Jisc Research Data Shared Service (RDSS) pilot project: summary of key issues from preservation tool testing
University of Glasgow

The University of Glasgow joined Jisc's RDSS pilot project in August 2017 to explore the practical implementation of digital preservation, evaluate current provision within the University for managing and preserving digital records, and share best practice with other pilot participants. During the pilot, we have tested a range of functions in the Jisc versions of Archivematica and Preservica, covering ingest, file format identification, format migration and metadata management, and have evaluated user access, error notification and reporting. Where possible, we ran the same test on both tools, using identical data. Testers have full user rights but do not have access to system administrator functions and logs. Some outcomes may depend on the way our workflows were configured, so other testers may obtain different outcomes. Alongside the testing, we have considered how digital preservation workflows could be integrated with systems and processes within our organisation. Our pilot project focused on the preservation of University corporate records and archival material, as systems are already in place to manage research data.

Overall, each tool successfully ingested a variety of file types, carried out preservation actions on them and produced archival information packages. The issues we encountered related mainly to configuring workflows, managing metadata and ingested material, validating processes and understanding messages generated by the tools. We have not tested integrations with other systems or storage to date, so this summary focuses on the preservation tools themselves. Jisc's forthcoming hosted third-party service for the RDSS pilot will move participants to Preservica version 5.10 and Archivematica v1.8. These updated versions may address some of the points raised in this report.

The aim of this document is to help foster discussion about user needs and priorities for the Jisc RDSS preservation service and contribute to the development of the preservation tools.

1. Transfer and ingest

Key issue: processing zipped content

1.1 Zipped content

Both preservation tools encountered difficulties uploading zipped content. Archivematica was able to extract and process the contents of zipped folders nested within an unzipped directory but Preservica does not unzip certain types of zipped folder at all e.g. 7-z format. Neither tool can upload a zipped directory via their standard transfer workflow as these are viewed as single files, not directories. We have tried uploading zipped directories to Archivematica using the BagIt specification:[1] this worked for certain zipped formats but not all. We have not tested this yet in Preservica. Currently, University of Glasgow researchers forwarding large quantities of completed research data for storage place it in zipped folders for transfer, so this is an issue which we will need to investigate further to ensure that material can be transferred easily.

2. File format identification

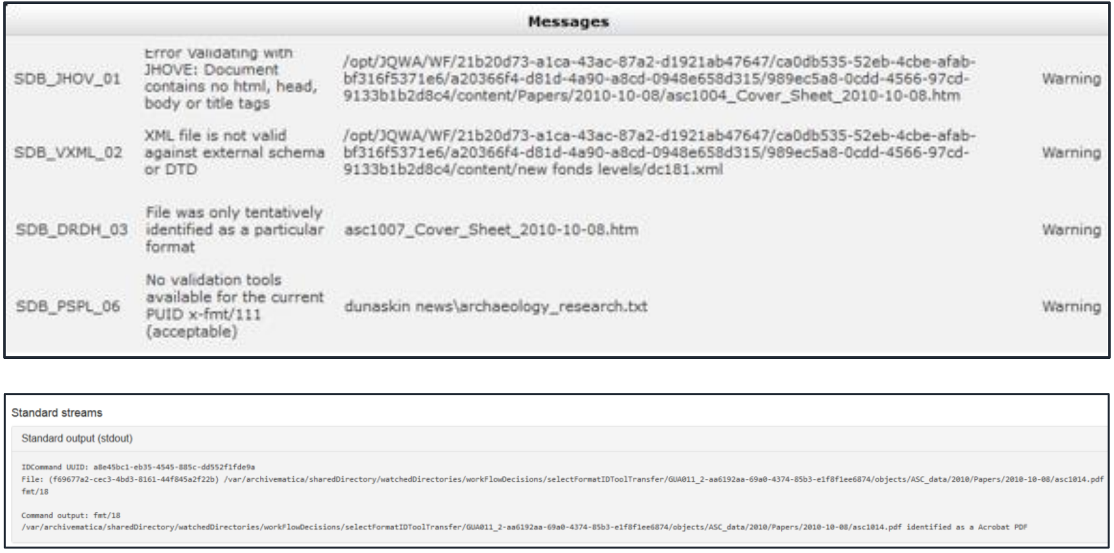
Key issues: accurate identification of file formats; user-friendly supporting processes

2.1 Effectiveness

The format identification tools used by Archivematica and Preservica were not able to accurately identify all of the file formats ingested, although they were all widely-used document, image and audio-visual formats. In Archivematica, each of the three identification tools available generated different results for some formats. Sometimes, Preservica's analysis differed from Archivematica's. Consequently, before ingesting digital content, we may need to run trials to identify any potentially problematic formats in advance, especially with commonly-used file types.

2.2 Reports on file identification processes

Neither Archivematica or Preservica produces an easily-readable, accessible report of format identification outcomes for an ingested directory. The results cannot be sorted or searched and the order in which the files are listed in the report does not mirror their arrangement within the directory, making it difficult to find the outcome for any specific file, especially in a large dataset. Preservica's pop-out window cannot be made fullscreen for ease of reading, while the completed ingest report does not flag up the presence of format identification warnings and errors. Neither tool enables the file identification report to be saved independently or exported, which would be a useful feature, as this would make it easier to collate and interrogate the results and identify problematic formats within the organisation.



Messages			
SDB_JHOV_01	Error validating with JHOVE: Document contains no html, head, body or title tags	/opt/JQWA/WF/21b20d73-a1ca-43ac-87a2-d1921ab47647/ca0db535-52eb-4cbe-afab-bf316f5371e6/a20366f4-d81d-4a90-a8cd-0948e658d315/989ec5a8-0cdd-4566-97cd-9133b1b2d8c4/content/Papers/2010-10-08/asc1004_Cover_Sheet_2010-10-08.htm	Warning
SDB_VXML_02	XML file is not valid against external schema or DTD	/opt/JQWA/WF/21b20d73-a1ca-43ac-87a2-d1921ab47647/ca0db535-52eb-4cbe-afab-bf316f5371e6/a20366f4-d81d-4a90-a8cd-0948e658d315/989ec5a8-0cdd-4566-97cd-9133b1b2d8c4/content/new fonds levels/dc181.xml	Warning
SDB_DRDH_03	File was only tentatively identified as a particular format	asc1007_Cover_Sheet_2010-10-08.htm	Warning
SDB_PSPL_06	No validation tools available for the current PUID x-fmt/111 (acceptable)	dunaskin news\archaeology_research.txt	Warning

Standard streams

Standard output (stdout)

```
IDCommand UUID: a8e45bc1-eb35-4545-885c-dd552f1fde9a
File: (f69677a2-cec3-4bd3-8161-44f845a2f22b) /var/archivematica/sharedDirectory/watchedDirectories/workFlowDecisions/selectFormatIDToolTransfer/GUA011_2-aa6192aa-69a0-4374-85b3-e1f8f1ee6874/objects/ASC_data/2010/Papers/2010-10-08/asc1014.pdf
fmt/18

Command output: fmt/18
/var/archivematica/sharedDirectory/watchedDirectories/workFlowDecisions/selectFormatIDToolTransfer/GUA011_2-aa6192aa-69a0-4374-85b3-e1f8f1ee6874/objects/ASC_data/2010/Papers/2010-10-08/asc1014.pdf identified as a Acrobat PDF
```

Preservica (top) and Archivematica (bottom) file identification outcome reports

2.3 PRONOM registry entries

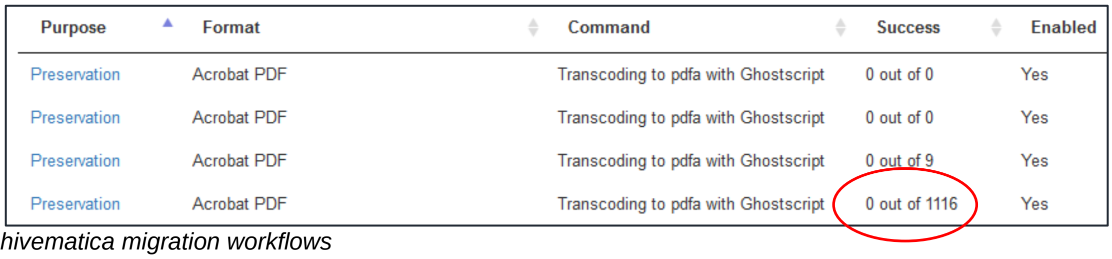
The UK National Archives' PRONOM registry[2] supports the identification and preservation of file formats. It is possible to submit updates and new content to PRONOM but this is a time-consuming process and National Archives staff have limited capacity to process submissions. It would be worth considering whether Jisc and/or the RDSS community could work with the National Archives to support the update process for PRONOM.

3. Migration to new formats (normalisation)

Key issues: process to create migration workflows; reporting and validation

3.1 Running and configuring migration workflow

Archivematica comes with a number of pre-configured migration workflows, which can be enabled or disabled, depending on user preference. The migration pathways can be modified or new ones created, although we have not tried this as it requires some programming. Usefully, the Archivematica dashboard shows which migration workflows have been enabled and whether they are running successfully. For example, this has highlighted that migration workflows converting different versions of Acrobat pdf format files to pdf/a format are not functioning correctly, which we have fed back to Artefactual.



Purpose	Format	Command	Success	Enabled
Preservation	Acrobat PDF	Transcoding to pdfa with Ghostscript	0 out of 0	Yes
Preservation	Acrobat PDF	Transcoding to pdfa with Ghostscript	0 out of 0	Yes
Preservation	Acrobat PDF	Transcoding to pdfa with Ghostscript	0 out of 9	Yes
Preservation	Acrobat PDF	Transcoding to pdfa with Ghostscript	0 out of 1116	Yes

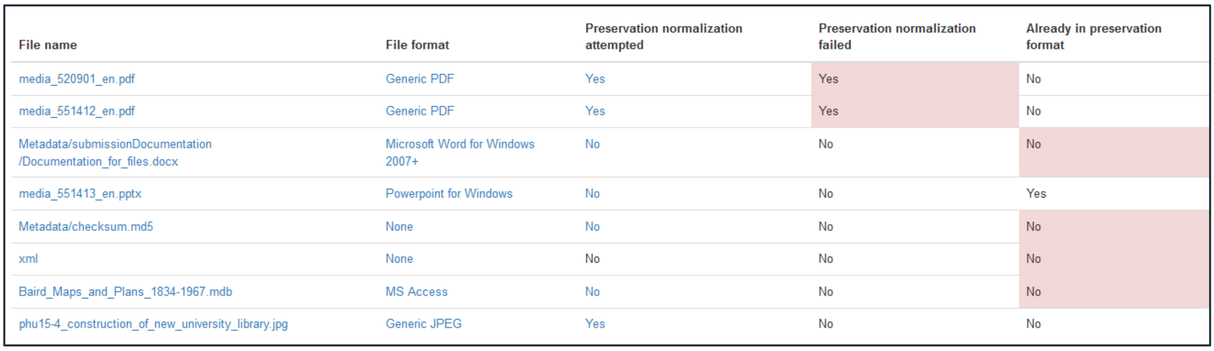
Archivematica migration workflows

No migration workflows come pre-configured with the Jisc RDSS version of Preservica but they are relatively straightforward to set up. However, each time we run these workflows, we have to enter certain criteria, including the PRONOM unique identifier for each file format we want to migrate from, which means collating that information before running the workflow. It may be that there is a way to set up the workflows which eliminates this step and this is something we will check with Jisc and Preservica.

As RDSS participants will probably have similar requirements, rather than each organisation configuring their own migration workflows, this may be something that the Jisc community could develop together and share.

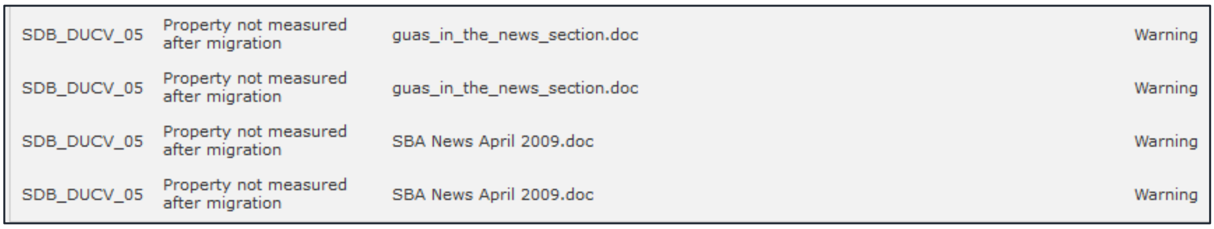
3.2 Validating and reporting on data migration

Preservica validates each file which it has migrated to a new format and informs the user if the integrity of the migrated version is less than 80% accurate. We have encountered the warning 'property not measured after migration' for some migrated data and are currently liaising with Preservica/Jisc's RDSS helpdesk to understand the implications of this. Archivematica does not carry out any post-migration validation, so users require their own validation workflows. Neither tool's format migration workflow report is straightforward to interpret. Archivematica's report states that normalisation did not fail but does not say explicitly whether it succeeded. Similarly, Preservica's workflow report presents a list of both original and migrated files, but does not describe the actions which have been carried out. Greater clarity would assist system users to interpret migration outcomes and quickly identify issues.



File name	File format	Preservation normalization attempted	Preservation normalization failed	Already in preservation format
media_520901_en.pdf	Generic PDF	Yes	Yes	No
media_551412_en.pdf	Generic PDF	Yes	Yes	No
Metadata/submissionDocumentation /Documentation_for_files.docx	Microsoft Word for Windows 2007+	No	No	No
media_551413_en.pptx	Powerpoint for Windows	No	No	Yes
Metadata/checksum.md5	None	No	No	No
xml	None	No	No	No
Baird_Maps_and_Plans_1834-1967.mdb	MS Access	No	No	No
phu15-4_construction_of_new_university_library.jpg	Generic JPEG	Yes	No	No

Archivematica migration outcome report



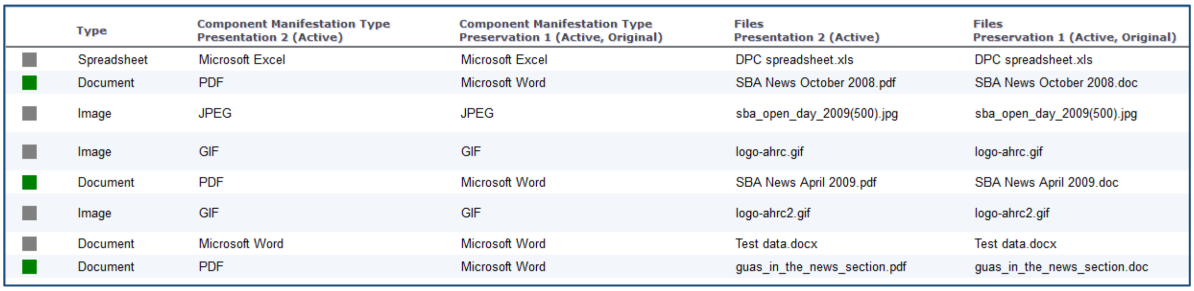
SDB_DUCV_05	Property not measured after migration	guas_in_the_news_section.doc	Warning
SDB_DUCV_05	Property not measured after migration	guas_in_the_news_section.doc	Warning
SDB_DUCV_05	Property not measured after migration	SBA News April 2009.doc	Warning
SDB_DUCV_05	Property not measured after migration	SBA News April 2009.doc	Warning

Preservica migration outcome report

3.3 Management of file versions (original, preservation and access)

Archivematica does not show the relationship between original and migrated versions of digital material inside the software. However, within its Archival Information Packages, the original file and its preservation version are saved in the same folder, presenting the two versions side by side. The METS file documents all actions carried out on ingested files, but the quantity of information within the METS file can make it hard to find specific elements.

We found that Preservica does not always clearly demonstrate the relationship between original and preserved versions of files. In the Explorer dashboard, within a deliverable unit's properties, the *Compare* tab provides an overview of all of the manifestations which exist for each file, showing clearly those which have a preservation or access (presentation) version. However, when viewing the properties of a preservation version, the metadata shows that the file was generated using the post-ingest migration workflow but does not provide information on the original file or contain a link to it. The metadata for the original version makes no reference to any migration actions.



	Type	Component Manifestation Type Presentation 2 (Active)	Component Manifestation Type Preservation 1 (Active, Original)	Files Presentation 2 (Active)	Files Preservation 1 (Active, Original)
■	Spreadsheet	Microsoft Excel	Microsoft Excel	DPC spreadsheet.xls	DPC spreadsheet.xls
🟩	Document	PDF	Microsoft Word	SBA News October 2008.pdf	SBA News October 2008.doc
■	Image	JPEG	JPEG	sba_open_day_2009(500).jpg	sba_open_day_2009(500).jpg
■	Image	GIF	GIF	logo-ahrc.gif	logo-ahrc.gif
🟩	Document	PDF	Microsoft Word	SBA News April 2009.pdf	SBA News April 2009.doc
■	Image	GIF	GIF	logo-ahrc2.gif	logo-ahrc2.gif
■	Document	Microsoft Word	Microsoft Word	Test data.docx	Test data.docx
🟩	Document	PDF	Microsoft Word	guas_in_the_news_section.pdf	guas_in_the_news_section.doc

Preservica: details of versions (manifestations) of files within a deliverable unit

4. Metadata management

Key issues: uploading metadata as metadata; accessing metadata with preserved digital material

4.1 Uploading metadata and managing metadata schemas

The two preservation tools have quite specific configuration requirements to enable uploaded metadata to be recognised and read as metadata. In Archivematica, metadata must conform to the Dublin Core metadata schema and be uploaded in csv files. Preservica requires metadata to be uploaded in xml format but different metadata schemas can be used to validate the metadata, for example, Datacite. However, each schema needs to be uploaded and configured in Preservica first of all, which is quite a lengthy process. Uploaded schemas are not automatically updated when a new version is released so they will require ongoing management.

Again, this is an area where the Jisc community could work together to produce guidance on transforming metadata into the required formats, and share configured schemas ready for use, rather than each institution developing these independently.

4.2 Discovering and accessing preserved metadata

Within both preservation tools, we found that metadata embedded with ingested material can be difficult to find and access. Archivematica has limited search functionality and not all metadata is searched. In Preservica, while metadata is discovered during searching, it cannot always be viewed within the context of its directory. At present, we have only carried out limited testing on uploading metadata via a schema but it appears that metadata uploaded this way is more discoverable. For example, metadata uploaded to Archivematica is stored within the METS file. When an Archival Information Package is downloaded, metadata within the METS file can be discovered by standard operating system search tools. We plan to carry out more testing around metadata.

5. Functionality

Key issues: Error notification unclear. Difficult to navigate search results.

5.1 User control and access

Preservica offers granular user control, enabling system administrators to manage which functions each user group has access to and the records which they can view and modify within the system. However, in Archivematica, every registered user has full access to all standard functions. If we use automated ingest processes, this may not be a problem, as preservation staff may be the only people required to log into the system. However, if interactive transfers are envisaged, data privacy could become an issue or risks associated with inadvertent mistakes could be higher. Ensuring that user access is managed appropriately will be an important consideration as we plan and develop our preservation workflows.

5.2 Reporting

Preservation staff will need to verify that individual ingests have been processed correctly, manage ongoing preservation and collate statistics. Preservica offers a useful range of reports, covering both individual ingests and data on all files processed by the tool. Outputs can usually be tailored by setting parameters e.g. a specific date span, and there is a choice of output formats. While it is possible to write customised reports, this does require knowledge of Java and has not been tested.

Archivematica does not provide any reporting functionality which collates information about ingested material. Some reports are generated during ingest processes, for example, file format identification outcomes, but these reports cannot be saved separately, so their wider use is limited. Users may therefore need to find another way of collecting this data.

5.3 Error notification

For both tools, we found that error messages are not easy to understand and errors are not always clearly signposted to users. For example, in many of the tests we ran in Preservica, the ingest report indicated that the file identification process had been completed successfully. However, the identification workflow report turned out to be full of warnings and error messages. When an ingest failed to complete, Preservica did not provide any e-mail notification, so we needed to actively verify each ingest.

In Archivematica, when a process encounters problems, it may be highlighted in red within the dashboard but again, the user is not always notified so pro-active checking is necessary. Conversely, for some failed processes, failure notifications are generated and are e-mailed to all of the institution's registered users, not just the active user.

We feel that notification of errors to the relevant user and clearer explanations of the problem encountered would assist users to deal with errors more quickly and effectively.

5.4 Searching and sorting

In Archivematica, only some hitlists can be sorted e.g. rules. Submission Information Packages in the Transfer and Ingest dashboards are not listed in any order and cannot be sorted or searched for, so the user has to scroll down the list to find a specific entry. Where search functionality is available, it is limited to searching the contents of one dashboard at a time. Preservica has good search functionality but again, it is not possible to sort the search results to manage the results and identify relevant entries. Working with a limited number of records for the pilot, it is possible to browse lists and find relevant material; in an active preservation system, this becomes less effective, so being able to find contents efficiently will be important.

6. Integrations with other software

To date, we have not tested how these two preservation tools will integrate with our existing software. As we develop a clearer idea of where digital preservation processes will fit into existing University workflows, then we will identify what integrations we will require. Jisc has already developed integrations for certain software, so again, this may be an area where pilot institutions can work together to establish priorities for further integrations and support their development. We have spoken to colleagues at other institutions about the software integrations which they have in place, which was very helpful; if other pilot participants are able to share their experience of managing integrations with the Jisc community, that would be very beneficial. Key for us will be integrations which are straightforward to set up and which work seamlessly for users.

7. Jisc RDSS model

At present, no firm details are available about what Jisc's preservation service will offer and the costs involved. Our priorities are for a service which is efficient and cost-effective, user-friendly for both preservation staff and researchers and which integrates with the University's existing systems. We welcome a collaborative approach, encouraging participants to contribute feedback and expertise to create shared outcomes and support development.

Conclusion

Participation in Jisc's RDSS project has enabled us to develop a working knowledge of a couple of digital preservation tools and improved our understanding of digital preservation principles. We have met with other pilot participants and shared experiences and best practice. Within the University, we have assessed how digital records are currently managed and preserved and the risks which they may face long-term. Seeing how the preservation tools work has enabled us to evaluate how digital preservation functions might integrate with current workflows. We have developed a digital preservation framework and are now looking at next steps to take forward digital preservation within the University.

[1] For information about the BagIt specification, see the Digital Curation Centre website http://www.dcc.ac.uk/resources/external/bagit-library .

[2] National Archives UK PRONOM registry https://www.nationalarchives.gov.uk/PRONOM/Default.aspx .