded it into the catalogue. The RefNo field generated the tree hierarchy automatically, and the whole catalogue structure fell miraculously into shape, with a bit of pushing and pulling in places. The data can then be further edited online via Epexio, until it is ready to go live online (not quite yet!).

William Dunn Macray (clerk 1844-1850, chaplain 1856-1870, fellow 1891-1916) (O1/P1/1/f52). Courtesy of the President and Fellow, Magdalen College, Oxford.





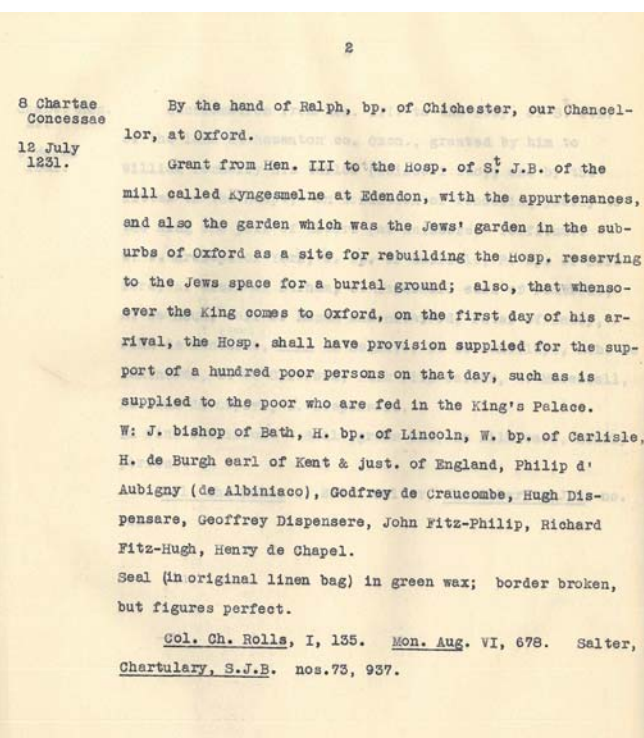
19th century manuscript catalogue entry by Macray for Chartae. Courtesy of Magdalen College Archive Concessae 8

Inevitably some problems arose due to 19th century cataloguing practices. The deeds had been individually catalogued and assigned a number on loose slips of paper, which were then bound into volumes in date order (therefore resulting in a random and meaningless sequence of RefNos). When the catalogue was typed up in the 1930s, a page/calendar number was also introduced, and this has occasionally been used by readers in citations. An extra AltRefNo has now sometimes had to be introduced to make the online catalogue better reflect the hard copy's original structure. Three volumes have been inputted manually due to problems with their

typescript layout and numbering. Other issues came up through using Excel, such as problems with post-1900 dates converting themselves instead into number sequences, and pre-set, auto-fill and auto-correct settings overwriting text. However, these points were minor and were easily resolved with much appreciated help from Gillian Sheldrick, who runs excellent courses in using Excel for archive cataloguing.

This retro-conversion project will transform how we understand this key collection. How many deeds do we have altogether? With seals? In French? Which is the





1930s typescript catalogue entry for Chartae Concesse 8, now digitised into pdf format by TownsWeb. Courtesy of Magdalen College Archive

oldest deed? Which manors have the most deeds? Now that the workflow has been tested and finalised, I have written a detailed 'How To Guide', as it is not impossible that a successor may finish the work on the catalogues in future years. Five catalogues are completed, only 43 or so to go...

“*Lockdown has now happily sped up the process of migrating the Macray catalogue, enabling intensive editing to take place over a short time period*”

## A new way to organise and search the archives: automated topic detection in Archives Portal Europe

The Archives Portal Europe (APE) are investigating the challenges and opportunities of topic searching within their portal, as told by **Marta Musso** (APE and King's College London), Kerstin Arnold (APE) and Federico Nanni (Turing Institute).

APE aggregates and connects via a single research point the catalogues and digitised material of over 1,100 archival institutions in and about Europe. It currently hosts material from more than 30 countries, from a variety of archival institutions (such as State archives, city archives, university and parish archives, private institutions, and more). It is maintained by the Archives Portal Europe Foundation, an international consortium of national archives and other archival institutions that aim to connect the archival material of single institutions into one digital repository; for the UK, the representatives are Archives Hub and The National Archives (TNA). The aim of APE is to allow universal access to the archival heritage of Europe, and to promote new forms of archival research that go beyond national or local boundaries.

One of the research tools made available by APE is a search by topics. This allows users to select archival collections held in archival institutions from different countries and in different languages, which are pertinent to the same subject (organised according to the classification index of The United Nations Educational, Scientific and Cultural Organisation (UNESCO) Thesaurus, and the main classifications of strands of history research). Archives are traditionally not organised by their subject content, but around the entity (person, family, or organisation) that created and/or collected the documents in

## Explore our topics:



The current Topic Cloud. Courtesy of Archives Portal Europe Archives Portal Europe

*“The research is extremely complex as it involves a series of hindrances both on a technical and conceptual level”*

the course of their activities. While this is an undisputed pillar of archival management, the availability of digital repositories for archival research requires new tools for digital archival research, particularly when different archival traditions from different countries and different types of institutions are merged into a unique research portal.

Topics then become a fundamental resource to support archival research. However, they are currently added manually by the archivists ingesting their material in APE, and the vast amount of archival material ingested in the portal, in more than 20 languages, makes it impossible to have a comprehensive body of topics that describe the whole of the APE repository in a fully consistent way. Furthermore, archival institutions participating in APE have shown different levels of engagement with the topic tool for different reasons, ranging from scarcity of resources to lack of taxonomic traditions.

King's College London is currently conducting research to devise a tool for automated topic detection in APE; it aims

*“The availability of digital repositories for archival research requires new tools for digital archival research”*

at creating an approach able to work in a multilingual environment, where human-created taxonomies act in combination with a machine learning algorithm to aggregate relevant material around a specific topic. First, a list of relevant entities and key concepts was pulled from the existing topics, creating for each topic a vocabulary of entries uniquely associable with one topic, through a combination of automated analysis on the finding aids available for one topic, and human control. Each vocabulary is linked to Wikidata in order to make it available across all 32 languages currently represented in the portal, and to allow a specially designed algorithm to model the entire repository of the portal. This is used, in combination with cross-lingual word embeddings from the project MUSE by Facebook Research, to both assign unlabelled documents to their most relevant topic and to retrieve relevant material given a user query across languages.

The research is extremely complex as it involves a series of hindrances both on a technical and conceptual level. Firstly, topics are organised on different scales, from the macro (“Economics”) to the micro (“German Democratic Republic”). Secondly, archival descriptions largely vary in the way they are written and organised from day-to-day language. We are tackling different archival traditions across Europe that have evolved over the centuries – simple vocabulary translations are a limited tool, and would require special monitoring and human intervention for each of the languages represented. A third hindrance is the small number of natural language processing toolkits available for many languages, which again strongly limits the possibilities for automatic key concept extraction in languages outside English and a handful of the more popular European languages.

A proof of concept will be released in August. While optimal results that provide perfectly comprehensive topics may be a long way off, this research is an innovative application of digital technologies to archival research, with the potential to change the way in which archives are organised (and searched for) in the digital space.

More information about the APE can be found at [www.archivesportaleurope.net](http://www.archivesportaleurope.net).



# User experience and access to born-digital data produced by publishers: the case of the Carcanet Press

**Lise Jaillant**, of Loughborough University, considers the restrictions of technology and confidentiality for archives, and archive users, interested in improving accessibility to born-digital records.

What can we do to make born-digital data more accessible? As a publishing historian, I rely heavily on access to primary sources, which are increasingly becoming digital. This short article presents my work on Carcanet Press, a leading poetry publisher based in Manchester. It addresses two issues at the core of the 'dark' archive situation: first, technical issues to make born-digital records available; and second, issues relating to the confidentiality or sensitivity of these documents. This work, funded by an AHRC Leadership Fellowship, builds on an earlier project that aimed to bring together archivists and scholars, for which I received a British Academy Rising Star Engagement Award. After three years of discussions and collaborative work with archivists, I am convinced that we need to move fast (and avoid breaking things). Open data respectful of privacy is possible, and the first step is to quickly build prototypes to give access to archival records.

Since the late 1970s, the John Rylands Library in Manchester has acquired the Carcanet Press archive on a yearly basis. The vast majority of the paper archive is uncatalogued and closed to researchers; and the digital part of the collection is a 'dark' archive, open only to a handful of staff. My AHRC grant allowed me to employ a Project Archivist for a couple of months in 2019. She was based at the Rylands Library and had access to the entire collection. In Summer 2019, she prepared a selection of 200 emails that she thought would be interesting for me to see. She then submitted the selection to Michael Schmidt, the founder of Carcanet Press, for approval. Schmidt requested that some materials be closed or redacted for confidentiality reasons. The

redacted selection of emails was then sent to me as a PDF, with email attachments in a separate ZIP folder.

For archivists, only basic technical skills are necessary to provide access to emails and other born-digital archives. There is no need to build a complicated system, or to buy expensive tools. Creating a PDF is enough to allow users to see content that will be useful for their research. I am not saying that this is perfect: as a researcher, I wish I could download thousands of emails and do some data analysis. But even a small selection of data is better than no data at all. Technical issues at the core of the 'dark' archive problem can be easily resolved if we change our mindset and embrace imperfection. A prototype can be improved over time, whereas a closed archive remains static and inaccessible.

The second issue, the confidentiality and sensitivity of some born-digital documents, can also be addressed with a change of mindset. It is essential to embrace risk and trust that researchers will make good use of the data they access. And for users, it is important to respect privacy. Each time I saw 'CLOSED' or '[.....REDACTED.....]' on the PDF, I wished I could see the entire message. Yet, I also realise that I would not want people to access all of my emails. Closure and redaction are reasonable measures, as long as the user is informed of the withdrawal of information.

Many libraries and archival collections are now experimenting with new systems to make their digital collections more accessible. For example, ePADD (an open-source software) is a valuable tool to discover born-digital materials,

but researchers still need to travel to Special Collections to consult relevant records. For archival repositories with limited staff time and funding, one solution is to create PDFs based on certain themes and to make them available to users after obtaining permissions. This is a low-tech solution that nearly all institutions could implement rapidly to respond to user needs. "Our users are crying out for faster access," argued Mark Greene and Dennis Meissner in their influential article 'More Product, Less Process'. To resolve the 'dark' archive problem, archivists need to start with the users and quickly work backwards. But unlocking born-digital data is not a one-way process. We need more collaboration between archivists and users. We also need more empathy: the ability to understand the concerns of archivists, and the needs of users of born-digital collections.

For more information on these AHRC projects, see: Jaillant, Lise, "After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives". *Archives and Manuscripts*, vol. 47, no. 3, (Sept. 2019), pp. 285–304.

For more information on ePADD, see: Schneider, J., et al, "Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives", *Archives and Manuscripts*, vol. 47, no. 3, (Sept. 2019), pp. 305–26.

For more information on low-tech solutions, see: Greene, Mark, and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing", *The American Archivist*, vol. 68, no. 2, (Sept. 2005), p. 235.

# An exploration of web archiving and Bash coding

An adventure in DIY web archiving from **Christopher Prince** of the Greater Manchester County Record Office.

During the COVID-19 lockdown, I have been looking into archiving the websites of the councils across the Greater Manchester area, which has taken me on a journey down a rabbit hole of web archiving and Bash scripting.

Prior to setting off on this coding journey, I attempted to simply save these web pages as pdf or html documents, with no luck. The files failed to accurately represent how the sites were displayed, did not handle hyperlinks well and only captured a single page at a time. So, I quickly decided I would need something a bit more dependable and usable.

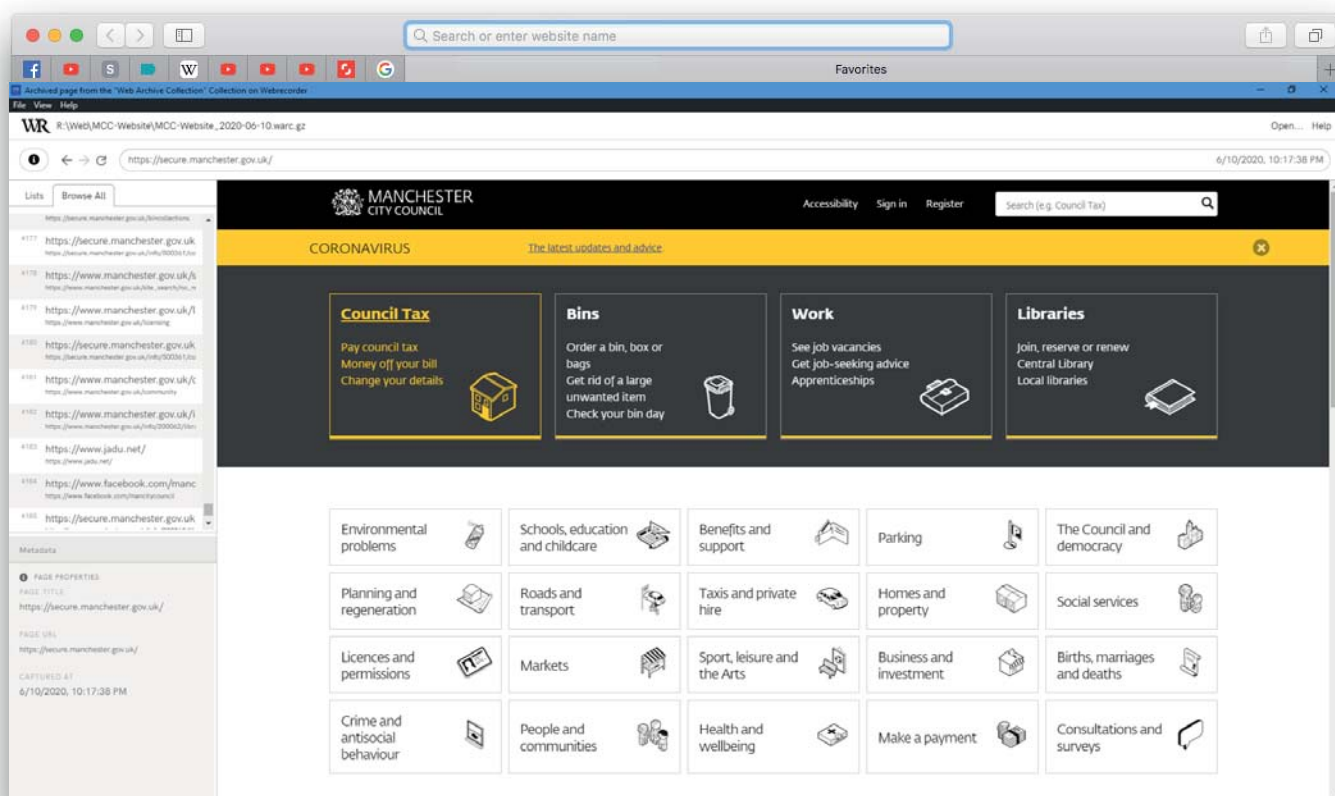
At this point in time, my knowledge of web archiving was scrappy. I knew that there existed, in some distant corner of the internet, software called web crawlers, which might hold the answer to my web archiving

prayers. But I had little idea how to use such tools, or even where to find them.

Venturing deeper into google searches, I stumbled upon the treasure trove that is the ArchivesTeam wiki. Here, I discovered the whereabouts of these elusive web crawlers and how to put them to good use. I also learnt that such crawlers operate by working their way through the hierarchies of a web address to capture all the links required to give an accurate representation of the site.

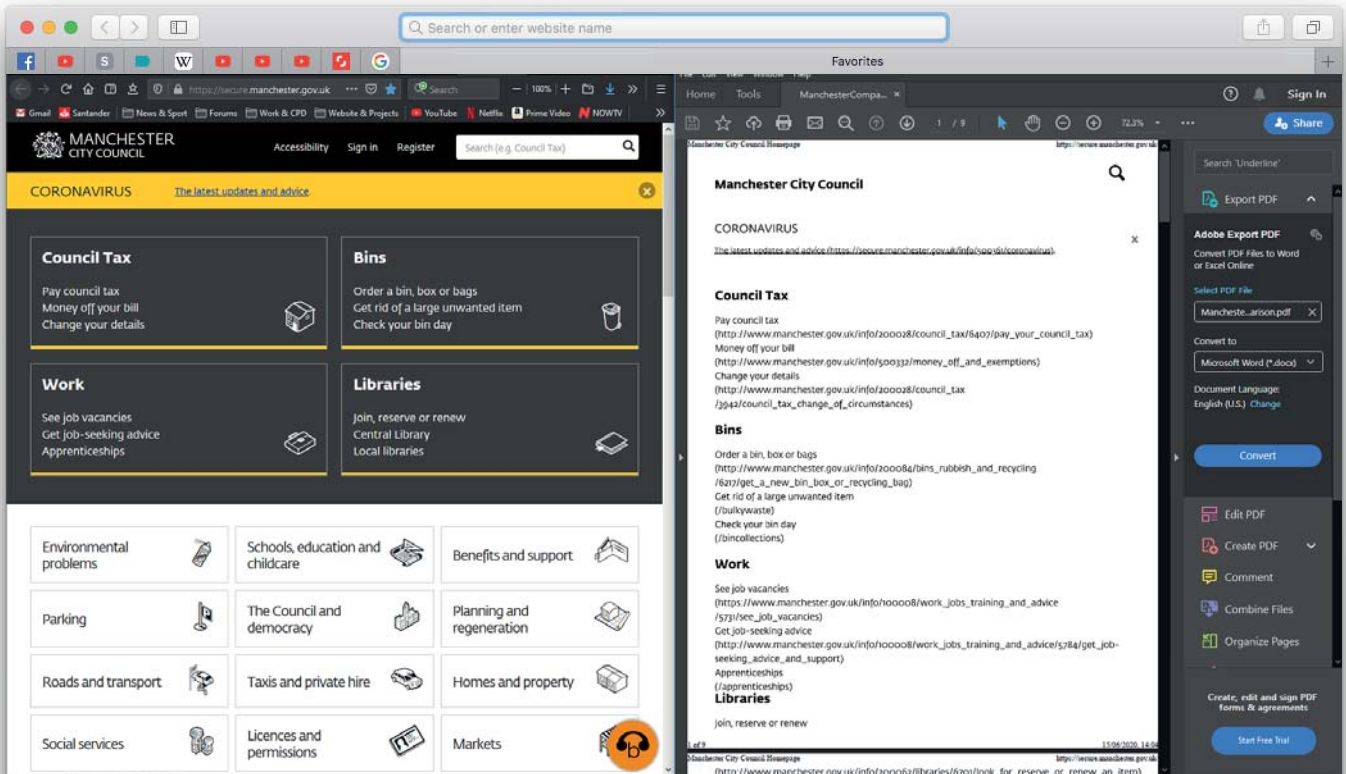
I also found that these tools were Linux based command line programmes. As fortune would have it, I am a big computer nerd and conveniently had a spare Linux PC ready to go!

After getting everything setup, which took a bit of tinkering, I set course to create my first Web ARChive

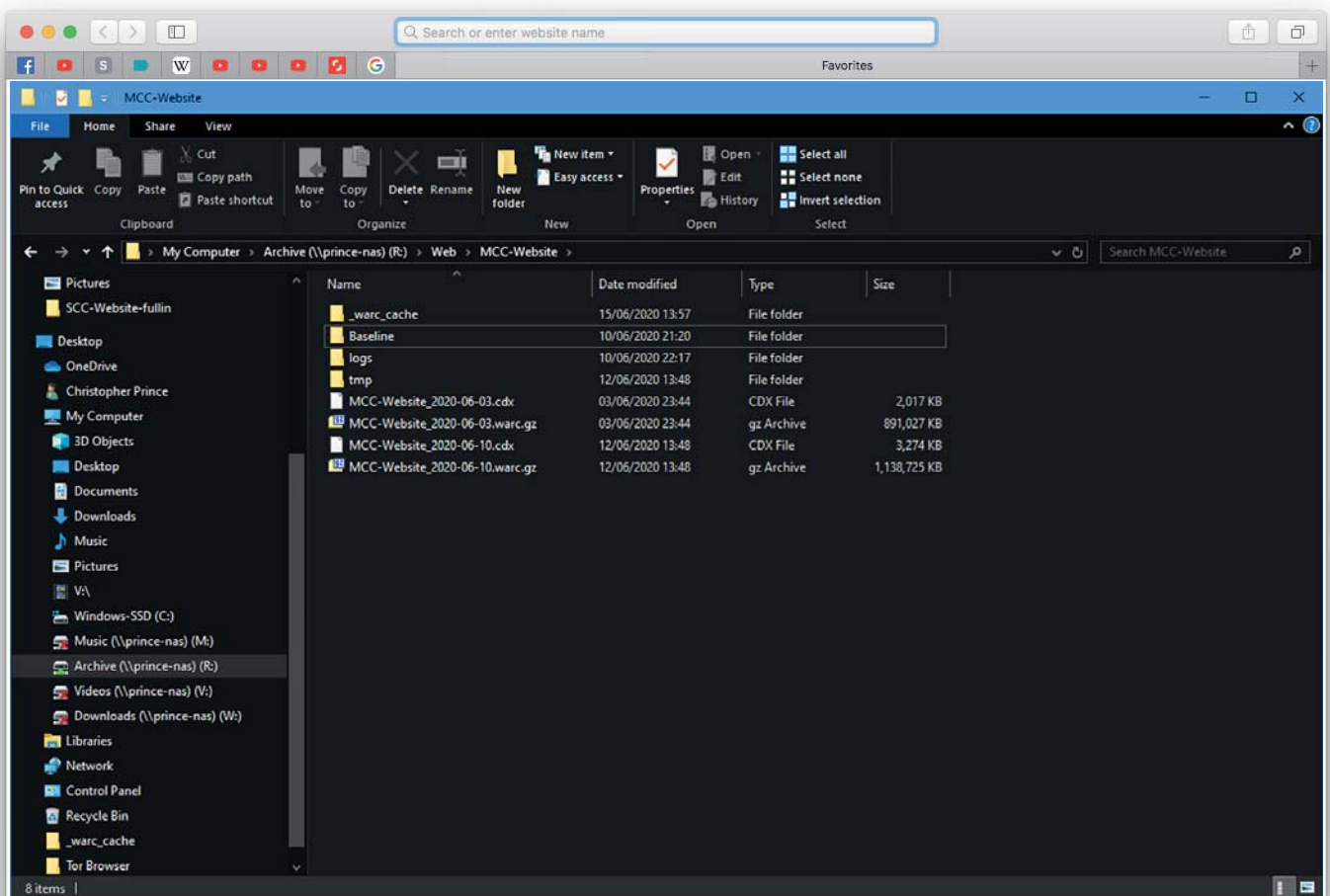


The finished WARC! Courtesy of Christopher Prince (Pyweb - displaying WARC by Conifer)





A comparison of the MCC Council on a browser and as a PDF. Courtesy of Christopher Prince



The files and folder structure that I've set out - as they appear from Windows on the server. Courtesy of Christopher Prince



*I knew that there existed, in some distant corner of the internet, software called web crawlers, which might hold the answer to my web archiving prayers*



format (WARC) using a tool from the ArchiveTeam called 'wpuull'. It took some trial and error to get used to the command line interface and get all the configurations correct, but by using the 'wpuull' tool, I was able to accurately capture a number of the Greater Manchester council's websites, with the display of the site and its hyperlinks intact.

However, 'wpuull' has its limitations. It does not provide a way to manage the newly created WARCs, nor does it provide a way to manage the configuration settings within the program itself. It also does not provide a way to set a schedule if you want to repeat the process. So, to attempt to address these issues, I set course into the seas of Bash scripting.

You might be wondering: what is a Bash script? Put simply, Bash is a way of communicating with your computer; what is known as a programming language. It allows you to input commands directly to the PC and receive the output of these commands. A script, on the other hand, is a set of those commands typed out and saved as a file, so that the instructions can be referred to without having to type them out again. What I had to figure out was the set of instructions and the order of the instructions that the computer needed to address the issues I was facing.

```

GNU nano 4.8 WARC_Script.sh
function appendinit {
while true; do
    echo "Debug: ConvarCheck=$convar"
    if [[ $convar == "y" ]]; then
        break
    else
        read -p "Do you want to run append?: Y/N " convar
        case $convar in
            [yY][eE][sS][sS])
                app="--warc-append --warc-dedup=$WARCTITLE.cdx -a $LOGLOC.db"
                logap="-a"
                break
                ;;
            [nN][oO][oO])
                echo "No selected - Exiting Operation"
                sleep 1
                exit 1
                ;;
            *)
                echo "Invalid option - please confirm" >&2
                ;;
        esac
    fi
done
}

function schdcheck {
while true; do
    if [[ $scptit -lt $scmpschr ]]
    then
        echo "WARC File init on:$fmsch"
        echo "Schedule not due until: $fmsch"
        sleep 0.5
        echo "Running appendinit"
        appendinit $0
    fi
done
}

break
}

Get Help Write Out Where Is Cut Text Justify Cur Pos Undo Mark Text To Bracket
Exit Read File Replace Paste Text Go To Line Redo Copy Text Where Was

```

My script in action! Complete with debugs! Courtesy of Christopher Prince

```

OpenSSH SSH client
0 updates can be installed immediately.
0 of these updates are security updates.

Last login: Fri Jun 12 11:44:43 2020 from 192.168.0.9
localadmin@princenas:~$ cd WebArchive
localadmin@princenas:~/WebArchive$ cd Scripts
localadmin@princenas:~/WebArchive/Scripts$ ./WARC_Script.sh secure.manchester.gov.uk MCC-Website fullex inf 60 www.manch
ester.gov.uk,manchester
debug filecheck:/pool/str/Archive/Web/MCC-Website/MCC-Website_2020-06-10.warc.gz
debug filedate:2020-06-10
debug schedule date: 2020-08-11
MCC-Website_2020-06-10.warc.gz already exists
Running schdcheck
WARC File init on:2020-06-10
Schedule not due until: 2020-08-11
Running appendinit
Debug: ConvarCheck=
Do you want to run append?: Y/N y
Running precheck
Settings Confirmation: LevelCheck=inf, ModeSettings=--span-hosts-allow linked-pages,page-requisites -r -l inf, AppendChe
ck=--warc-append --warc-dedup=/pool/str/Archive/Web/MCC-Website/MCC-Website_2020-06-10.cdx -a /pool/str/Archive/Web/MCC-
Website/Logs/MCC-Website_2020-06-10.db, Acceptlist=secure.manchester.gov.uk,www.manchester.gov.uk,manchester
Debug: PrevarCheck=
Are these settings correct?: Y/N y
Intialising wpuullinit, profile: fullex
Running loginit
Running schdinit
Running wpuullinit

```

A small section of my script, just to give you an example - this 'append check' portion gives me a prompt to tell me that a file exists and whether I want to run an 'append' on the existing WARC. Courtesy of Christopher Prince

```

OpenSSH SSH client Windows PowerShell
00 OK. Length: 44679 [text/html; charset=utf-8].
INFO Fetching 'https://www.google.com/url?http://togethermcrcr.com&source=gmail&ust=1598564907981008&usg=AFQjCNGLG05MITLkzo7i_BmpEvLj0u1g'.
[ 0 ] 5.0 B 0:19:43 -0.0 B/s
INFO Fetching 'https://www.google.com/url?http://togethermcrcr.com&source=gmail&ust=1598564907981008&usg=AFQjCNGLG05MITLkzo7i_BmpEvLj0u1g': 200 OK. L
ength: unspecified [text/html; charset=utf-8].
INFO Fetching 'http://togethermcrcr.com/'.
INFO Fetching 'http://togethermcrcr.com/': 301 Moved Permanently. Length: 0 [text/html; charset=utf-8].
INFO Fetching 'https://togethermcrcr.com/': 301 Moved Permanently. Length: 0 [text/html; charset=utf-8].
INFO Fetching 'https://www.instagram.com/static/images/payments/fbpaylogofooter.png/84d68686fde3.png': 200 OK. Length: 143150 [text/html; charset=utf-8].
100.0% [=====] 139.0 KiB 0:19:51 126.8 B/s
INFO Fetching 'https://www.instagram.com/static/images/payments/contactinfo.png/930cfc393ff6.png': 200 OK. Length: 3232 [image/png].
INFO Fetching 'https://www.instagram.com/static/images/payments/paymentinfo.png/d2083fc98d03.png': 200 OK. Length: 3618 [image/png].
100.0% [=====] 3.5 KiB 0:19:53 -14.8 B/s
INFO Fetching 'https://www.instagram.com/static/images/payments/paymentinfo.png/d2083fc98d03.png': 200 OK. Length: 3618 [image/png].
INFO Fetching 'https://www.instagram.com/static/bundles/metro/DonateCheckoutPage.js/images/payments/document.png': 404 Not Found. Length: 0 [text/html;
charset=utf-8].
INFO Fetching 'https://www.instagram.com/static/images/payments/fbpaylogofooter.png/84d68686fde3.png': 200 OK. Length: 143150 [text/html; charset=utf-8].
100.0% [=====] 3.2 KiB 0:19:54 95.5 B/s
INFO Fetching 'https://www.instagram.com/static/images/payments/securedonation.png/203133c6984a.png': 200 OK. Length: 2572 [image/png].
100.0% [=====] 2.5 KiB 0:19:54 -59.9 B/s
INFO Fetching 'https://www.instagram.com/static/images/payments/securedonation.png/203133c6984a.png': 200 OK. Length: 2572 [image/png].
INFO Fetching 'https://www.instagram.com/static/bundles/metro/DonateCheckoutPage.js/images/payments/fbpaylogoheader.png': 200 OK. Length: 143150 [text/html;
charset=utf-8].

```

wpuull in action fetching resources, as seen from the Command line. Courtesy of Christopher Prince (wpuull by ArchiveTeam)



Now, as someone without extensive experience in scripting, this took some time to get used to. Sparing you the technical details of my scripting trials and tribulations, I eventually built up my script bit-by-bit to a (semi-)finished state. Ultimately my script performed a number of functions: it allowed the use of pre-set configurations for wpull to refer to; it created a folder structure for the WARC files to be stored in; and it allowed a schedule to be set for the websites.

At the end of my journey, I was quite happy with what I achieved using 'wpull' and Bash and I would hope to continue using these tools and the knowledge I've gained to further build our web archiving capacity in Greater Manchester.

For anyone interested in finding out more, please see the following links to resources discussed above and additional ones that were particularly useful on my journey:

WARC/Webrecorder's Reader: [conifer.rhizome.org/](http://conifer.rhizome.org/)

ArchiveTeam: [archiveteam.org/index.php?title=Main\\_Page](http://archiveteam.org/index.php?title=Main_Page)

Wpull: [github.com/ArchiveTeam/wpull](https://github.com/ArchiveTeam/wpull)

Wget: [www.gnu.org/software/wget/](http://www.gnu.org/software/wget/)

For getting started with Bash, the questions and answers section on stack overflow proved a great resource: [stackoverflow.com](http://stackoverflow.com)

For understanding statements in Bash: [linuxacademy.com/](http://linuxacademy.com/)

For a general Bash introduction: [ryanstutorials.net/bash-scripting-tutorial/](http://ryanstutorials.net/bash-scripting-tutorial/)

For understanding Conditions and Parameters: [likegeeks.com/linux/](http://likegeeks.com/linux/)

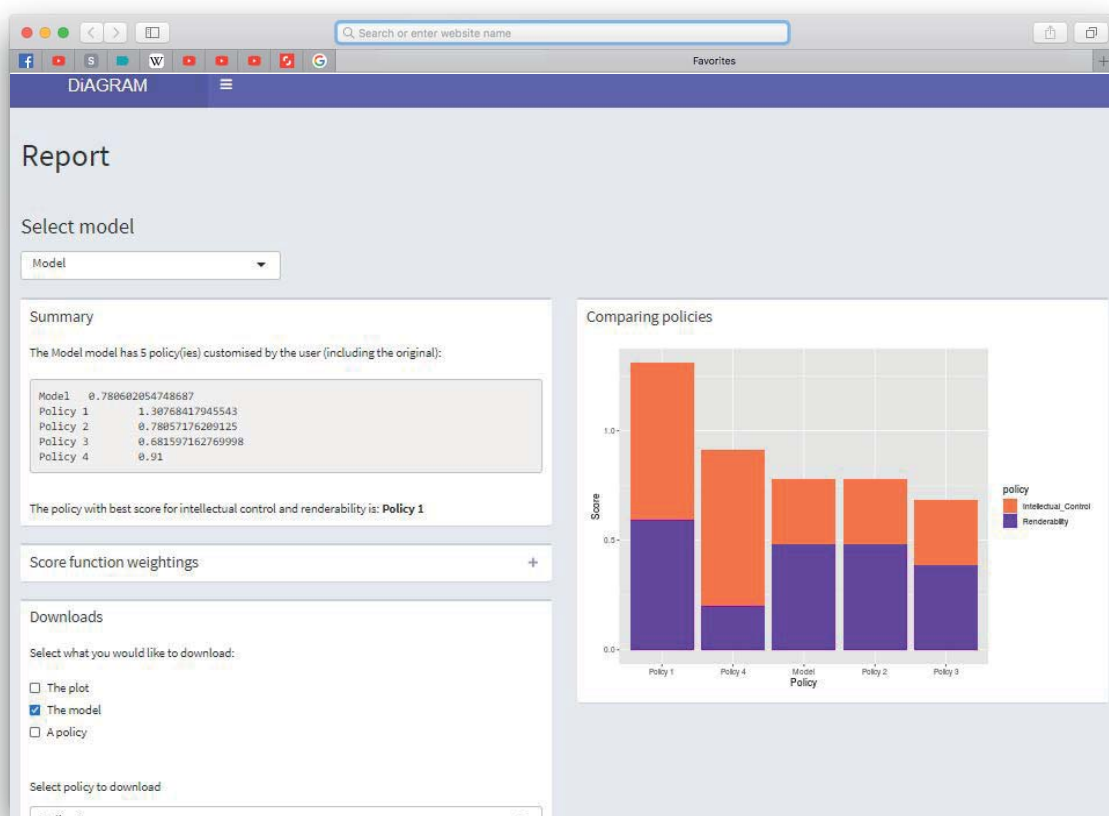
# Trying to work out your digital preservation risks? Now you can DiAGRAM them

**Hannah Merwood** and **David Underdown**, of The National Archives UK (TNA UK), describe the development of a new, accessible tool for measuring an archive's digital preservation risks.

TNA UK has been working with UK archive partners and academics from the University of Warwick to develop a statistical model that quantifies the risks involved in digital preservation. With funding from The National Lottery Heritage Fund and the Engineering and Physical Sciences Research Council, the project began in January and we now have a prototype version of this web-based tool freely available, with the first official version due to be released in the autumn.

We know that preserving digital materials involves protecting them from a wide (and evolving) set of risks. These risks are diverse, from the technical (obsolescence of equipment or file format, lack of skills to identify file formats and technical metadata) to more general risks of lack of resource and inadequate cataloguing. It's therefore very hard to directly compare the different risks and prioritise mitigating actions, or to build business cases and gain stakeholder support for funding.

Staff from TNA UK, Dorset History Centre, Gloucestershire Archives, Transport for London Archives, University of Brighton Design Archives, and University of Leeds Special Collections have been collaborating with statisticians at the University of Warwick's Applied Statistics & Risk Unit to produce DiAGRAM (the Digital Archiving Graphical Risk Assessment Model): a free online tool, where archivists can measure and compare their digital preservation risks. Just answer a few questions about your digital holdings such as the proportion of digital surrogates and born digital records you have and the type of data storage you use, and the tool will report on your current risk levels. You can then experiment by applying different changes to your digital preservation practices to see the effect on your overall risk level. The different risk scores are graphed so you can make comparisons and include the outputs in business cases or reports. Unsurprisingly, COVID-19 has impacted the way we've had to run the project. Back in January 2020, we began with an in-person workshop for



A snapshot of the DiGRAM report page, comparing the scores of different policies. Courtesy of The National Archives UK (TNA UK)

the project partners at TNA to develop the basic model, and explore the interconnections between different risks. We also determined that our key outputs would be risk scores for the renderability of the digital materials being preserved, and the degree of intellectual control the archive had established over the material.

As we entered March, we were right in the middle of planning our next workshop. This time we were aiming to obtain data for the model from the knowledge and experience of digital preservation experts as the data was not available from any other source. This workshop would use a scientific technique known as the IDEA protocol - to convert expert judgement to quantitative data. The arrival of the pandemic meant that there was no way we could carry out the workshop in person. With over 20 people involved it was not simple to convert to an online event, particularly with the amount of discussion required to ensure that everyone had a similar understanding of the questions and could share the reasoning behind their initial answers. However, we were pleased that we managed to hold the event on the same dates as planned and with more experts than we had expected for the physical gathering, which makes our data even more robust.

We processed this data and finished the prototype model, ready for the team to then test the tool and open it up to other archivists to explore. We had expected to be attending various digital preservation conferences to introduce the model including at the ARA Conference 2020, as well as running workshops of our own, but alas the coronavirus changed our dissemination plans too.

*“Feedback from the first online workshop suggests that users could really see the value of the model and they gave great feedback”*

The Digital Preservation Coalition (DPC) has a formal role in the project around organising events (and carrying out evaluation), so we worked closely with them to adapt our plans for physical workshops into virtual ones. This did pose new challenges, including trying to teach people a new tool without seeing their screens and explain complex statistics without seeing their non-verbal feedback, but it has enabled us to reach a wider audience including international organisations, and all three of our online events sold out weeks in advance. Feedback from the first online workshop suggests that users could really see the value of the model and they gave great feedback, both to improve the tool, and to make the second and third workshops even better. We hope that this level of interest will continue, and that the tool really will be of great use to archivists. It also highlights areas where we have not been collecting data as a community, so hopefully will inspire us to do this now and we can feed this new information into the model, to have an even more solid evidence base for digital preservation. For more information and a link to the tool itself, see: [bit.ly/DiGRAM](https://bit.ly/DiGRAM).





National Museum of Ireland Central Archive, Collections Resource Centre, Swords, Co. Dublin.  
Courtesy of the National Museum of Ireland

# Remote working at the Archives and Records Service, National Museum of Ireland

**Emer Ní Cheallaigh** of the National Museum of Ireland (NMI) recounts the challenges of, and opportunities emerging from, the adjustment to remote working.

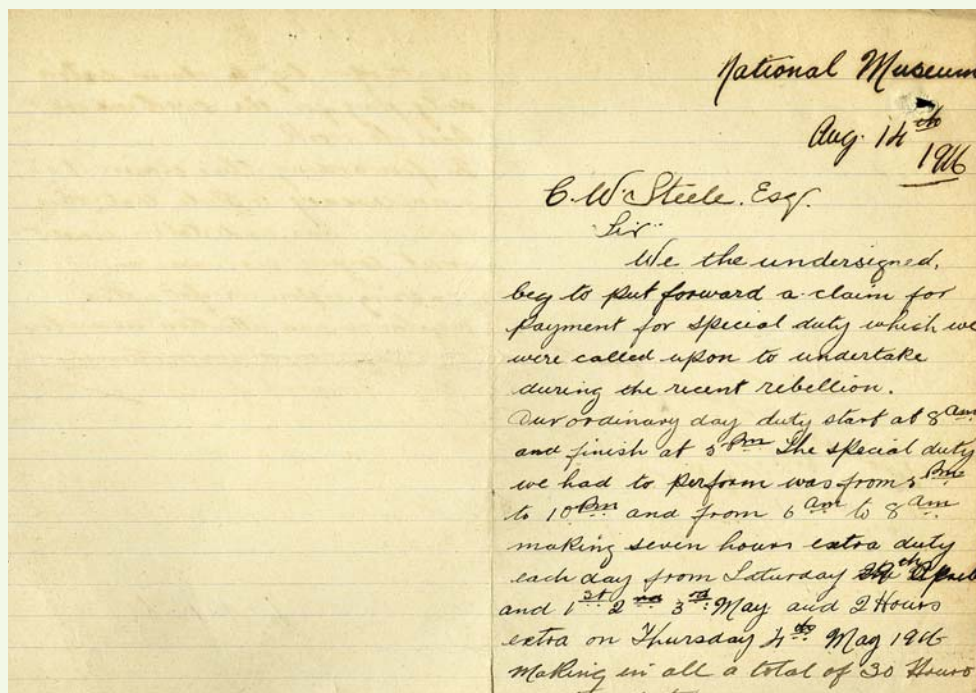
Little did I know when I was employed as Archives and Records Manager at NMI in 2008 that I would one day find myself working remotely from the wilds of County Donegal! My family and I had travelled from Dublin to my mother's house in Donegal for the Saint Patrick's weekend when the travel restrictions were announced, and having weighed up all the pros and cons and following a discussion with my line manager, I established the first ever Donegal branch of the NMI's Archives and Records Service!

NMI is a multi-site, multi-county operation. Prior to the introduction of the restrictions however, the facility of remote working did not exist for us. Our colleagues in

ICT were in the process of introducing the infrastructure required to support remote working via government networks but the shutdown of all NMI sites in March introduced Remote Working (RM) at the speed of a runaway train. It is fair to say that RM hit us before we were fully prepared!

In my case, I had to install WIFI in my mother's deliberately technology-free-zone, get my work laptop sent from Dublin and set about reviewing my business plan to identify potential work packages for my team members and for myself. As the only archives and records staff member at NMI I rely on procurement of archival services from external companies for project-





Extract of a letter from the 'A Files', Archives of the Directorate, National Museum of Ireland requesting special payment for staff on duty during the 1916 Easter Rising. Courtesy of the National Museum of Ireland

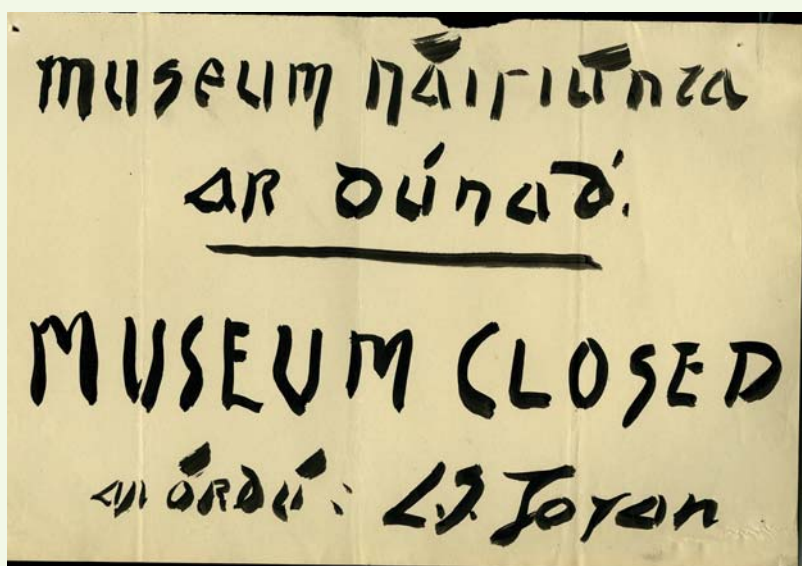


Natural History Museum Archive, Collections Resource Centre, National Museum of Ireland. Courtesy of the National Museum of Ireland

“The project has been underway for one month and has worked well so far, with 500 records ready for online publication”

based work. My team normally consists of archivists (outsourced), interns, volunteers and a digital imaging officer. Their work is typically focused on tackling the backlog of archival listing. Our volunteering and internship programmes were suspended so the priority was to redirect the services of our two archivists to produce a viable and useful project that would be possible to execute with one archivist locked-down in Enniskillen and the other in Dublin! I was fortunate that both of my colleagues had antique work laptops, access to WIFI and that they were willing to engage in new projects remotely.

I immediately thought of a project that I had been tackling solo and slowly. As part of NMI's Online Collections Project (OCP), I had been preparing catalogue records for online publication through the use of middleware software drawing from all NMI catalogues (MS Access, Adlib Museum, Archive and Library, and from Portfolio Extensis). Knowledge Integration (KI) were engaged to produce an Application Programme Interface (API) to extract data from all NMI finding aids with a view to publishing them online via NMI's recently upgraded website. One of my first tasks was to prepare a test record for the KI to allow them to trial a test extraction of data from Adlib Archive (this extraction is scheduled but has not yet taken place).



Notice of closure from 'A Files', Archives of the Directorate, National Museum of Ireland, placed on the door of the museum on the outbreak of 1916 Easter Rising



Transfer of archives to National Museum of Ireland Central Archive, Collections Resource Centre, Swords, Co. Dublin. Courtesy of the National Museum of Ireland



“Prior to the introduction of the restrictions... the facility of remote working did not exist for us”

I selected 1,500 records of the archives of the NMI Directorate for the project. I produced a project plan and audit/edit checklist for use by the archive team containing data to be reviewed, standardised and, if required, upgraded. The data and fields are checked against an in-house minimum cataloguing standard that I tailored to ensure compliance with ISAD(G). We hold team meetings to ensure that we are sufficiently familiar with the project and its requirements, including an agreed daily target of 30 records to be reviewed. The team also identifies major global edits required and records which are not appropriate for online publication. I review all global edits and sensitive records and in the case of records that should be closed, tick the in-built 'web exclusion' tick box. This facilitates a double layer of scrutiny.

The project has been underway for one month and has worked well so far, with 500 records ready for online publication. It has not been without its challenges! Poor internet connection for one colleague has led to continuous locking of records. For me, balancing remote working with the challenge of homeschooling has not been easy and has led to irregular work patterns. The repetitive and sometimes tedious nature of the work can be magnified in the remote working environment; without face to face interaction with colleagues and I am conscious of trying to maintain a remote sense of 'team'.

In addition to holding online meetings via MS Teams we have had a couple of virtual cuppas, which have helped support the team and our sanity! On the upside, while COVI-19 has turned the work of NMI on its head, we have been presented with an opportunity to focus on a project that will support our statutory function as a place of deposit under the National Archives Act, 1986. The preparation of the records for online publication will extend the reach of the archives of the NMI Directorate and will enhance the accessibility of the collection. It will also function as a test case for the preparation of all NMI Archive catalogues for online publication going forward.

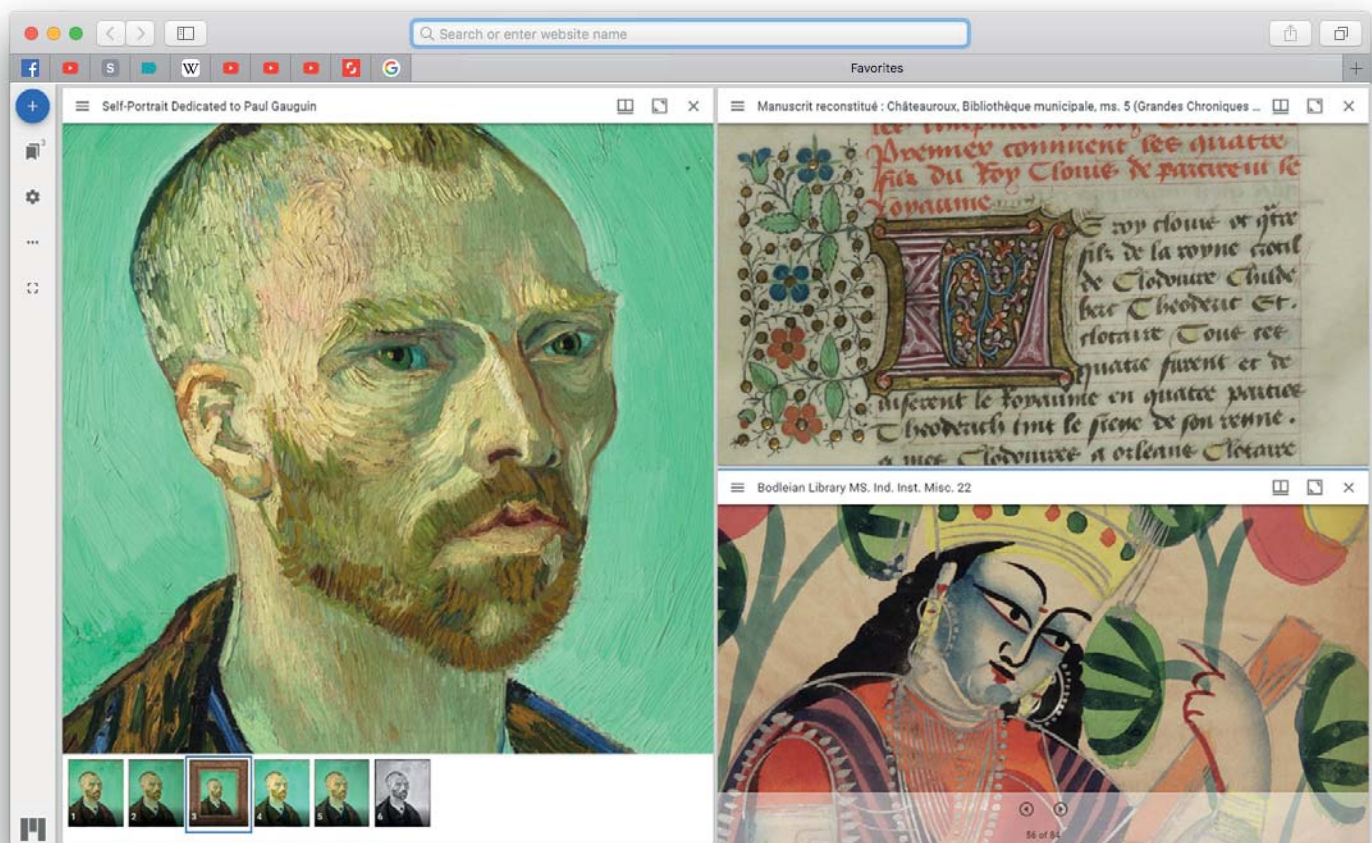
# From self-isolation to self-development

**Alison Harvey**, of Cardiff University, reviews free technical training offered during the lockdown period, including editing for Wikipedia, machine-learning, and the IIIF digitisation standard.

The national shift to remote-working in March took us all by surprise. Like many busy services, it was not something the Special Collections and Archives team at Cardiff University were accustomed to, or able to perfectly prepare for. Challenges abound, but one redeeming feature of the times we are living through, is the ability to take advantage of more self-development opportunities than would normally be possible with limited travel budgets and staff cover. A huge range of technical training has been made available for free over the last few months, which has opened doors to future online engagement.



Contributors at a public Wikipedia Edit-a-thon to mark International Women's Day



IIIF viewers can be embedded in web pages to analyse and compare images from any repository side-by-side

**Transkribus** Register Login

# Transcribe. Collaborate. Share...

...and benefit from cutting edge research in Handwritten Text Recognition!

Download version 1.11 Download version 1.11 for Mac Wiki » How-to guide (pdf) »

## Scholars

Are you transcribing historical documents? Handwritten or printed, from the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and reliable way? And do you appreciate to get support from automated tools such as Handwritten Text Recognition and Layout Analysis?

[View details »](#)

## Archives

Are you responsible for large collections of handwritten and printed documents? Do you believe that digitisation paves the way to realise new opportunities to access, enrich and explore archival material? And are you open to involve humanities scholars and volunteers so that they can work with these documents in an effective way – producing data which can also be integrated in your repository?

[View details »](#)

## Volunteers

Are historical letters, postcards, manuscripts or medieval documents fascinating for you? Do you enjoy deciphering handwriting – this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to scholarship and science?

[View details »](#)

## Scientists

Are you a computer scientist and working in the fields of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and publishing? And you would really enjoy to get some reference data in a well-acknowledged format, such as PAGE?

[View details »](#)

Transkribus software can be freely downloaded. Volunteers can access a web version



In April, Wikipedians Kelly Foster (Public Historian and founding member of TRANSMISSION) and Jo Pugh (Digital Development Manager at the National Archives) organised a free webinar aimed at archive employees ([bit.ly/wikipedia-archives](https://bit.ly/wikipedia-archives)). Rather than a practical session on how to edit Wikipedia, it focused on exploring the specific opportunities and challenges for archives.

We know that sharing information about collections via Wikipedia can massively boost search engine findability, but archive employees need to be cautious. Wikipedia enforces conflict of interest rules aimed at preventing business promotion, which made me wonder how I could add contributions as a paid employee. Kelly and Jo explained that there is scope for archive workers to write about collections - carefully - by signing up as an individual rather than with an institutional account, being transparent about affiliation, and focusing on content about collections rather than the parent institution.

I was also unclear how to provide adequate citation when dealing with archives. Guidance for Wikipedia editors stresses the importance of supporting every contribution with citations from respected published sources, such as academic textbooks. How can information that exists in archives be contributed, if it has never been published? The answer perhaps shouldn't have surprised me as much as it did - we should cite from our archive catalogues. The detailed records that archivists create in the course of our descriptive work, and host online, are in fact 'respected, published sources'. There are ways in which archives can contribute without contravening guidelines; enhancing records for individuals and organisations represented in our collections, as well as engaging students in innovative assessment activities conducive to remote teaching.

In May, C. Annemieke Romein of Huygens Institute offered two free, in-depth sessions on Transkribus ([bit.ly/transkribus-huygens](https://bit.ly/transkribus-huygens)). This machine-learning software looks for patterns in manuscript transcriptions, in order to read examples of the same handwriting. If your archive holds a large quantity of documents written by the same person, this is well worth looking into. Transkribus can host images of manuscripts and provide a line-by-line transcription interface. Once as few as 50 pages have been transcribed, it is possible to start 'teaching' a model how to read the author's handwriting. The more pages added, the better the accuracy.

I transcribed the minimum 50 pages from our collection of 500 letters from the poet Edward Thomas, and the resulting model was able to transcribe unseen letters with 92% accuracy. The results can be corrected and fed back into the model to improve it further. A

*“Focusing on self-development has been critical to helping me maintain a positive, forward-looking outlook throughout an extremely challenging few months”*

web interface enables anyone, anywhere, to contribute to transcriptions and corrections; yet another way in which we could promote remote, crowdsourced engagement with our collections, and enhance their findability.

Perhaps the ultimate means of sharing content with the world is the International Image Interoperability Framework (IIIF). IIIF-compliant image and AV files can be re-used by anyone, anywhere online without losing their digital provenance. In June, the IIIF Consortium offered all their annual conference workshops for free ([bit.ly/iiif-week](https://bit.ly/iiif-week)). Following the conference, I had the opportunity to take part in a week-long course supported by IIIF experts around the world.

I learned how to install and use a local image server to convert standard digital images to deep zoom-enabled, IIIF-compliant files. I also discovered it was possible to embed a Universal Viewer in a webpage, to preview, annotate, and compare IIIF content held anywhere in the world. This is all possible for free via GitHub, with no infrastructure required beyond a personal computer. These skills will allow us to enhance our digital provision, remote teaching, and research support service, by enabling interaction with collections held in IIIF-compliant institutions worldwide, including the National Library of Wales, British Library, Harvard, Stanford University, and the Smithsonian.

I would like to thank all those individuals and organisations who have contributed their time and expertise for free during difficult circumstances and acknowledge their impact. Focusing on self-development has been critical to helping me maintain a positive, forward-looking outlook throughout an extremely challenging few months for myself and my colleagues. In addition, these learning opportunities are opening up new areas for us to explore as a service, as we consider how best to adapt to and navigate the changing landscape of our sector.



Fleadh na Leanbh [Children's Fete, part of annual Oireachtas competition] 1909, complete with children in costume. Courtesy of James Hardiman Library Archive, NUI Galway: Conradh na Gaeilge collection: G60

## Conradh na Gaeilge collection

**Niamh Ní Charra** and **Aisling Keane** of the National University of Ireland (NUI), Galway, discuss the creation of a digital exhibition and blog from the records of an historically significant collection.

In 2018, Conradh na Gaeilge (The Gaelic League) donated their archive to NUI Galway. Project archivist Niamh Ní Charra has had the task of processing this collection consisting of the equivalent of 850 bankers boxes, which is due to be completed in 2021.

Conradh na Gaeilge was founded in 1893 to promote the Irish language in Ireland and abroad, and the collection encompasses material from founding through to 2018. Following on from several 19th century organisations promoting the Gaelic Revival, Conradh na Gaeilge became the main organisation to spearhead the Irish language revival. Its existence has subsequently coincided and overlapped with the revolutionary years,

up to and including the War of Independence, The Civil War and the foundation of the Irish Free State, the early years of the Republic, the turbulent "Troubles" which saw human rights issues and political strife come to the fore in Northern Ireland, through to the emergence of the Celtic Tiger and beyond. It therefore gives a unique insight into the history of the island of Ireland, with many of its members prominent in a wider context.

“Following on from several 19th century organisations promoting the Gaelic Revival, Conradh na Gaeilge became the main organisation to spearhead the Irish language revival”





Window display 1944 of Craobh an Chéitinnigh [Keating branch]. Courtesy of James Hardiman Library Archive, NUI Galway: Conradh na Gaeilge collection: G60

The main body of material comprises records generated by the organisation's headquarters, although it also includes material returned to them by various branches in Ireland and abroad, most notably the London branch.

Staff were actively involved in promoting and observing the use of Irish across all aspects of everyday activities, and as such the collection includes material relating to the running of classes and events, to Irish language publications, to correspondence with businesses, Government departments, and members of the public, and to research carried out on minority languages elsewhere.

Material also covers several language rights and fundraising campaigns. Conradh na Gaeilge was instrumental in community campaigns which led to the creation of Irish language radio and television stations, to the enactment of the Official Languages Act, and to Irish becoming an official language of the European Union. Documents also relate to



Máire Ni Rathallaigh [Mary O'Reilly]'s membership card for Parish of North Cork branch, 1904/05. Courtesy of James Hardiman Library Archive, NUI Galway: Conradh na Gaeilge collection: G60

“These blogs, while giving readers a sneak peak into the contents of the collection, have become a vital tool in expectation management while work continues”

prisoners' rights and civil rights in Northern Ireland, in particular during the period of the Hunger Strikes.

Due to the size of this collection the overall project was divided into three sections:

- The First Pass: Box-listing, appraising and rehousing the collection.
- Arrangement: None existed and it was therefore necessary to impose one.
- Cataloguing: Collection and Series Level descriptions are complete and work continues on listing to lower levels of hierarchy.

Because of the interest in and vastness of this collection, rather than waiting until the entire collection is described fully, the decision was made to release parts of the collection in tranches, as they are completed.

A blog for the project was set up and several entries have been written over the course of the project to inform readers of the work taking place behind the scenes before the collection is made accessible, and of the value and need for this work. These blogs, while giving readers a sneak peak into the contents of the collection, have become a vital tool in expectation management while work continues.

An extension of this outreach work was the decision to design an online mini-exhibition with descriptions of the collection, the imposed arrangement and each of the 45 individual series. This has allowed researchers to both understand the context of what is already accessible and to be aware of what is yet to come.

Working closely with Niamh who provided descriptions, metadata and just under 50 carefully selected images, digital archivist Aisling Keane took on the exhibition task. Though the images were compressed for the purposes of the digital exhibition, they were digitised in high resolution in the first instance, to allow for them to be more securely preserved and managed in the Library's digital repository.

The exhibition is hosted on NUI Galway Library's exhibitions platform which is built on an Omeka web-publishing site. The ISAD(G) fields created by Niamh were mapped to corresponding Dublin Core fields. In all cases, the Title, Description, Identifier, and Publisher fields were completed and for the series now accessible, an additional field, Relation, was included, populated with a link to NUIG's archives catalogue to enable further exploration.

The bilingual site was launched on 17 March 2020 with an accompanying blog advising users on how best to use both the exhibition and the archives catalogue.

While work continues on cataloguing this amazing collection, the online exhibition and the blogs all contribute to users' understanding of it, the work of the archivist, and the context of each release. It has also proved invaluable during the COVID-19 lockdown, as a means of promoting the collection and future access.

The exhibition site can be found at [exhibitions.library.nuigalway.ie/s/cnag/](https://exhibitions.library.nuigalway.ie/s/cnag/) and the accompanying project blog at [nuigarchives.blogspot.com/2020/04/conradh-na-gaeilge-digital-exhibition.html](https://nuigarchives.blogspot.com/2020/04/conradh-na-gaeilge-digital-exhibition.html).

# The 'What was here?' mobile app

**Sam Bartle** of the East

Riding Archives recounts the development of a mobile app connecting modern users with geographies of the past.

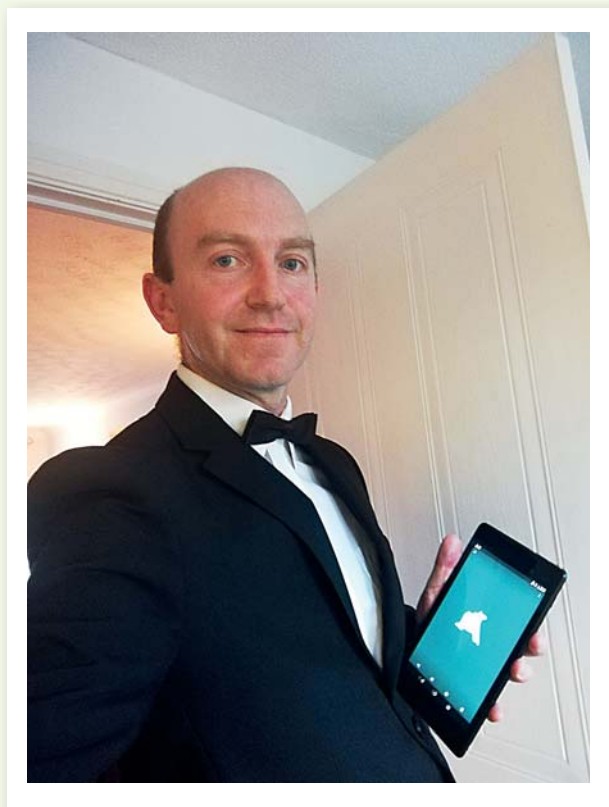
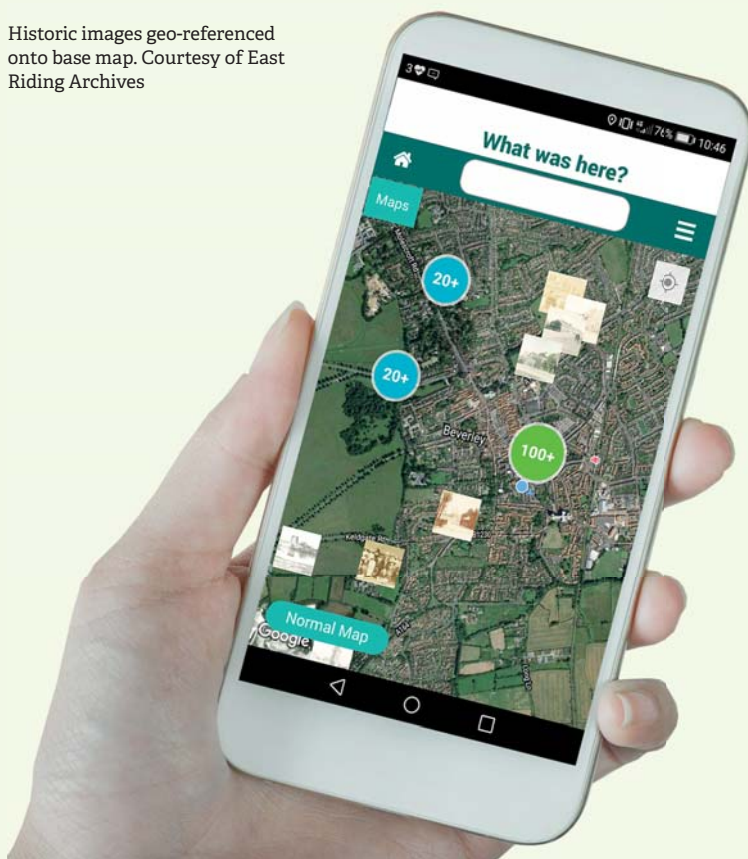
Everyone should enjoy archives at our fantastic research facilities, but it's a harsh (and perhaps uncomfortable) truth that many people will never make that physical trip to visit us. In this digital age the 'hard copy' nature of most original archives inevitably creates physical barriers to access. Traditionally, we wanted to entice people to visit us and explore our archives - but what if we turned that around by taking archives out of the repository, and encouraged people to explore their surroundings?

Through my remit for Digital at East Riding Archives (Treasure House, Beverley, East Yorkshire), I wanted to break down those physical barriers and deliver archives into the hands of potential users. In June 2015, I found my answer: A location-based mobile solution for accessing geo-referenced historic photos and maps, through self-directed exploration and discovery of what existed in the user's environment (with the inclusion of heritage trails). I called it 'What was here?' My ambition was for an intuitive mobile discovery experience, empowering users with archives, be that for general tourism, ancestral tourism (family history), formal education, supporting the vulnerable, encouraging good health and wellbeing, or simply local interest.

It took 3 years to reach the delivery stage, working with support from the Group Manager and ICT business analysts on producing technical specifications and drafting business cases to gain approval for procurement. The work finally began in April 2018 with outsourced developers, but there were a number of technical challenges. I naively assumed that for the historic maps we could supply raw image files, but I quickly learned that the developers needed me to crop the images (100+ files) and obtain accurate decimal coordinates for each tile to be overlaid onto the base map. I did this using Google Earth and, being something of a



Historic images geo-referenced onto base map. Courtesy of East Riding Archives



'Virtually' attending the Tourism awards (at home)!  
Courtesy of East Riding Archives

map geek, actually quite enjoyed aligning our maps (as closely as possible) to the modern satellite imagery!

However, there was still an issue where the map overlays would freeze and become unloadable, so our focus turned to the historic photographs element. We already had these geo-referenced so it was a case of sending the data and 1422 images of the East Yorkshire region (also including Hull and Scarborough) to the developers, who uploaded the files for me to edit in the Wordpress CMS.

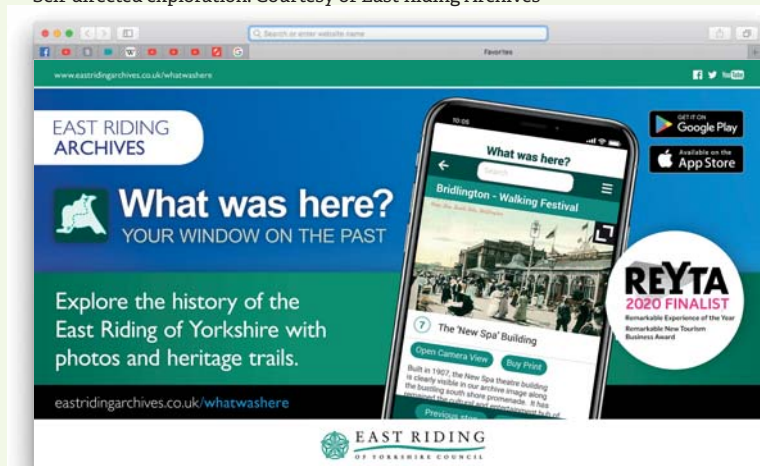
The further addition of a 'Camera View' on heritage trails enhanced the past/present comparison by using the device camera to overlay historic images onto the modern scene and allowing users to toggle the transparency.

After a year of development we finally launched (minus the maps) in April 2019 on Android and iOS platforms, together with an accompanying desktop experience.

Meanwhile, background work continued to bring the archive maps onto the user interface. We reduced the file sizes and launched the maps in April 2020, in the middle of lockdown, with a timeline including an 1855 composite of Ordnance Survey sections, region-wide maps stretching back to 1610, and 18th and 19th century enclosure maps. The app included an opacity slider tool for making the past/present comparison.



Self-directed exploration. Courtesy of East Riding Archives



The 'What was here?' mobile app. Courtesy of East Riding Archives





Enclosure map overlaid onto modern satellite imagery. Courtesy of East Riding Archives



Enclosure map, mid-transition to Google Maps (with opacity slider). Courtesy of East Riding Archives

It's satisfying to think many 'What was here?' users will be discovering items like enclosure maps for the first time. Archives don't have to, nor should they, be the preserve of the discerning researcher, and they can have direct mainstream uses for anyone with an interest in the past. The 'What was here?' app demonstrated this by trending in the Google Play store and earning 'Finalist' nominations among top local businesses at the 'Remarkable East Yorkshire Tourism Awards (REYTA) 2020', namely the 'New Tourism Experience of the Year' and the 'New Tourism Business of the Year' categories. The merits of the app's tourism value were assessed alongside popular established attractions including a multi-million pound concert venue, gin school, local breweries, and a national trail. It was fantastic to see the app hold its own, and I think this represents the power of archives; that a low profile archives service can compete in the razzmatazz world of the tourism industry with a single digital

engagement device. There is enormous untapped potential in archival information; all it needs is a way to package and deliver it digitally.

At East Riding Archives we now have a free digital platform to distribute archival content instantly onto mobile devices, making us more discoverable and increasing our digital presence.

New partnerships have been established with local stakeholders in the tourism and heritage sectors, and as we all navigate through a COVID-19 world, where social distancing is vital, the delivery of archives onto mobile platforms and other digital channels as a means of access is ever more desirable.

The app is free on Google Play and the App Store. For desktop users, go to [www.whatwasthere.org](http://www.whatwasthere.org).





Coventry University's 1932 Lanchester car parked outside the library (appropriately named the Frederick Lanchester building). Courtesy of Lanchester Interactive Archive/ Coventry University Archives, Creative Commons (BY-NC-ND) 4.0 license

# Who was Frederick Lanchester?: a car manufacturer, engineer, scientist or inventor?

**Gary Collins** of Coventry University explains how Frederick Lanchester was all these and much more...

Frederick Lanchester (1868-1946) created the first all-British four-wheel petrol driven motor car in 1895, published pioneering work in aerodynamics, and designed powered flight in the 1890s before the Wright brothers' first successful flight in 1903. His mathematical theories on military combat and strategy formed the basis for business operations models, and he was a government advisor on military matters in the First and Second World Wars. He was a true polymath and among his other interests were musical notation, pneumatic-framed buildings, radios, loudspeakers, and gramophones.

The Lanchester Motor Company was formed in 1899 in Birmingham and Fred spent much of his life and career in the West Midlands, a region that had particular significance for the development of the automotive industry.

So how do you bring to life the work of an extraordinary man often referred to as a modern-day Leonardo da Vinci? This is the task facing the Lanchester Interactive Archive (LIA) at Coventry University, which celebrates the hugely varied work of Lanchester, his life and his inventive mind.





*A free, engaging and easy-to-use augmented reality app brings Lanchester's inventions to life using target images in the exhibition space*



The LIA uses augmented and virtual reality technology to tell his remarkable story and remind the world of his innovations, so often ahead of their time. We like to think that Fred would be fascinated by such technology, especially as his interests also included optics, field of vision and colour photography.

In early 2016 the LIA started to digitise much of the material in the collection, the largest Frederick Lanchester archive in the world. The LIA public exhibition space in the university library opened in April 2017 and includes a car that visitors can sit in, which was built by Lanchester expert and historian Chris Clark. The project's first phase (funded by the National Lottery Heritage Fund and others) ended in April 2019, and the university now continues to support the LIA.

What sets the LIA apart from many archives is their use of technology to reach a wider audience through its outreach programme. A free, engaging and easy-to-use augmented reality app brings Lanchester's inventions to life using target images in the exhibition space, which can also be taken out to events. They can be printed at home or used from a screen, making the experience very portable. A further games app combines a fun approach to solving



An augmented reality tablet in a steering wheel frame being used in the Lanchester Interactive exhibition space at Coventry University Library. Courtesy of Lanchester Interactive Archive/Coventry University Archives, Creative Commons (BY-NC-ND) 4.0 license



Bringing the Lanchester story to life in virtual reality is an immersive experience and enables users to visit Lanchester's workshop. Courtesy of Lanchester Interactive Archive/Coventry University Archives, Creative Commons (BY-NC-ND) 4.0 license

some serious engineering puzzles, while players learn more about the principles behind the designs. For an even more immersive experience, both apps have been combined into virtual reality for the Oculus Go VR headset. In this experience you are transported to Lanchester's workshop and you can explore the sights and sounds

of his machines surrounding you.

Because of its experience before the COVID-19 pandemic lockdown, the LIA was well placed to move more of its outreach programme online. Examples include video features produced using contemporary reports about events at the famous Brooklands





Courtesy of Ugwueze, O 2017, 'Flight Simulation Model of the Lanchester 1897 Manned Aerial Machine', MSc Thesis, Coventry University, Coventry

**“A free, engaging and easy-to-use augmented reality app brings Lanchester’s inventions to life using target images in the exhibition space”**

racetrack in the 1920s, and monthly sketchbook club challenges using themes from Lanchester’s life and work. Information from the collection and elsewhere is used to highlight significant events and facts for ‘On this Day’ items, and for social media, such as providing material for Twitter themes such as #Archive30, #ArchiveZ, #HistoryBeginsAtHome, #ExploreYourArchive and others.

Plans are underway to combine these new technologies with the university’s recent purchase, a 1932 Lanchester 15/18 car. The car will be equipped with projectors, tablets, VR headsets, copies of documents and other contemporary objects, to provide a unique experience of what the archive has to offer, especially during 2021 Coventry City of Culture.

Over 21,000 images will be available via the university’s online archives catalogue allowing access to correspondence, sketch books, note books, patent applications, blueprints, published works and a large collection of contemporary photographs of Lanchester cars and other vehicles. Unlisted material will be added to the catalogue in the future. This

includes items relating to other Lanchester family members and their achievements, which have been donated by individuals and organizations such as the Lanchester Trust, a charity that supports the University’s Lanchester collection work.

The LIA aims to expand the potential for research in various subjects. One Coventry student was inspired by an 1897 Lanchester aircraft patent and used it as part of his MSc in aerospace engineering. A flight simulation model of the manned flying machine (never built at the time) was created and software then proved that Lanchester’s machine would not just have flown, it would have been more aerodynamically stable than the Wright brothers’ machine! The simulation was among several designs showcased by Coventry University students at the 2018 Farnborough International Airshow.

More information about the collection is available on the LIA website [www.lanchesterinteractive.org](http://www.lanchesterinteractive.org), and the online catalogue is accessible at [catalogue.lanchesterinteractive.org](http://catalogue.lanchesterinteractive.org). With thanks to the LIA team: Anthony Hughes, Paul Henderson, and Paul Nolan.

EXPLORED  
DISCOVERED  
FOUND  
DETECTED  
CONNECTED  
ENGAGED  
UNEARTHED  
LEARNED



CONNECTED  
ENGAGED  
UNEARTHED

[www.archivescard.com](http://www.archivescard.com)