

# Jumping in at the Deep End

## Updating a Digital Preservation Policy as New Digital Archivists

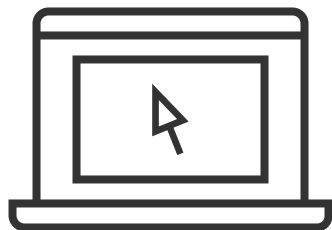


**Date:** June 2023

**Prepared by:** Jennifer Pearson & Katie Kettle  
Digital Archivists – Global Archives, HSBC

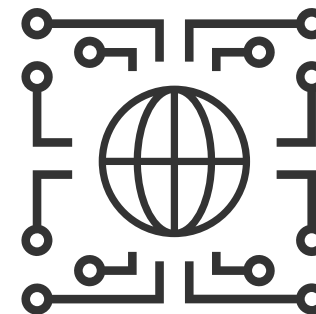


# Upgrading the HSBC Global Digital Archive



**HSBC GDA**

- ◆ HSBC's Global Digital Archive (GDA) has been in operation since 2015
- ◆ The digital collections comprise >12TB material, across our 5 regions
- ◆ The system provides continual preservation and management of digital assets, and the descriptive and technical metadata of the global collection



**HSBC GDA – Upgrade!**

- ◆ In 2022/23, the HSBC digital archives team is upgrading the GDA
- ◆ The new system will streamline and improve many of our digital archiving processes
- ◆ It will require:
  - ◆ Moving digital content from the “old” to the “new” platform
  - ◆ Implementing new archival cataloguing functionality
  - ◆ Training the wider Global Archives team so that the new system is integrated into day-to-day archival work
  - ◆ And... upgrading the Digital Preservation Policy!

# Importance of the Digital Preservation Policy (DPP)

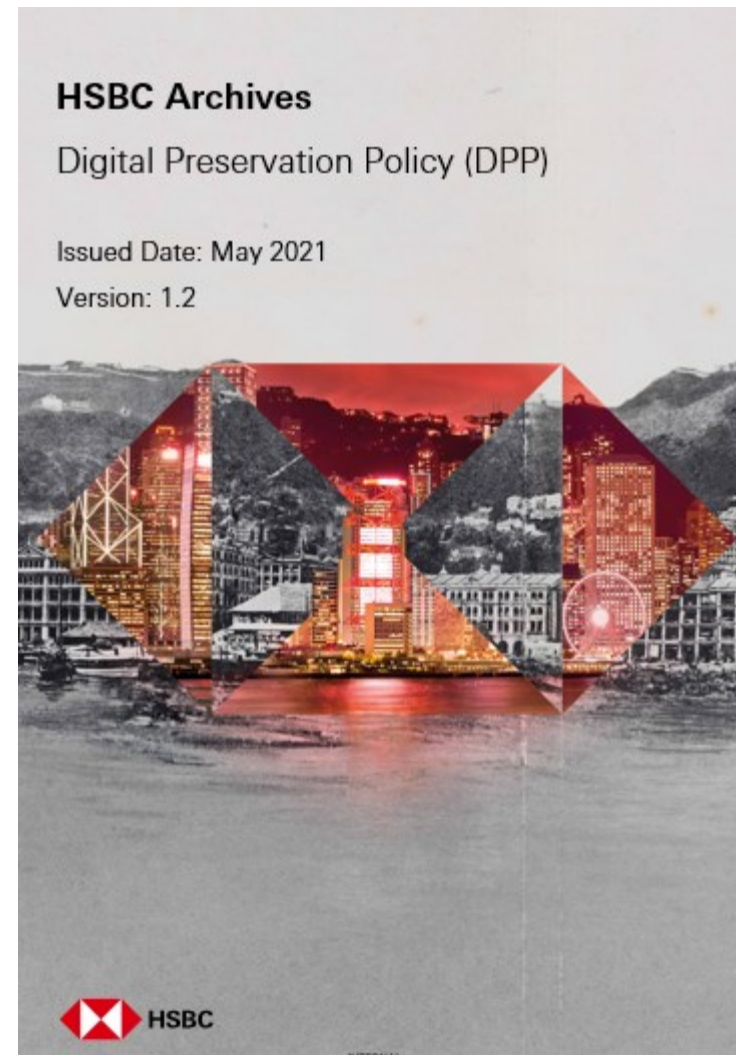
Amongst many key benefits, HSBC's DPP serves two core purposes, providing:

- ◆ A mission statement to reinforce the value of our digital archives function

**Why we do what we do**

- ◆ Detailed guidance for ensuring long-term preservation of our digital assets

**How we do it**



# Upgrading and Implementing the HSBC DPP

Our DPP upgrade plan includes:

1. Updating our policy recommendations so that required preservation and access representation formats for different file format types align with current best practices
2. Ensuring all of the required pathways to those required preservation and access representation formats are available in the new system
3. Implementing all of these preservation workflows to ensure that all existing and new content in the GDA is preserved according to the DPP



# In This Session



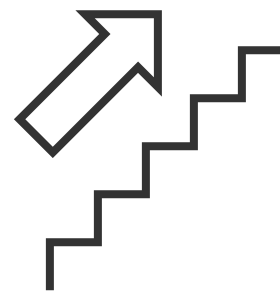
## Process

- ◆ Step-by-step approach for updating and testing the DPP
- ◆ Progress-to-date



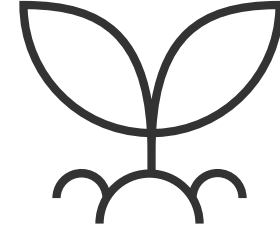
## Insights and mysteries

- ◆ Lessons learned
- ◆ Mysteries to solve



## Next steps

- ◆ Roadmap towards project completion



## Note about our backgrounds

- ◆ All of these topics, insights and learnings are being shared from our perspectives as relatively new entrants to the digital preservation profession
- ◆ Much of this presentation might be well-known to experienced colleagues, and we welcome any advice or expert insights
- ◆ We also hope that other new digital archivists will see this as a possible project approach as they consider their own similar projects



# Process



# Summary of DPP Project Steps

- 1. **Updated DPP guidance:** Specific file format recommendations were identified for each type of file in our system, specifying:
  - a) The format to which each file type should be migrated to ensure long-term preservation, and
  - b) The access representation format that should be created for each file type
- 2. **File format spreadsheet:** We downloaded this report from our system, which provided a list of all of the file formats in our system, alongside their format code, and the number of records corresponding to each
- 3. **Matched the file format report to our DPP guidance:** For each file format in our spreadsheet, we assigned the corresponding preservation and access rules from our DPP
- 4. **Checked availability of preservation and access rules in the system:** We ran a search for each file format code in our system’s set of migration rules to check that the DPP-required pathways were available

	A	B	C	D
1	File Format	Version	Format	Frequency
2	Acrobat PDF/A - Portable Document Format	1b	fmt/354	33943
3	Microsoft Word Document	97-2003	fmt/40	11898
4	Acrobat PDF 1.4 - Portable Document Format	1.4	fmt/18	11448
5	Acrobat PDF/A - Portable Document Format	1a	fmt/95	11405

fmt/45

+ Add rule for another format

Rich Text Format 1.0-1.4

Normalise document to P

Do not migrate

Normalize document to DOCX using LibreOffice

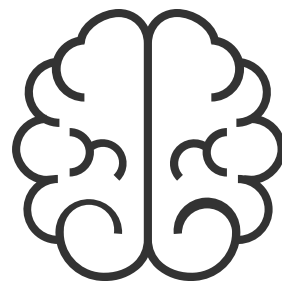
Normalise word processing document to ODT using LibreOffice

Normalise document to PDF using LibreOffice

Normalise document to PDF/A (1A) using LibreOffice

# Side Note 1: The Brain of an Archivist

Storytelling,  
People



Spreadsheets

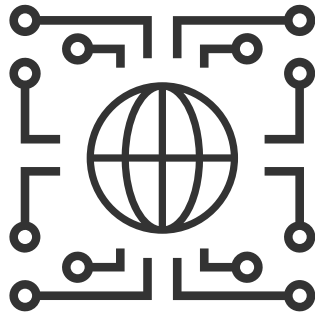


At each step in this project, our findings for each file  
format type were captured in our project  
spreadsheet



# Side Note 2: The Importance of a Test System

Our GDA has two separate versions



Live version

- ◆ Full global digital archive
- ◆ >12 Tb of catalogued digital records

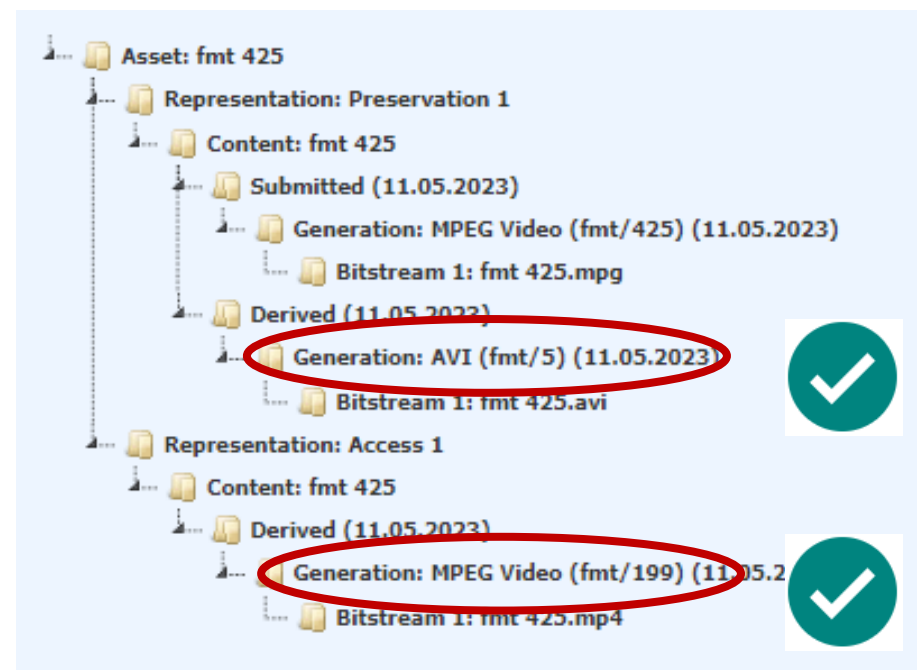
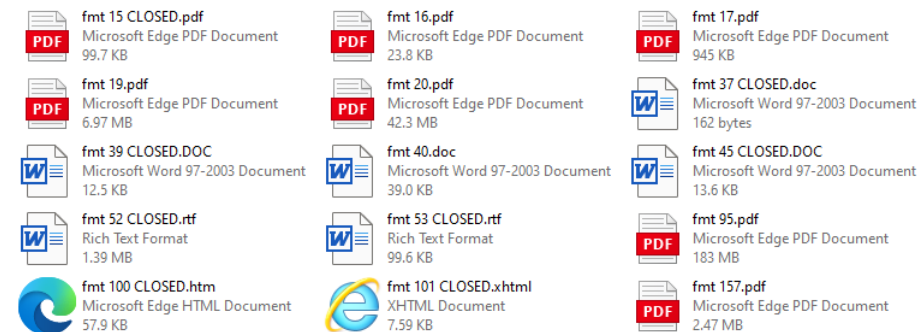


Test version

- ◆ A separate, standalone portal for testing out new features and processes, without imposing any risks to our full GDA
- ◆ **We check all of our new DPP rules on the Test system to confirm everything works as expected, BEFORE implementing anything on the Live system**

# Summary of DPP Project Steps cont'd

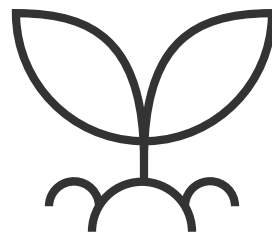
5. **Added the preservation and access rules to the Test system:** In most cases, this meant going through the file format list and bringing up each format code, one-by-one, to ensure the correct rule was selected
  - According to these rules, these actions would now be automatically applied to new files as they are uploaded (ingested) to our system
6. **Created a “Test Folder” of files to ingest:** We downloaded a copy of a file corresponding to each specific file format code from our original spreadsheet
7. **Ingested the “Test Folder” to the Test system:** With the rules in place, this should mean that all of the ingested files would have new formats created for preservation and access representations, where applicable; we checked each individual ingested file’s metadata to assess the outcome
8. **Live system implementation:** Having confirmed that all of the rules worked properly on the Test system, the next step was then to repeat these steps on the Live system



# Insights and Mysteries



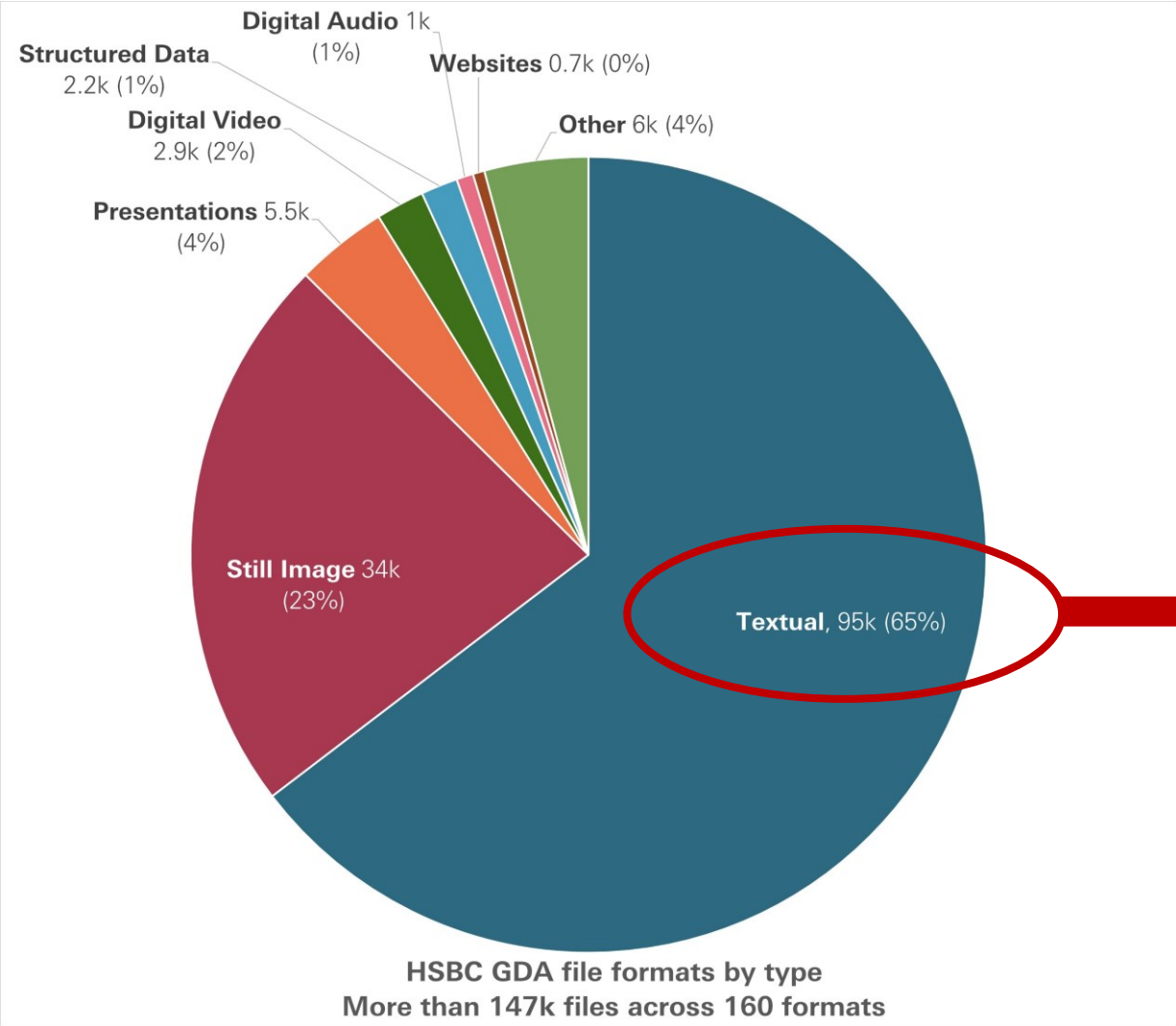
# 10 Insights and Mysteries to Share



## A reminder about our backgrounds

- ◆ As new digital archivists, each step in this project has presented fascinating new, hands-on learning opportunities
- ◆ The lessons learned and insights presented here may be obvious to many of our more-experienced colleagues, but our hope is that other new digital archivists will learn from our journey
- ◆ **If you recognise or have a solution that might address any of our mysteries, please share your insights!**

# 1. Insight: A Pivot on DPP Guidance for PDFs

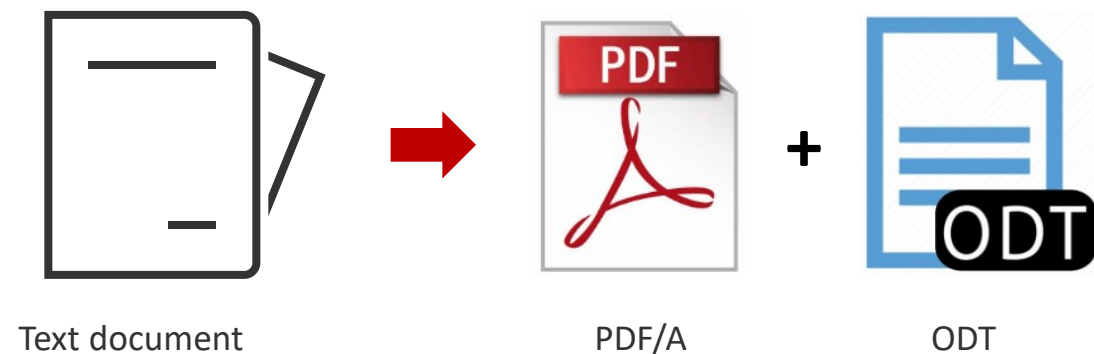


Our DPP guidance for digital text records is particularly significant because this represents 65% of the digital files in our GDA



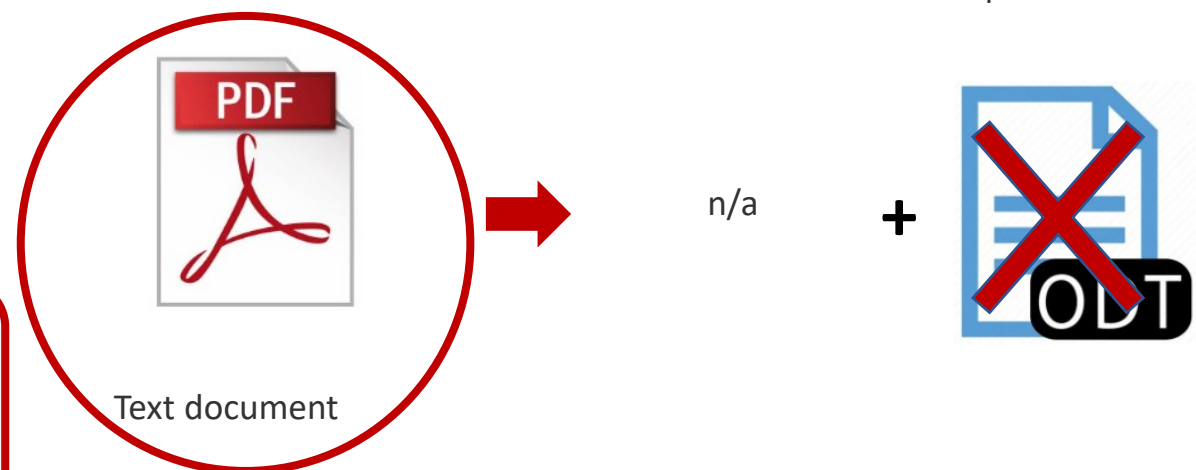
# 1. Insight: A Pivot on our Original Policy Guidance for PDFs

- ◆ Having updated the requirements for text documents in our DPP, the recommended file formats were PDF/A for preservation and ODT format for an access representation



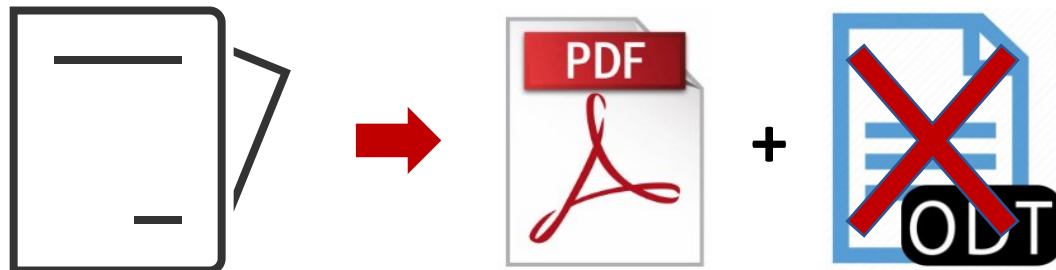
- ◆ However, in Step 4 of our project, we found that for text documents that already exist as PDFs, **there is no rule for creating a new ODT representation from that PDF**
- ◆ This meant that if a document was ingested as a PDF, there was no rule for creating a corresponding ODT access copy

**Insight: We updated the DPP requirements so that documents that already exist as PDFs do not require ODT access copies. PDF can serve as the access copy, given that it is itself a preferred file format**



## 2. Insight: Our Access Representation is Created From the Preservation Manifestation

- ◆ We found that most non-PDF text file formats had the correct preservation and access representation rules available, which seemed to indicate that the desired PDF/A and ODT would be created when these file formats were ingested
- ◆ However! In Step 7, checking that each file's migrations had worked after ingest, file after file of this type had failed to create the ODT!
- ◆ **Read the fine print:** This is how we learned that “new representation migrations apply to the Preservation representation of content only”
  - ◆ Setting up a rule to create an ODT from fmt/40 was never going to work... because the ODT needed to be created from fmt/95 (the preservation representation)!
  - ◆ As shown on the previous slide, there is no PDF to ODT rule, all of which meant the ODT was never going to happen



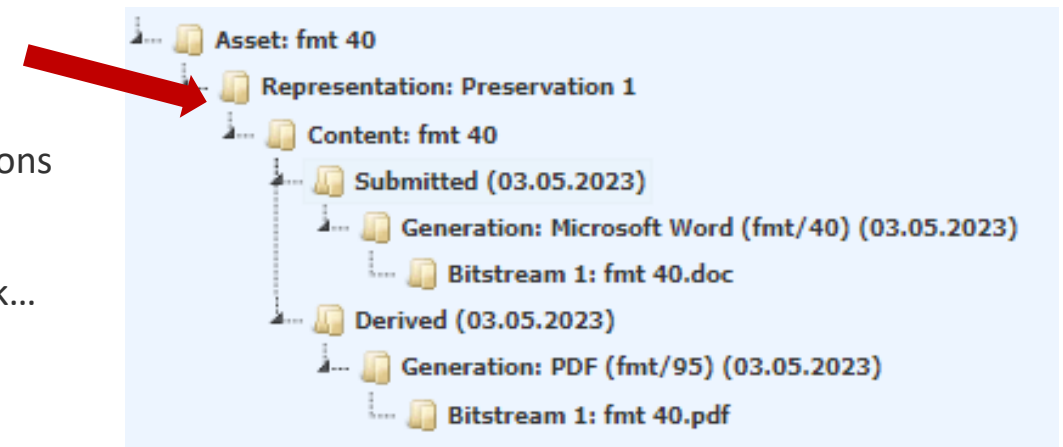
PUBLIC

Format: Microsoft Word Document 97-2003 (fmt/40)

Migration Business Rule: Normalise document to PDF/A (1A) using LibreOffice

Format: Microsoft Word Document 97-2003 (fmt/40)

Migration Business Rule: Normalise word processing document to ODT using LibreOffice



**Insight:** We redrafted the DPP requirements once more, to remove the ODT access copy requirement altogether. PDF can serve as the access copy, given that it is itself a preferred file format

### 3. Mystery: File Formats With “No Rules”

- While assessing available migration and representation workflows in our system, we encountered a number of formats for which there were simply no rules available

File Format	Version	Format F	Frequ	Which rule applies?	Normalisation Rule Availability (Preservation)	Possibl	New Representation Rule Availability (Access)	Possibl
Graphics Interchange Format	89a	fmt/4	51	Still Image	Normalise to JPEG-2000, PNG or TIFF using ImageMagick	Y	Migrate image to JPEG using ImageMagick (with other options, but has JPEG covered)	Y
Adobe InDesign Document	CS6	fmt/552	44	Presentation	No rules for format fmt/552 Adobe InDesign Document CS6	N	No rules for format fmt/552 Adobe InDesign Document CS6	N
Microsoft Powerpoint Presentation	4	x-fmt/88	43	Presentation	No rules for format x-fmt/88 Microsoft PowerPoint Presentation 4.x	N	No rules for format x-fmt/88 Microsoft PowerPoint Presentation 4.x	N
Rich Text Format	1.5-1.6	fmt/50	42	Textual	Normalise document to DOCX, PDF or PDF/A (1A) using LibreOffice, Normalise word processing document to ODT using LibreOffice	Y	Normalise document to DOCX, PDF or PDF/A (1A) using LibreOffice, Normalise word processing document to ODT using LibreOffice	Y

File formats for which we need to find migration pathways

Investigate further

# 4. Mystery: File Formats With Incorrect Rules

- ◆ When assessing available migration and representation workflows in our system, we also encountered a number of formats for which the available pathways did NOT match out DPP recommendations

File Format	Version	Format ID	Frequency	Which rule applies?	Normalisation Rule Availability (Preservation)
Windows Media Audio		fmt/132	64	Digital Audio	Do not migrate
MPEG-1 Video Format		x-fmt/385	31	Digital Video	Normalise video to Matroska with FFV1 and WAV using Ffmpeg
Microsoft Word for MS-DOS Document	4	x-fmt/274	16	Textual	Normalise document to DOCX, PDF using LibreOffice, Normalise word processing document to ODT using LibreOffice

DPP requires migration to WAV as preservation manifestation, but this is not available for Windows Media Audio

DPP requires migration to AVI as preservation manifestation, but AVI is not available for this file format

DPP requires migration to PDF/A as preservation manifestation, but only PDF is provided as an option for this file format

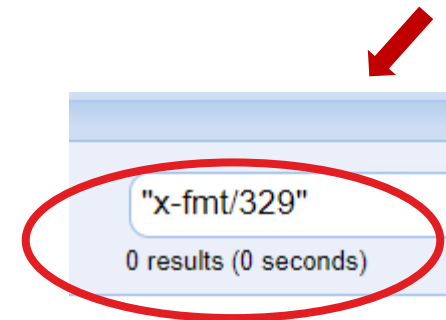
Investigate further

## 5. Mystery: “Ghost Formats”

- ◆ Our file format inventory listed all of the file formats appearing in the GDA

Interleaf Document	x-fmt/329	60
--------------------	-----------	----

- ◆ **However!** When we tried to download a copy of each file format, there were many format codes wherein the files themselves could not be located
- ◆ For example, there should be 60 files characterised as x-fmt/329, but a search for files corresponding to this format yielded 0 results





# 5. Mystery: “Ghost Formats”

- ◆ **What is happening?** There appear to be files that have a number of different associated formats; for example, the file is registered as x-fmt/129 as its primary format, but x-fmt/329 (our “ghost format”) is also associated with it
- ◆ There is only one piece of content (1 file), but that file downloads as format x-fmt/129 and only that primary format x-fmt/129 is actually preserved
- ◆ Do we require digital preservation for the rest of these file formats?

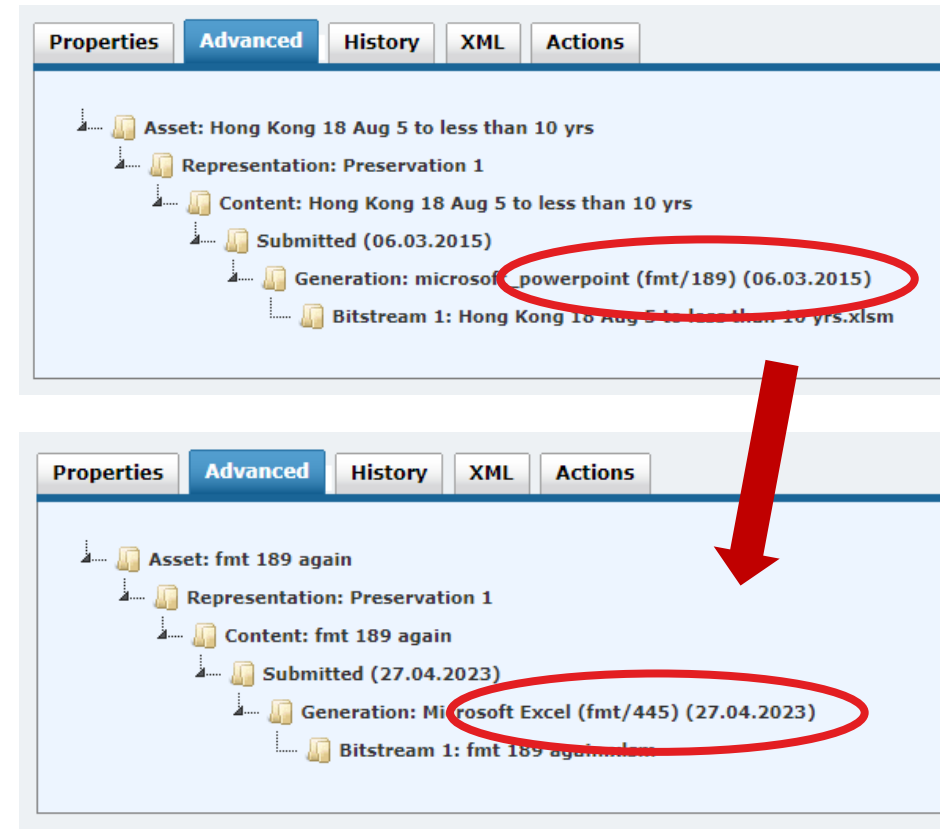
## EXE0513.DOC

Description	Technical Metadata	History
Formats		
Name	PUID	Version
Microsoft Word for Macintosh Document	x-fmt/129	X
Wordperfect Secondary File	x-fmt/42	5.0
Stationery for Mac OS X	x-fmt/131	
Microsoft Word for Macintosh Document	x-fmt/2	6.0
Microsoft Word for MS-DOS Document	x-fmt/273	3.0
Wordperfect Secondary File	x-fmt/43	5.1/5.2
Interleaf Document	x-fmt/329	
Fixity		
Name	Value	

Investigate further?

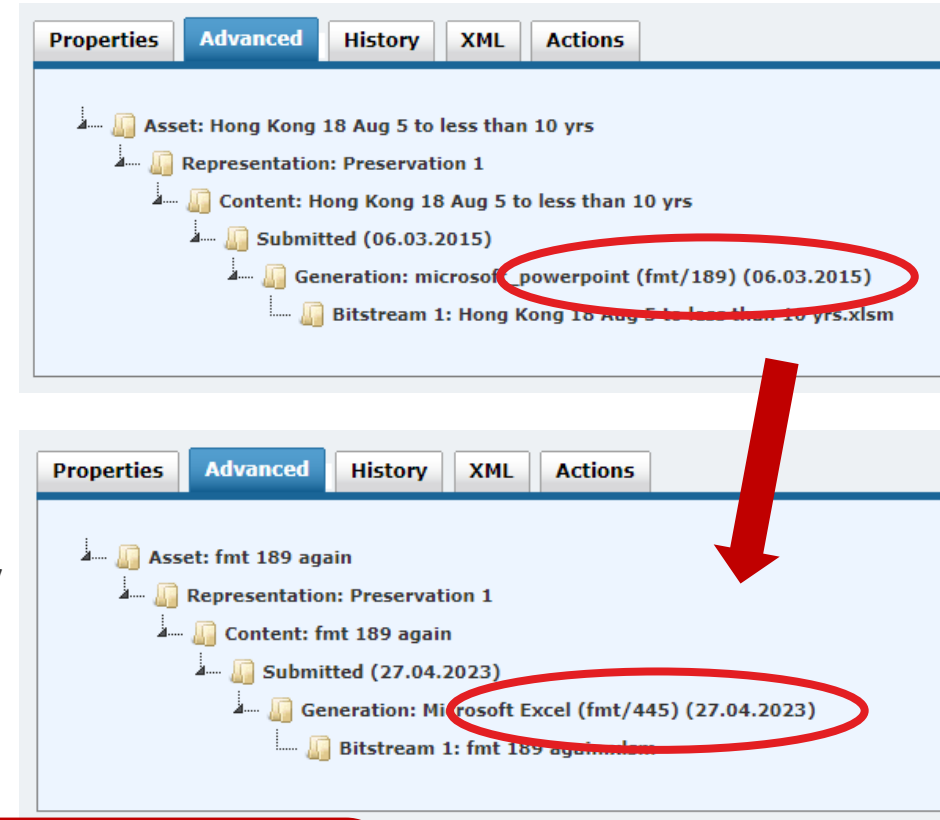
## 6. Insight: Re-characterise Before Implementing the DPP

- ◆ When ingesting our file format samples to the Test version, a number of files did not ingest as the same format that had been downloaded
- ◆ As shown here, fmt/189 was re-characterised as fmt/445, and **this was not an isolated observation**: many file formats reflected updates after ingest



## 6. Insight: Re-characterise Before Implementing the DPP

- ◆ **Re-characterisation required:** A lot of the content on the GDA was last characterised on ingest, which could be as long ago as 2015, and it was now being ingested into a digital preservation system with more up-to-date format identification
- ◆ All of which sparked a number of important realisations:
  1. **Start from an accurate and up-to-date file format list:** Our original file format project spreadsheet reflected “old” characterisation in many cases, meaning we were checking for rules and migrations for out-of-date formats
  2. **A number of our mysteries might be solved by simply re-characterising:** “No rule” formats, formats missing migration pathways, and “ghost formats” might all disappear by simply tidying up the database with an up-to-date re-characterisation



**Insight:** As a first step in our project, we should have completed a full re-characterisation to ensure our file inventory starts from an accurate and up-to-date point. **Our plans now are to go back to address this as the next step in our process!**

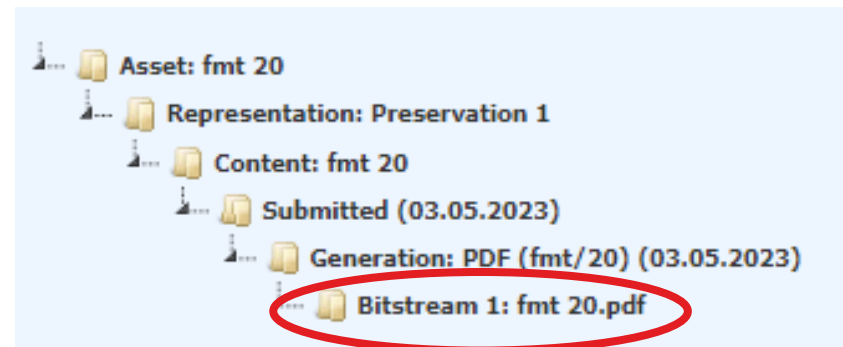
# 7. Mystery: Failed Migrations

- ◆ The majority of our rules were executed successfully, with the desired preservation format and/or access representation being correctly applied upon ingest
- ◆ However, a number of tests showed that the chosen rules did NOT work, indicating areas that we must explore further before implementing the DPP

For example, we definitely had this rule set for the migration of fmt/20 to PDF/A

**Format:** Acrobat PDF 1.6 - Portable Document Format 1.6 (fmt/20)  
**Migration Business Rule:** Normalise document to PDF/A (1A) using LibreOffice

The file was not migrated upon ingest



Investigate further

## 8. Insight: Tidy Up Zip Files

- ◆ Our DPP does not have preservation guidance for ZIP format, and our team actively discourages users from ingesting ZIP to the GDA; there are no migration pathways for ZIP in the system
- ◆ Despite this, the file format inventory shows a relatively small number of ZIP in the GDA
- ◆ As expected, no migrations occur for this format; the ZIP file can be downloaded and then unzipped to reveal its contents (in this case, a set of Word docs), but the contents themselves do not undergo preservation workflows while “zipped”
- ◆ **How do we ensure that the contents/files within these ZIP files are adequately preserved according to our policy?**

The image shows a screenshot of a file format inventory table and a file properties window. The table has columns for File Format, Version, Format ID, Frequency, and Normalisation Rule Availability (Preservation). A row for 'ZIP Format' is highlighted in orange, with a message: 'No rules for format x-fmt/263 ZIP Format'. Below the table is a file properties window for 'Asset: xfmt 263 CLOSED'. The 'Advanced' tab is selected, showing a tree structure: 'Representation: Preservation 1' -> 'Content: xfmt 263 CLOSED' -> 'Submitted (15.05.2023)' -> 'Generation: ZIP (x-fmt/263) (15.05.2023)' -> 'Bitstream 1: xfmt 263 CLOSED.zip'. A red arrow points from the 'Bitstream 1' entry to a file icon labeled 'xfmt 263 CLO...'. Another red arrow points from the file icon to a list of files: 'Bring Your Own Device -', 'Bring Your Own Device -', and 'BYOD - poster design -'.

File Format	Version	Format ID	Frequency	Normalisation Rule Availability (Preservation)
ZIP Format		x-fmt/263	376	No rules for format x-fmt/263 ZIP Format

Properties Advanced History XML Actions

Asset: xfmt 263 CLOSED

Representation: Preservation 1

Content: xfmt 263 CLOSED

Submitted (15.05.2023)

Generation: ZIP (x-fmt/263) (15.05.2023)

Bitstream 1: xfmt 263 CLOSED.zip

xfmt 263 CLO...

Bring Your Own Device -

Bring Your Own Device -

BYOD - poster design -

**Insight:** Our current thinking is that we will need to have a separate side project at a later time to “tidy up” these ZIP files - - downloading each, unzipping them, and then re-ingesting the contents to ensure these are all properly preserved. More investigation is needed.



## 9. Mystery: Failed Chinese Language Migrations

ften used during Lunar New Year... in the doorways or walls to create festive atmosphere... The two Fai Chun I have selected are 開工大吉 ("Good luck to our work," which people usually hang in offices) and 大吉大利 (Good luck and all (photos, above right) "is very rare in Hong sells incenses and other items to worship Fai Chun around the new year time." And Chinese horoscope, Tin Pui also shared a

**This is our original Word file**

≠

Year... in the doorways or walls to create festive atmosphere... The two Fai Chun I have selected are 開工大吉 ("Good luck to our work," which people usually hang in offices) and 大吉大利 (Good luck and all go smoothly)." The Fai Chun shop (photos, above right) This small shop usually sells incense ancestors and gods, and they include time." And, as it is the Year of the Rat Pui also shared a photo (left) of her pe

**Preservation PDF is missing important information**

- ◆ If we download the original ingested DOC file from our system, it still has the correct characters, so it's not the ingest process
- ◆ If we "print to PDF" on this original DOC file, the PDF is created with the correct characters, so it's not the PDF creation
- ◆ This also happens for LibreOffice migration workflows to ODP and ODS for PowerPoint and Excel
- ◆ This also happens for select other language characters

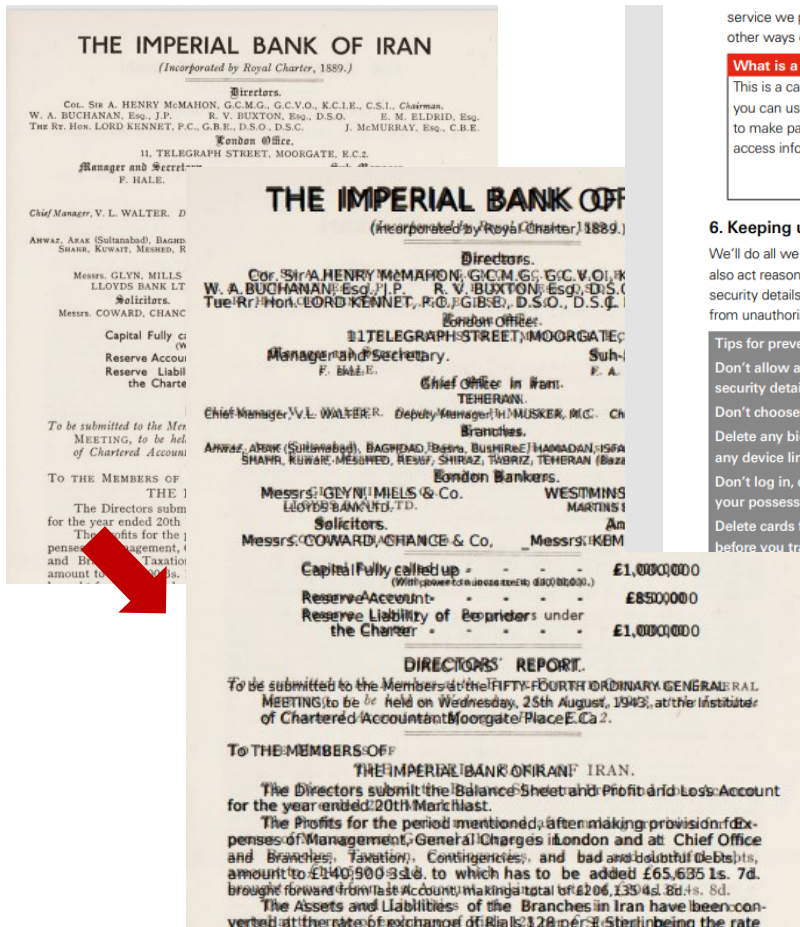
### Our hypothesis:

- ◆ The issue appears to be that the preservation workflow is NOT migrating or reading these characters

Investigate further

# 10. Mystery: Migrations Not Preserving Original Formatting for PDFs

- ◆ In our tests, the preservation rules for certain PDF file types appeared to have successfully migrated to PDF/A
- ◆ **However!** when the preservation manifestations were viewed, there were a number of distortions



service we provide for accessing your accounts) or with a payment device. We may add other ways of checking your identity in the future.

What is a 'payment device'?	What are 'security details'?
This is a card, e-wallet or another device that you can use (usually with security details) to make payments, give us instructions or access information about your account.	These can be passwords, PINs, security codes, memorable data, biometric data (such as a fingerprint or voice ID) or any other information or details that you have set up with us for use with a payment device or Telephone, Online or Mobile Banking.

## 6. Keeping up security

We'll do all we reasonably can to prevent unauthorised access to your account. You must also act reasonably to prevent misuse of your account by keeping your payment devices and security details safe. If you don't, we may block access to your account to protect you and us from unauthorised use. If you ask, we'll tell you when the block will end.

Tips for preventing misuse of your account:

- Don't allow anyone (other than a TPP) to use any payment device or tell them your security details.
- Don't choose security details that can be easily guessed by anyone else.
- Delete any biometric access to a payment device (such as fingerprint, face or voice ID) or any device linked to a payment device by anyone else.
- Don't log in, or stay logged in, to a device if you're not in full control of it or it's not in your possession.
- Delete cards from a payment device or from any e-wallets accessed from a payment device before you transfer possession of it to someone else (for example if you sell or give it to someone for repair).

- In all other cases, we'll assume we're dealing with you and that we are authorised to take action if your security details are used with Telephone, Online or Mobile Banking (or another service we provide for accessing your accounts) or with a payment device. We may add other ways of checking your identity in the future.

What is a 'payment device'?	What are 'security details'?
This is a card, e-wallet or another device that you can use (usually with security details) to make payments, give us instructions or access information about your account.	These can be passwords, PINs, security codes, memorable data, biometric data (such as a fingerprint or voice ID) or any other information or details that you have set up with us for use with a payment device or Telephone, Online or Mobile Banking.

## 6. Keeping up security

We'll do all we reasonably can to prevent unauthorised access to your account. You must also act reasonably to prevent misuse of your account by keeping your payment devices and security details safe. If you don't, we may block access to your account to protect you and us from unauthorised use. If you ask, we'll tell you when the block will end.

Tips for preventing misuse of your accounts:

- Don't allow anyone (other than a TPP) to use any payment device or tell them your security details.
- Don't choose security details that can be easily guessed by anyone else.
- Delete any biometric access to a payment device (such as fingerprint, face or voice ID) or any device linked to a payment device by anyone else.
- Don't log in, or stay logged in, to a device if you're not in full control of it or it's not in your possession.
- Delete cards from a payment device or from any e-wallets accessed from a payment device before you transfer possession of it to someone else (for example if you sell or give it to someone for repair).

Additional tips for protecting your account and for keeping your payment devices and security details safe are on our website and in the Banking made easy brochure. Please check for any updates we may make from time to time.



Investigate further

But also! We updated our process to include quality checks for each test file



# Summary & Next Steps



# Next Steps in Implementing the DPP

Solve our mysteries before proceeding to full implementation on the Live system:

1. Complete a full re-characterisation
2. Go through our Test version project steps once again with this updated file format inventory
  - a) Assess whether all formats have the required preservation workflows; are there formats with no rules or the incorrect rules?
  - b) After assigning rules and running another test ingest, are any migrations failing?
3. Investigate the observed distortions and character loss arising from certain preservation workflows



**This will need to be solved**



**This will need to be solved**




**This will need to be solved**

**Once these solutions are found, and our re-tests are successful...**

**We will then implement the DPP on the Live version of our GDA!**

# Key Lessons Learned as New Digital Archivists

1. **Make policy pivots as required:** For example, DPP-required workflows for PDF migrations to ODT were not available, and we realised the PDF was sufficient in itself
2. **Our access representations are created from the preservation manifestation:** No need to set access rules for the original file format type
3. **Run a full re-characterisation at the start of the project:** Ensure the file inventory is up-to-date and accurate so that the correct preservation rules can be assigned and assessed
4. **Keep the project moving forward by maintaining a list of challenges to investigate later:** Learn to prioritise major challenges rather than trying to solve each question about each file format at each step
5. **Check the files:** In your test runs, do not simply take a successful preservation workflow confirmation as a sign that things have worked. Be sure to quality check the preservation manifestation to make sure it is correct. **Give some consideration to how much you want to rely on automated preservation processes**

 **Overall:** Tackling a big DPP implementation project has been an excellent and enjoyable way to get hands-on experience and learning as new digital archivists



# Thank you!



**Please visit –** [history.hsbc.com](http://history.hsbc.com)

**Contact –** [digitalarchives@hsbc.com](mailto:digitalarchives@hsbc.com)