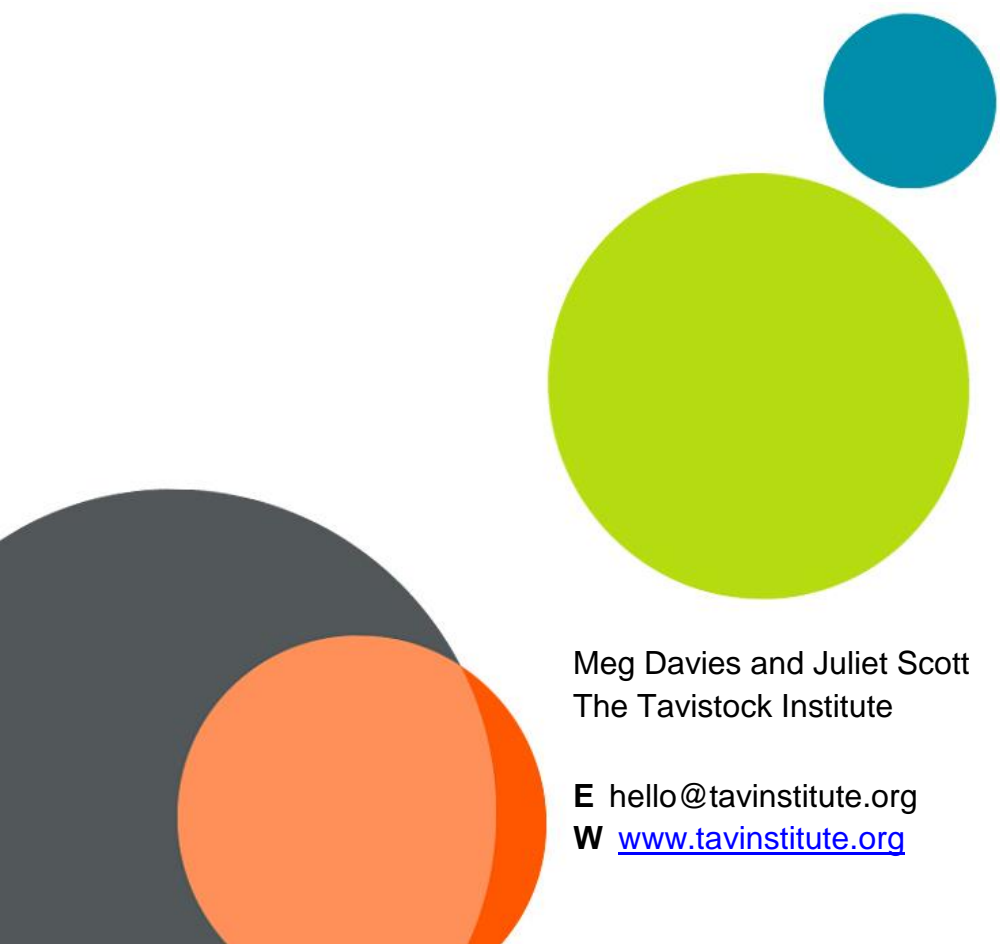


Digital Archive Scoping Report

Report written by Frank Owen with contributions by
Meg Davies and Juliet Scott

Max Communications Ltd and Tavistock Institute of
Human Relations (TIHR)



Meg Davies and Juliet Scott
The Tavistock Institute

E hello@tavinstitute.org
W www.tavinstitute.org

Frank Owen
Max Communications Ltd

E frank@maxcommunications.co.uk
W www.maxcommunications.co.uk

Contents

1	Introduction	2
1.1	Background	2
2	Project Methodology	4
2.1	Alignment	4
2.2	Definition	5
2.3	Design	6
2.4	Implementation	10
2.5	On Going	11
3	Findings and Recommendations for the TIHR Project	13
4	General recommendations for arranging files for Digital Preservation	14
	Appendix 1 Wellcome/TIHR Archive Review Criteria	15
	Appendix 2 Wellcome Accession Inclusion Criteria	16
	Appendix 3 Defined output structure for TIHR born digital file arranging	17

Table of Figures

Figure 1: View of the main screen of the prototype tool. Users can filter by search text, document type and section.	7
Figure 2: Detail of batch value changing tool. Users can select multiple documents with the check box and set batches file type and archival action to be carried out.	7
Figure 3: Desired directory paths represented in a spreadsheet table	9

1 Introduction

This report is for the Scoping Grant awarded to The Tavistock Institute of Human Relations (TIHR) and relates to their digital files up to 2009.¹ This phase was preceded by appraisal deposit and cataloguing of the paper archive, 1930s to 1990s that was carried out from 2012 to 2017.

Scoping born digital files presents different challenges than a paper-based collection. Digital collections tend to be larger in scale as there is less imperative to selectively save material due to limitations in physical storage space. There are also issues about duplication of multiple saved copies and obsolescent file formats.

Born digital collections also have the advantage of allowing automated auditing and manifest through some freely available software tools.

The work undertaken in this scoping project ran parallel with producing a practical solution for tackling the daunting task of arranging tens of thousands of files for ingestion into a digital preservation and archive management system.

This report aims to document not only an overview of the scope of the collection but also the processes, methodology and strategies TIHR are employing to select and arrange material for archiving, using multiple indexers.

1.1 Background

The Tavistock Institute for Human Relations faces a unique challenge. The Institute produces a great deal of research of academic interest, but is not part of an academic institution with the conduits and infrastructure for disseminating this material to the academic community.

To address this challenge the TIHR has entered a partnership with the Library at The Wellcome Collection for the management and dissemination of archive material.

The first stage of the project saw:

- The movement of paper based records to Wellcome
- The paper records being catalogued in Calm by a TIHR archivist seconded to Wellcome (2015-2017)
- Production of a preliminary retention policy
- Archive descriptions being disseminated at <http://archives.wellcomelibrary.org/>.
- Engagement and dissemination of the archive via the archive blog at <http://tihr-archive.tavistock.org>, art exhibitions, social dreaming events and TIHR's 70th anniversary festival in 2017.

An indication of the interest and significance of TIHR's outputs is the great success of the hard copy material already on Wellcome's online catalogue. TIHR articles are among the most frequently accessed in the Rare Materials Room.

¹ This is the cutoff date agreed by TIHR Management Team in consultation with the Wellcome for the first deposit of Born Digital material. Accruals will be every 5 years hereon.

Wellcome have the IP for the catalogue descriptions they have produced. Not all the archive has been catalogued. None of the content has, of yet, been digitised and OCR'ed.

Wellcome are set to take on the born digital material up to 2009 but the material must be "weeded" (non archivable files removed) and arranged (put in an ordered directory system) before accession. Wellcome has produced a set of criteria that files must meet to be accepted; see Appendix 2 - Wellcome Accession Inclusion Criteria.

There is a desire within the TIHR Management Team, to make the archive the central means of disseminating the work of The Institute. Stakeholders should see archiving as a project's final goal and a way to build their own personal professional authority and legacy.

In the second stage of the project TIHR face the challenges of archiving "born-digital" output created during the desktop publishing (DTP) revolution - the period from the late 1980s through to 2009. The issues TIHR faces are common to many organisations seeking to archive digital material of this period.

The opportunities that word processing and DTP offered to publish and disseminate individual and organisation's work, naturally lead to the creation of a large number of documents. Organisations rarely implemented sufficient processes for storing the documents systematically, leading to generally disorganised collections.

In some cases this lead to a wholesale "saving-of-everything", with individuals and teams often using idiosyncratic naming conventions. The result for the TIHR is a server with tens of thousands of differently named documents stored in a variety of directory structures. In his essay "Parsimonious Preservation" Tim Gollins outlines how the daunting nature of organising chaotic drives has become a barrier for digital preservation.²

The work needed to meet Wellcome's Accession Inclusion Criteria is fundamentally the same work needed to prepare documents for ingestion into digital preservation and archive management systems. Some of the methodologies and tools used for this project should be directly transferable to other projects.

² <http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf>

2 Project Methodology

The project was planned around the five phases outlined by The British Computer Society “IT Enabled Business Change” methodology; Alignment, Definition, Design, Implementation and Monitoring.

2.1 Alignment

This is the process of understanding the mission and goals of the organisation as a whole and understanding what the benefits are the project can provide the organisation. It involved auditing the existing material and a discovery process with stakeholders of content that might be of benefit to the project. A confirmation of the business case to guide the project and a set of expected outcomes came out of this process of stakeholder consultation.

Business Case for transfer of material to Wellcome

- Long term preservation of the Institute’s archive for future study.
- Dissemination of the Institute’s work to a wider research community via Wellcome.
- Operational cost benefit of managing server space at TIHR.
- TIHR’s IT provider can focus on day-to-day management of the IT systems rather than the long term preservation of records.
- A new archive directory structure benefits TIHR by
 - Making the archive the destination for work within the organisation.
 - Decreasing the work for future accruals of material for Wellcome.

Project Outcomes

- Tools and a methodology for multiple end users to sort files ready for ingestion into Wellcome’s digital preservation and dissemination systems.
- Tools and methodology for other organisations to carry out arranging of files before digital preservation or ingestion into archive management systems.
- Greater insight into the content of the archives to facilitate curatorial and editorial work for the TIHR archive blog; a key media for further publicising the popular Wellcome Online Catalogue of TIHR and increased research interest into the archive.
- Definition of a clear archiving and retention policy for born digital material.
- Design of an archiving pathway and file structure for ongoing operational use.

Securing the Data

The first step in the project was to take a copy of the original data as a snapshot for possible rollback. This snapshot covered the Institute’s full shared server with just under 1TB of data. TIHR’s IT provider copied the material on to an encrypted external hard drive which was then written to 2 LTO tapes to provide a solid backup.

Issues and lessons learnt

The file copy process carried out by the IT provider failed to copy all the files correctly. The tool TeraCopy provides a better confirmation of file transfers.

Audit of files

From the snapshot of the files we used a linux based server side application called exifTool to produce a manifest of the files. ExifTool provides a CSV (comma separated variable file, in effect a stripped down spreadsheet) with rows for each file containing a full pathname and extracts all the available metadata attached to the file.

There are hundreds of different metadata fields, most specific to certain file types. The metadata in a file is information about the contents of the file. The data columns chosen were deemed to be of use to the indexers:

- FilePath
- Size
- Author Title
- SubTitle
- CreateDate

The extent and content of the metadata for each file is dependent on how the original author set up their software. Various versions of Microsoft Office produce varying amounts and types of metadata.

Issues and lessons learnt

ExifTool does not extract data from txt files and omits any files without metadata.

A second manifest was produced using a linux “find” script that included all the file types missed by exifTool. This list was then amalgamated with the previously extracted metadata material.

2.2 Definition

From the Alignment process, requirements and Critical Success Factors for a solution became apparent. These were used to ensure that any developed tool was fit for purpose.

CSFs and Requirements for Tools and Methodology

The developed methodology had to allow -

- Non-destructive arranging of files by multiple end users. In TIHR case these are summer student interns and/or apprentices.
- Tagging of files for actions i.e. retain, archive, sensitive GDPR, review, destroy.
- TIHR moderator approval of students’ arranging work before actual file removal / archiving.
- Viewing of documents for review.
- Working strategy to guide users on which files to archive.
- Batch tagging of multiple files, directories and sub-directories where appropriate.

Defined Final Output Directories

A simple strategy to help staff make quick and accurate decisions on the suitability of thousands of documents was needed. A clearly defined target directory structure informs the decisions of the end users for retention and arranging. By defining an ideal directory structure for the material, staff are able to look at each document and see if it fits in any of the existing directories. If it does they

can label it “archive” and register it for the destination directory. If it does not then they can mark it for removal, or if it clearly needs archiving add a new folder to the final directory structure.

The TIHR archive team took a day out to define a directory structure that would cover all the documents that met Wellcome’s Archive Review Criteria. Juliet Scott, Megan Davies and Dr Elizabeth Cory-Pearce with Frank Owen went through all the material at a macro level and defined a directory structure to encompass all the files the organisation will archive.

The outputs of TIHR work were identified as being in three distinct areas. Professional Development, Governance and Central Administration and Research Projects. It was recognised that users should be able to add folders in some instances where an undefined number of directories might exist. E.g. Other professional development conferences and research projects.

The directories and subdirectories for the three sections are described in Appendix 3. Certain exclusions around bursaries, finance, and marketing were identified in the process and included for the indexing team’s reference.

An additional tag was made available to the indexers for GDPR sensitive material. It was felt this material would be of long term interest but that it should not be made available currently, as no consent had been given by the individuals concerned.

2.3 Design

Online Tool

An online tool has been used in the past at Max Communications for multi user indexing jobs. Some preliminary research work was carried out to test the suitability of developing an online tool for this project.

Advantages of an online tool are

- multiple users working remotely
- the file structures are not compromised as the files are sorted by proxy
- Having the metadata for the files in a relational database facilitates export for ingest into Wellcome’s Digital Preservation / Archive Management system.

The CSV manifest was imported into a database, and a front end web page produced.

Figure 1: View of the main screen of the prototype tool. Users can filter by search text, document type and section.

TIHR Document Classification Pre Archiving

Filter Documents

Any Section

All Action Types

All Doc Types

Filter

Search

Change the selected values

Do Not Change

Do Not Change

Change all selected values

Logout

#SourceFile	#Title	#Author	#CreateDate	#section	#docType	#action	Select All
CONFERENCES/LEICESTER 2006/LEICESTER 2006/Leicester Things to do.doc		msher	2005:03:12 15:23:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Programmes & Timetables/ConferenceProgramme(Members).doc	WORKING CONFERENCE PROGRAMME: 30 March – 12 April 2003	SHER	2005:06:26 11:39:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Programmes & Timetables/Daily staff timetables.doc	Thursday 6 1h April 2000	SHER	2005:06:26 12:38:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/LEICESTER 2006/STAFF/CONFERENCE MANAGEMENT AND STAFF.doc	CONFERENCE MANAGEMENT AND STAFF	Mannie Sher	2006:03:12 16:48:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Leicester Things to do.doc		msher	2006:03:12 15:23:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/DIRECTOR'S MATERIAL/Director's CONFERENCE opening address.doc	The Leicester Conference 2000	SHER	2006:03:12 17:22:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/DIRECTOR'S MATERIAL/IE Opening Statement.doc	Authority Leadership & Organisation The 'Leicester' Conference 5th _ 18th April 2000 University of Leicester	SHER	2006:03:12 20:08:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/DIRECTOR'S MATERIAL/JG Opening Address.doc	Authority Leadership & Organisation	SHER	2006:03:12 19:38:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/DIRECTOR'S MATERIAL/Seating @ Opening Plenary.doc	SEATING ARRANGEMENTS – OPENING PLENARY	SHER	2012:04:11 15:17:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/Leicester Things to do.doc		msher	2006:03:12 15:23:00	Professional	Undefined	Undefined	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/Programmes & Timetables/ConferenceProgramme(Members).doc	WORKING CONFERENCE PROGRAMME: 30 March – 12 April 2003	SHER	2005:06:26 11:39:00	Professional	Undefined	Undefined	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/Programmes & Timetables/Daily staff timetables.doc	Thursday 6 1h April 2000	SHER	2005:06:26 12:38:00	Professional	Undefined	Undefined	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Mannie's stuff/STAFF/CONFERENCE MANAGEMENT AND STAFF.doc	CONFERENCE MANAGEMENT AND STAFF	Mannie Sher	2006:03:12 16:48:00	Professional	Undefined	Undefined	<input checked="" type="checkbox"/>
/000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 2006/Marketing/Brochure letter general.doc	July 2005	LAO1	2005:08:05 14:17:00	Professional	Undefined	Undefined	<input checked="" type="checkbox"/>

Drop down menus on the top left of the screen allow filtering of the files by search, file type etc. The menus top right allow batch tagging of the files using checkbox flags.

Figure 2: Detail of batch value changing tool. Users can select multiple documents with the check box and set batches file type and archival action to be carried out.

Change the selected values

A* Clearly Archivable ▾

Conferences (Inst Archive) ▾

Change all selected values

[Logout](#)

#CreateDate	#section	#docType	#action	Select All
13:11:16 :15:10+00:0	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input checked="" type="checkbox"/>
09:12:01 14:02:00	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input checked="" type="checkbox"/>
15:09:09 :38:54+01:0	Professional	Conferences (Inst Archive)	A* Clearly Archivable	<input checked="" type="checkbox"/>
15:09:09		Conferences (Inst	A* Clearly	<input type="checkbox"/>

After discussing with other archivists involved in arranging files for digital preservation it was found the preferred solution for most was to use the Windows operating system File Explorer system for the sorting exercise.

However as we were looking for a non-destructive solution for multiple end users, moving the files directly wasn't feasible. It became clear that this was a key difference between paper archives and born digital material; as the Institute server is a live working system staff had to be able to continue working in the server without files disappearing/moving/appearing as archivists work.

Some prototyping work was undertaken to produce a graphical user interface with nested folder icons, based on the manifest path names. However other disadvantages became apparent with the online tool, beyond usability.

Disadvantages of an online tool

- Making direct links from the system to the actual files for viewing content is difficult. It requires uploading the entire drive to a web-server, which would constitute a security risk and require a large expense. The problem is linking local files to web pages.
- For other institutions to implement this solution would require a generalised version of the tool which could incur expense.

A Spreadsheet Solution

As an alternative to the online tool a second solution was developed using the two spreadsheet solutions Google Sheets and Excel.

Advantages of spreadsheets

- Multiple users working remotely can track their work in the same place if using Google Sheets online.
- The original file structures are not compromised as the files are sorted using details in the spreadsheet
- Direct links can be made to the source files using a downloaded version of a Google Sheet inside Excel.
- Technology and interface is readily available and familiar to most archivists in other organisations.
- Ability to add new directories and sections (e.g. a project) to the output folder

Disadvantages of spreadsheets

- Risk of users changing structure of sheet
- Risk of users adding data incorrectly if using free text

The goal remained to allow users to easily tag multiple files for archiving and assign those files to specific output folders. Defining the output folder for each document identified in the audit manifest was achieved with columns in the spreadsheet representing levels of directory. The path to the destination directory is represented by the shaded columns in the spreadsheet detail below.

Figure 3: Desired directory paths represented in a spreadsheet table

Current File Path	Top	Section	Sub1	Sub1 YEAR	Subsection2 ACTIVITY
./000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 1974/Joining Information.pdf	PROFESSIONAL_ DEVELOPMENT	GR_PROGRAMME	LEICESTER	1974	CONFERENCE_ PACK
./000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 1974/List of Members.pdf	PROFESSIONAL_ DEVELOPMENT	GR_PROGRAMME	LEICESTER	1974	PARTICIPANT_ LIST
./000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 1974/List of Staff.pdf	PROFESSIONAL_ DEVELOPMENT	GR_PROGRAMME	LEICESTER	1974	PARTICIPANT_ LIST

For each row the user selects the five directory levels for where a file deemed suitable for archiving should be moved to. For example on the first row above, the file identified in the manifest is currently at

./000 GR PROGRAMME/01 LEICESTER CONFERENCES/LEICESTER 1974/Joining Information.pdf.

The indexer has assigned this file to the path

PROFESSIONAL_ DEVELOPMENT/GR_PROGRAMME/LEICESTER/1974/CONFERENCE_ PACK/Joining Information.pdf

The combined manifest data from the exifTool and the Linux find command were imported into Google Sheets. Google sheets is a free to use, cloud based spreadsheet. Files can be worked on, via a browser, by multiple end users simultaneously. Some consideration was given to the risk involved in having the metadata in the Cloud. It was felt acceptable because the data is password protected and contains just the metadata file descriptions, not the actual sensitive material. The files will be removed from Google Drive at the end of the project.

A spreadsheet was produced for each of the three main subdivisions from the data identified during the definition process (as described in Appendix 3).

Using predefined data in the form of drop down lists, minimises typing errors in the data input. In spreadsheets this is achieved using a “protected range” for the input cell.

For each of the columns shaded in blue in the diagram above the relevant list of options was pasted into a separate tab on the Google Sheet we called “LISTS”.

A right click on a Google Sheet Cell, shows a contextual menu that allows the cell to have a “protected range”. This is defined by selecting the appropriate list from the “LISTS” sheet. The “protected Range” is then copied to the rest of the column. This in effect gives each cell a drop down menu of acceptable values. Standard “copy and paste” and “click and drag” functions within Google Sheets allow the user to assign paths to multiple files and whole directories quickly and intuitively.

Google sheets allow a single data source that multiple users can record their work on. However a downloaded excel sheet is required to enable direct linking from the sheet to the files. The users

therefore kept an excel sheet open for quick access to the files and a google sheet available for recording their work. The work could all be carried out on single Excel sheets by single users.

A new “links” column was produced in the Google Doc and the following formula pasted in to all the cells. (‘B2’ refers to the column with the original file path).

```
=HYPERLINK( concatenate ("file:///\\", SUBSTITUTE (SUBSTITUTE (B2, "/", "\"), ".\", "O:\") ) )
```

The Google Sheet can then be downloaded and the links will work to allow direct access to the individual files.

2.4 Implementation

Initial weeding of material by date and relevance.

As the material for archive is sitting on a live server there are a large amount of current files that need to be marked as “RETAIN”. There are also entire sections of the files on the server which do not meet the Wellcome selection criteria as shown in Appendix 1.

The TIHR Archive Team was able to identify which folders were irrelevant and mark entire directories as RETAIN, thus reducing a large number of required working days. This initial weeding was kept at a high level – only ‘obvious’ folders were removed e.g. folders containing financial records which Wellcome would not accept. The team also identified directories which were definitively post 2010 and these were appropriately marked as well.

Arranging by student volunteers

During the first phase of the archive project, as referenced in the introduction, students were used with great success to sort through the paper archive material and catalogue documents before they were either sent to Wellcome or destroyed. During the born digital phase it was felt that it would work to use students again as indexers; continuing on with the archive project’s aim of engaging researchers and working collaboratively across fields.

Recruitment focused on students from University College London – this was due to geographical reasons, UCL being in close proximity to the office, and also because UCL are one of three University’s in the UK that offer a dedicated Archiving course. The aim was always to recruit more than one student so that they were able to support each other during the process and share work load. Students were recruited based on their understanding of and enthusiasm for the project and training was provided on the job by the digital records management consultant, Frank Owen. Students are able to work flexibly and remotely after an induction week in the office; this allowed for a bigger pool of candidates applying.

The work has now started on indexing, tagging and arranging files and directories for archiving using the spreadsheet method. For remote working TIHR IT department has provided VNC remote access, allowing the indexers to access desktop machines within the office.

In addition to a “Sensitive material” column the indexers have added extra columns for project management information; “Notes/comments”, “Date of Indexing”, “Cataloguer” so that work can be tracked administratively.

Review of Non Standard Files

The manifest identified all the file types on the drive including several obsolete database types. From these files the types that may possibly contain archivable information are being examined.

Reviewing work of students before extraction

Content queries are fielded to Meg Davies on a day to day basis, and she has been pointing the indexers in the direction of the best suited stakeholders within the organisation to answer specific content queries. Organisational knowledge is key as files contents and titles are not always obvious to those who are unfamiliar with the work of the Institute. The work is also being reviewed and spot checked on a weekly basis.

The archive team will meet to go through the files marked for review and address any outstanding queries. The data sheets will then be redacted to show just the files that will be archived and the ones that will be destroyed. This can be passed to TIHR Management Team for approval.

Extracting files from server by tagged term.

The redacted datasheets showing which files to archives will be used to produce

- a manifest of files on the working server
- a manifest of the archival output directory
- the data to automatically produce “Powershell” commands to move files from the working directory to the archival directory

Inclusion of metadata for Wellcome’s digital preservation systems

Wellcome are in the process of migrating their digital preservation and digital archiving solutions. For open source solutions such as Archivematica and AtoM, metadata provided in a metadata.csv file can be included with the assets during importation. With further consultation with Wellcome some of this metadata can be provided from the exifTool produced manifests.

2.5 On Going

Organisation wide pathway for continuing digital preservation

Once the data has been extracted the defined archival directory structure should be available on the main working server. Processes need to be developed for regular archive preparation for accessions to be produced for Wellcome. Next steps will include:

- Applying the retain/destroy outcome of this archiving process to the Institute server; removing those files that were allocated destroy. It is noted that unlike physical documents the digital files could be stored on the server indefinitely but this contradicts the aimed outcome of this project, namely not holding onto every file and streamlining the saving of documents
- Ensuring that this archive project remains in line with the Institute’s GDPR policy and updating this policy if required. For example updating retention rates for documents that are not in the Wellcome criteria but are institutionally relevant e.g. financial documents
- Promoting the archive to research leaders within the Institute to ensure stakeholders remain aware of the joint Institute and Wellcome criteria for archiving – which documents should be kept, why these documents are kept, where should they be kept etc.

- Introducing a new protocol for signing off when a project is completed; this to include Project Directors producing a file of material for archive consideration at the end of every project
- Producing an 'Archive' folder on our server to hold research material that is for archive consideration and Professional Development/Governance files that are archive ready; responsibility for this to be held with the administration team.
- Introducing a naming protocol to avoid multiple saved versions of the same document with only minor alterations and to ensure that 'Final' versions are clearly marked – this to be done in consultation with our 3rd party IT provider and Institute management team

Accessions are only taken for material over 10 years old, but it is critical that the ground work should be carried out to tag material up to the present so that the Institute is not repeating the above process every 10 years.

Stakeholders should be encouraged to view the work involved in archiving as fundamental to producing their own personal legacy. This is consistent with the focus and orientation of the TIHR archive project overall as forming its own organisational development; recognising working with history as a dynamic process which needed sensitive excavation to support the Institute to move on from its illustrious past. This approach was carefully documented in the archive blog, comprising a series of reflections, accounts, speculations and analyses by project members, staff and associates on themes that varied from a nostalgic account of a mother cooking fish fingers to present experiences of sorting through the digital material. The work psychologically preparing the organisation to consider its own outputs as relevant for preservation and to take on a more considered approach to the arrangement of its digital filing.

Further sense was made of the project in a paper presented at the 2018 Annual Meeting of the Academy of Management as part of a wider symposium exploring the work of the archive as enabling customisation of contemporary practice.³

Monitoring uptake

Stakeholders should be given access to register and annotate their work with suitable cataloguing metadata e.g. title, date, author etc. This can be achieved with a shared spreadsheet as in this work, or a simple intranet page could be built for logging material. Either of these data repositories can provide simple management information on up take numbers.

³ <https://www.tavinstitute.org/news/tihr-storm-chicago/>

3 Findings and Recommendations for the TIHR Project

Significance of the Collection

The TIHR's research is significant both for its unique social science approach to working on societal challenges and because of the historically significant organisations it works with. Further because the TIHR is not affiliated with a University or academic institution this material has traditionally been difficult to access for interested researchers. The popularity of TIHR material already on Wellcome's online catalogue indicates the relevance and importance of the material. In November 2019 as part of the ongoing programme of work with the archive, TIHR and Wellcome are partnering on a symposium exploring research interest and engagement in the archive⁴ including E15 Acting School, research on personalisation in health, Tavistock methodologies as applied in Italy.

Current Status of the Digital Collection

Compared with many organisation's early born digital material the filing of the TIHR material is well organised and in the most part the material clearly labelled. Thus far the contracted indexers have been easily identifying material for archiving.

A full manifest of the files revealed a set of files in the format .db an obsolete database format. Investigation concluded that the material was of little interest so no further action for migrating the data was undertaken.

Digitisation and transcription of hardcopy documents and dissemination of born digital material

The partnership with Wellcome has made the catalogue information for the archives available to researchers. Having sustainable processes for making the full content of selected research, conferences and consultation available should be a goal of TIHR in the future.

⁴ <https://www.tav institute.org/news/archive-a-live/>

4 General recommendations for arranging files for Digital Preservation

Secure the data

Take a snapshot of the server using Teracopy to save a facsimile of the data to LTO.

Audit

Take a manifest of all the files and extract the useful metadata. Use the exifTool to achieve this. On windows machines exifTool will run under "GitBash". Use the find linux command to make a comprehensive manifest and combine this with the output from exifTool. The "dir" command can be used on Windows with Powershell.

Define files for saving

Identify all the stakeholders who will potentially be contributing material for the archive. As a working group define a set of target directories. This helps making retention decisions quickly and can be used for ongoing archiving of digital material.

Get multiple users to sort the files using a spreadsheet

The sheer number of files for sorting means that a team rather than an individual need to tackle these problems. By using a Google Sheet, team members can register files for archiving and define which folders they are to go to, without touching the locations of the originals. They can all keep their data in the same place and archivist or supervisor can review their work before committal.

Use command line tools

Make a manifest with pathnames to the original files. Define the final output folder structure for each file to be archived. Use "concatenation" in a spreadsheet to automatically produce rsync commands to move the files into their respective directories.

Catalogue while sorting

For this project Wellcome will be cataloguing the material but where this kind of resource is not available, the arranging of files for archiving in a spreadsheet presents an opportunity for indexing of further metadata, such as Author, Subject, Precis etc.

Appendix 1 Wellcome/TIHR Archive Review Criteria

These are the guidelines produced by Wellcome for transfer inclusion for the Library Collection.

Pre 2010.

TIHR Institutional Archives

- Minutes inc minutes and other notes of internal meetings, especially of units and working groups, not only CASR and SSG, but all units, as they developed
- TIHR /IOR/HRC/COOR numbered docs
- Annual publications e.g. reports and accounts
- Other publications e.g. journals and newsletter
- Leicester and other significant conferences and research meetings: conference pack (if produced), programme, list of participants, conference publication (if produced), keynote presentations, record of significant discussions

NB routine admin records do not need to be kept

TIHR Research Projects

- Proposal including research methodology
- Reports (interim and final)
- Working notes
- Fieldwork notes
- Minutes of project meetings

Appendix 2 Wellcome Accession Inclusion Criteria

This is the work that must be undertaken before files can be ingested into the Wellcome Collection Library.

- We must undertake an initial “weeding” of the material based on the accession criteria for non-digital material (.e.g Janet Fosters criteria). This includes
- Sorting by agreed cut-off date
- Removing files that are not relevant to the archive
- Removing files that may contravene GDPR legislation or are institutionally sensitive e.g. financial data
- Ensuring stakeholders have involvement and oversight of the process

Wellcome require a manifest of the files in CSV format and the files in their original directory structure in a “Cleaned” state. “Cleaning the files” includes

- Virus and File Corruption Check
- Removal of nonstandard characters in file names (e.g. dots)
- Checksums
- Removal or conversion of non-archival file formats
- Renaming of excessive long path names (260 characters)
- Removal of passwords on protected material

Appendix 3 Defined output structure for TIHR born digital file arranging

This is the directory structure defined by the Management Team. This represents the final destination for archive files. It provides a framework and guide for inclusion. If files do not fit into this structure, either they will be omitted or the structure will need to be adapted.

PROFESSIONAL DEVELOPMENT	000 GR PROGRAMME	LEICESTER	1974	conference pack
		BELGIRATE	1975	programme
		OTHERS	1976	participant list
			1977	publications outputs
	CORE PD PROGRAMMES	C4L_05_COACHING	1978	keynote presentations
		DBL	1979	record of discussions
		LYL_08	1980	
		P3C_07	1981	
		SUPERVISION_06	1982	
			1983	
			1984	
	OTHER CONFERENCES	KINGS	-etc--	
		USER ADDED	2009	

GOVERNANCE & CENTRAL ADMINISTRATION				
		1974	ANNUAL_REPORT	
		1975	BUSINESS_DEV	
		1976	COUNCIL_MINUTES	
		1977	COUNCIL_REPORTS	
		1978	COUNCIL_WORKING_PAPERS	
		1979	EDRU	
		1980	MANAGEMENT_TEAM	
		1981	OCTI	
		1982	PROFESSIONAL_DEVELOPMENT	
		1983	RESEARCH_AND_CONSULTANCY	
		1984	STAFF	

		-etc--	STRATEGY
		2009	OTHER

RESEARCH PROJECTS	1974	<i>USER ADDED PROJECT</i>	FIELDWORK
	1975		MINUTES_PROJECT_MEETINGS
	1976		PROPOSAL
	1977		REPORTS_AND_OUTPUTS
	1978		CASE_STUDIES
	1979		DATA_SETS
	1980		PRESENTATIONS
	1984		WORKING_NOTES
	-etc--		WRITE_UPS
	2009		OTHER