“A Tartan Rather Than a Plain Cloth”

Building a Shared Workflow to Preserve the Regional Ethnology of Scotland Project Archive

|  |
| --- |
| **Sara Day Thomson** |
| *University of Edinburgh*  *UK*  *Sara.Thomson@ed.ac.uk*  *0000-0002-3896-3414* |
|  |

**Abstract – This paper provides a case study of a shared workflow to preserve the oral history recordings created through the Regional Ethnology of Scotland Project. This workflow has been developed as the first production run of a semi-automated integration of Archivematica at the Centre for Research Collections (CRC) at the University of Edinburgh. This experience demonstrates that digital preservation, above all, is about people and to succeed requires the input of a range of perspectives and skillsets.**

**Keywords – oral history, workflow documentation, Archivematica, automation**

**Conference Topics – Community; Resilience**

# Background

In an interview from 20th August 2012 [1], Robert McQuistan from Carsluith tells his interviewer, Mark Mulhern, a lead researcher with the Regional Ethnology of Scotland Project (RESP) [2], what he has learned as a fieldworker collecting the stories of his neighbors:

“'Oh, well, ah’ve learnt how intensely people feel about their own history, they, once they get into it and, one person says ‘Oh ye know, it’s taken me back and it’s made me rethink and relive my past and stuff that ah’d forgotten about has come back.’

“So, in that personal sense, it’s quite powerful for them but from my point of view just the flow of a person’s life, just how it developed and evolved over the years. And the changes, just the remarkable changes from five to ten years, to twenty years, just how it all piles up.

“All of that,” he says, 'it’s like a rich tapestry, it’s like a tartan rather than a plain cloth.”

The way McQuistan describes the accounts of individual lives in his village, the “remarkable changes”, also reflects the challenges faced by digital preservation. While the relentless evolution of technology poses a risk to continued access, the inevitable transformation of staff, organizations, and end users also poses a risk of digital resources dissipating in the mists of time. The fieldworkers, volunteers, researchers, ethnographers, curators, and archivists involved in projects like the RESP Archive move on to other things. The institution looking after the resource - the files, documentation and all their idiosyncrasies - undergoes restructures, staff come and go, and priorities evolve. Digital preservation aims to anticipate these nebulous and unpredictable changes so that unique resources such as the RESP Archive have the best possible chance of persisting into the future.

This paper provides a case study of a shared digital preservation workflow built to withstand those changes. It reflects the steps taken to preserve the oral history recordings created through RESP, a project managed by the European Ethnological Research Centre [3]. This workflow has been developed as the first production run of the bespoke implementation of Archivematica [4] at the Centre for Research Collections (CRC) at the University of Edinburgh [5], the custodians of the RESP Archive. This experience demonstrates that digital preservation, above all, is about people and to succeed requires the input of a range of perspectives and skillsets. A digital archivist can help bridge the gap between a systems developer and project archivist. The nuanced understanding researchers and project archivists possess about the digital resources has a direct impact on the effectiveness and robustness of a digital preservation strategy. Long term access to these life stories from across Scotland is best assured by different types of practitioners working together, implementing digital preservation measures as early as possible.

# Automating Digital Preservation with Archivematica at the Centre for Research Collections

The digital preservation system at the CRC - encompassing archives, rare books, arts collections, museums, and reader services - uses Archivematica to process digital content and create AIPs and (in some cases) DIPs. This system, built by a dedicated developer (Hrafn Malmquist) in 2018-19, automates the transfer of AIPs and DIPs to DSpace [6] and ArchivesSpace [7] using the integration *DSpace via REST API* [8]. The system also integrates Archivematica with *Tivoli Storage Manager* (TSM), a proprietary tape storage system by IBM [9], used as primary preservation storage.

These Archivematica integrations fulfill the need to automate the transfer of digital archival materials directly to preservation storage and to the archives discovery system ArchivesSpace. The AIPs created by Archivematica are pushed to DSpace Collections and to TSM so that the AIP is duplicated. DSpace, running on disk storage, provides easy access to AIPs for investigating issues and evaluating processes. The DIPs created by Archivematica are pushed to DSpace Collections and then to the catalog record where they appear as Digital Objects. In sum, 1 Archivematica SIP = 1 Archivematica AIP/DIP = 1 DSpace Item = 1 ArchivesSpace Digital Object Record. The objective of this implementation was to create a seamless workflow from ingest to preservation to access.

These Archivematica integrations were built based around University Court Senate Records. These archival records mainly comprise meeting minutes in MS Word and Adobe PDF formats. This corpus of data posed relatively few technical complications for the envisioned workflow, aside from establishing different levels of access restrictions (which was ultimately handled outside the preservation workflow). The processing of the RESP Archive provided the first corpus of data to test the implementation with content that had not informed development.

# A Programmatic Workflow for the RESP Archive

The RESP Archive includes oral history recordings (audio and moving image files), transcriptions (PDF), and photographs (JPG) created through RESP, a project that enables communities across Scotland to work together to record information about their local life and society. This work is carried out on a regional basis by conducting fieldwork interviews. To maximize the usability of the collections for researchers and others, detailed summaries for each item are provided in the ArchivesSpace catalog and all interviews have been transcribed in full. RESP considers the collection to be the creation of those who have made the recordings. As such, it is a central aim of the project that the recordings are made freely available in an easily accessible way, presented under Creative Commons. The RESP Archive, as a result, contains a collection of fieldwork interviews rich in detail about all aspects of life, place, and memory from different regions of Scotland.

While the RESP interviews follow a relatively uniform model, the process also involves a good deal of organic growth and deviation, leading to a relatively complex archival structure. The RESP Archive is arranged by Region, then by Fieldworker, each Fieldworker comprising a series of interviews, some including only a single interview and others comprising closer to ten. Some Fieldworkers, like McQuistan, have also been interviewed to gather reflections on the methodology and experience. Some Fieldworkers are individuals and some are entire groups, like the Campie Primary School P5 pupils [10]. The hierarchy (and deviations across the collections) created a challenge for developing a workflow for a semi-automated Archivematica implementation developed for a relatively flat record series (the University Court meeting minutes). While automation has the power to exponentially speed up processing, it also requires materials and metadata to be structured in a particular way.

The structure of the RESP Archive wasn’t the only challenge for the workflow. The ArchivesSpace interface did not support the accessibility and usability required by the target end users of the collection, who are not expected to be familiar with archival research. In ArchivesSpace - built as an archives catalog not a digital repository system - discovering content and browsing the fully digital collection was cumbersome and opaque compared with the web-based discovery most people expect. The RESP Archive team, with the developer who built the Archivematica implementation, opted to build a website that would automatically pull metadata from ArchivesSpace and corresponding files from DSpace.

More fundamentally, the workflow suffered from a heavy dependency on developer support, as the workflow was initially developed with no digital archivist in post. The original developer, who had detailed knowledge of the systems involved, had moved to a completely different project. The project archivists had no training in Archivematica and digital preservation processing was outside their remit. These circumstances led to long delays in establishing the workflow, deciding requirements for preservation metadata, and processing content.

# Adapting a Programmatic Workflow to a Manual Workflow

To reduce dependency on developer support, the workflow needed to be re-designed so that it could be implemented by the digital archivist with support from the RESP Archive team. First, the programmatic workflow had to be documented (to a basic level) and broken down. The main tasks in the workflow which had been carried out programmatically included:

* Creation of a metadata file encoded in json to instruct Archivematica where to send the AIP and DIP
* Creation of directories in the required structure
* Transfer of files from network storage (Data Store) to the staging area for ingest into Archivematica
* Execution of the Archivematica process (referred to as a ‘transfer’)

Some parts of the workflow were already manual (or only partly automated):

* Creation of Collections in the dedicated DSpace repository
* Suppression of unredacted files in DSpace
* Deletion of unredacted files in ArchivesSpace
* Quality assurance of the website

Performing any of these previously automated tasks manually is more labor-intensive, but not prohibitive. The more serious problem is that manual processing creates a greater risk of human error. However, because the RESP Archive team no longer had to wait for availability from a developer, the content could be transferred in smaller batches as it became available. Furthermore, members of the team are able to check each other's' work throughout the workflow.

Due to the delays created by the developer dependency in the initial workflow, the RESP Archive team was keen to take on more parts of the digital preservation workflow. During the process of transforming the workflow, it was well-documented. As a result, it was relatively straightforward to identify those parts of the digital preservation workflow that could be handed over to project archivists upstream. The main barrier to handing these tasks over was knowledge of the systems and technologies involved. The digital archivist provided some basic training and step-by-step documentation for DSpace, alongside a general overview of Archivematica. While the metadata file encoded in json could be produced by hand in Notepad or a similar program, the team decided this approach entailed too high a risk of human error.

Narrowing down the remaining barriers to the workflow enabled the team to provide tightly scoped requirements to the development team. Fortunately, the head of the development team was able to include the creation of a lightweight, web-based tool as part of a larger internally funded project. The result is JSON Convert, a simple form with metadata fields that transforms input text into the json structure required for the automation of the Archivematica workflow. Using JSON Convert, the RESP Archive team has taken over a time-consuming task, creating the metadata file for each SIP alongside the creation of the ArchivesSpace catalog. This gives the project archivists, who work more closely with the fieldworkers and researchers, control over the basic metadata included with the AIPs.

Distributing parts of the digital preservation workflow has raised awareness of digital preservation and provided an opportunity for project archivists to upskill in digital methods and systems. Furthermore, the previously specialized, programmatic tasks in the workflow have now been transformed to manual tasks that a wider set of practitioners have the skills to perform. As the workflow has been devised and tested and revised, the RESP Archive team has documented the digital preservation tasks so they can be implemented by practitioners with little or no experience of Archivematica or encoding information in JSON.

# Lessons Learned and Next Steps

While it may seem like a backwards step to transform a programmatic workflow into a manual one, the process has allowed the practitioners involved to evaluate the process in detail. Before expounding on future plans, it must first be acknowledged that the resource to process and catalog this collection so quickly, to such a high quality, is a real privilege made possible by external project funding. The majority of CRC collections do not enjoy this level of staff resource. As a result, the model created by the RESP Archive will not be easily applied to other archive collections (though it does provide valuable precedent).

The new, frontend workflow for the digital preservation of the RESP Archive has become well-established to work effectively. However, there are a number of improvements that have been identified. A new platform or digital repository purpose-built for discovering and engaging with digital archives would circumvent the need for creating project websites to improve accessibility and usability, removing an integrated system from the workflow. Furthermore, the CRC remit includes large digitized collections, digital artworks, and datasets. ArchivesSpace will not support access for these non-archival formats.

Before leaving, the implementation developer submitted a request to Artefactual (the developers of Archivematica) for functionality to automatically suppress unredacted files (i.e. allow passing an authorization policy when depositing to DSpace). This automation would prevent the need to manually suppress or remove these files after the completion of an automated workflow [11]. The digital archivist aims to re-introduce automation more generally, for example to create structured directories. Though the manual workflow has helped to refine requirements for future automation.

Most significantly, the aim is to further develop JSON Convert so that it can be used for the submission of metadata and files directly to digital preservation. With time, this transfer method could be rolled out not only across the CRC, but across the library and the wider university. Future development might be modeled on the AVP Exactly tool [12] which currently uses FTP transfer or the web form created for the CRC’s Collecting Covid-19 Initiative which transfers files over HTTPS [13].

This workflow, developed for a remarkable collection of recorded memories and life stories, has provided a basis for the expansion of digital preservation at the CRC. The progress made on the RESP Archive shows great promise for what's to come, in the words of Robert McQuistan, “the changes, just the remarkable changes from five to ten years, to twenty years, just how it all piles up”.

# REFERENCES

1. Interviews of Robert McQuistan, EERC/DG/DG14/3. <https://collections.ed.ac.uk/eerc/record/165128/archival_object>.
2. RESP Project Website. <https://collections.ed.ac.uk/eerc/>.
3. EERC at the University of Edinburgh. <https://www.ed.ac.uk/literatures-languages-cultures/celtic-scottish-studies/research/eerc>.
4. H. Malmquist, ‘Automating OAIS compliant digital preservation using Archivematica and DSpace’ (26 November 2019). DOI: <https://doi.org/10.5281/zenodo.3554060>.
5. Centre for Research Collections, University of Edinburgh. <https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc>.
6. DSpace Wiki. <https://wiki.lyrasis.org/display/DSPACE/Home>.
7. ArchivesSpace. <https://archivesspace.org/>.
8. Artefactual, Archivematica Storage Service 0.13.0, Documentation, Administering the Storage Service. <https://www.archivematica.org/en/docs/storage-service-0.13/administrators/#dspace-via-sword2-api-or-dspace-via-rest-api>.
9. IBM, IBM Tivoli Storage Manager Documentation. <https://www.ibm.com/docs/ko/tsm>.
10. Interviews of Campie Primary School P5 pupils, EERC/EL/EL21. <https://collections.ed.ac.uk/eerc/record/190167/archival_object>.
11. Github, archivematica/issues, ‘Feature: Allow passing an authorisation policy when depositing to DSpace #1316’. <https://github.com/archivematica/Issues/issues/1316>.
12. AVP, Exactly. <https://www.weareavp.com/products/exactly/>.
13. Centre for Research Collections, University of Edinburgh, ‘Collecting Covid-19 Initiative’. https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc/collecting-covid-19-initiative (live URL) or https://web.archive.org/web/20220305173433/https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc/collecting-covid-19-initiative (archived URL, 2022-03-05).