IJDC | *Brief Report*

Capturing the Cloud: Towards SharePoint Transfer at UK Parliament

Nicole Hartland
UK Parliament

Emily Chen
UK Parliament

Rosemary Reynolds
UK Parliament

Abstract

Since 2020, UK Parliament has moved towards cloud-based ways of working and collaboration, with colleagues across both Houses increasingly storing and sharing most of their information in Microsoft SharePoint. In response to this shift, the Parliamentary Archives sought to establish an end-to-end process to transfer information of archival value out of SharePoint and into the Digital Repository. Three challenges, unique to the cloud-based and collaborative nature of this environment, arose: defining the authoritative version, extracting files from the cloud with properties and metadata intact, and validating and authenticating files in the cloud. This brief report outlines the Parliamentary Archives efforts to explore and test the transfer and authentication of archival data from the cloud and into their digital repository with a focus on building trust and transparency.

Submitted 31 May 2024 ~ *Accepted* 10 June 2024

Correspondence should be addressed to Nicole Hartland. Email: hartlandn@parliament.uk

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v18i1.987
DOI: 10.2218/ijdc.v18i1.987

Background Context and Drivers

From 2013/14, Parliament began an initial rollout of Office 365 with the option to 'opt-in' to using SharePoint 2013. When SPIRE, our previous Electronic Document and Records Management System (EDRMS), was reaching end of life in 2016, teams across Parliament moved to using SharePoint. This was part of a multi-year project (2017-2020) to move Parliament away from network sharing and personal drives hosted on aging on-premises services to a predominantly cloud-based way of collaborative working. The previous EDRMS had Parliament's Authorised Retention and Disposal Policy (ARDP) 'baked in', and the move to using SharePoint as Parliament's corporate repository established a new system of site and library design architecture developed by the Information and Records Management Service (IRMS) and the Productivity and Collaboration (P&C) team. This shift introduced greater flexibility for users with different uptake of functionality such as metadata-based organisation and application of retention labels across different business areas. Crucially, IRMS's involvement in the Microsoft 365 (M365) rollout enabled a focus on core information management documentation and policies alongside the greater flexibility the shift to M365 provided.

Lessons Learnt

Alongside rolling out this cloud-based collaboration platform across UK Parliament, we had to transfer and migrate all of our old information and data from SPIRE into the new system – tens of thousands of records. There were two key lessons learnt as part of this work, which we identified were applicable to establishing a transfer process for SharePoint now. Firstly, to take an iterative approach to process development – especially important with live data. Secondly, collaboration between the Information and Records Management Service and the Digital Preservation team was crucial to the success of this programme of work.

Disposal in SharePoint

Despite the many benefits to this shift towards cloud-based storage and sharing, we faced a number of distinct challenges. Parliament's risk-averse approach to information disposal resulted in a large volume of information waiting for human approval to delete. The amount of information sitting in this 'pending disposition' queue in Microsoft Purview, Microsoft's native disposal tool, meant that IRMS encountered difficulties in searching, filtering, and reviewing information in line with applied retention labels. This necessitated a more manual approach to Disposal in SharePoint, harnessing the expertise and trust of Record Officers (ROs) and Information Asset Owners (IAOs) across different business areas in Parliament. Our aims for SharePoint transfer were to build upon the foundations of the Disposal project to ensure the transfer of digital information from SharePoint to the digital repository is a trusted process, with a high degree of transparency between Digital Preservation and IRMS as well as the wider business. As available commercial products did not fit our business needs and risk-averse approach, we needed to explore native functionality to export information from the SharePoint environment and build an initial transfer workflow. This 'proof of concept' could then lay the foundations for scaling up SharePoint Transfer.

Drivers

There were three key drivers to establishing, testing, and piloting a process to transfer information from SharePoint to the Digital Repository for permanent preservation:

- Business need – increasingly, teams and departments across UK Parliament were approaching IRMS, identifying information within their SharePoint sites which should be in the archive, but still remains on the site, and asking what to do with it.

- 'We've just got to do it' – as information professionals, we know this information does need to be permanently preserved. This is around a third of all information across Parliament's SharePoint sites, so this problem is not going to go away. We needed to take a proactive approach despite the challenges in establishing a transfer process.

- System end-of-life? – our biggest driver for transfer with SPIRE was that the EDRMS was coming to end of life and falling out of support. The impact of this was huge, and at one point we were transferring tens of thousands of records out of the system per day. The likelihood of a similar situation occurring with SharePoint is low, given the size and stability of Microsoft as a company.

Building a Trusted Process

From this foundation, we worked towards building a process for identifying and extracting information of archival value from SharePoint. Three key challenges to ensuring trust and transparency in SharePoint transfer, unique to cloud-based environments, were identified which we tested and integrated into our transfer processes.

Defining the Authoritative Version

Defining which version of a file is authoritative within the cloud-based environment was a key challenge. The nature of cloud-based collaboration software is to promote simultaneous editing and sharing of files. Information management policies, guidance, and support are central to managing this challenge, namely:

- Clearly defined disposal and transfer instruction in the ARDP, and the provision of ongoing guidance and business support to enable ROs to manage this at a local level.

- Guidance around file sharing and access, controlled through M365 'copy link' functionality and access groups – discouraging multiple copies and ad hoc access.

- Clearly marking files for disposal as 'read-only' and restricting access to control further modification within the SharePoint environment.

- Clear ownership within IRMS for exporting metadata and folder structures to enable transfer, management of source data deletion post-transfer, and ongoing communication with ROs and IAOs.

Validation Criteria and Methods

Once we have established which file is the 'authoritative' version for transfer, we needed to ascertain the criteria through which we would validate these files. We wanted to ensure we preserved the bitstream and content during the transfer. For this, we decided we would need to check their hash values, ensure metadata remained unchanged, and conduct a visual check of the files. We also wanted to ensure our processes for validating file pre and post transfer were repeatable, and that we could re-import information back into the original cloud environment (SharePoint) to ensure there was no missing context from that information environment.

Generating Hash Values in SharePoint

Typically, to quantitatively validate data we run a hashing algorithm against the original files which generates a hash value that acts as a unique alpha-numeric "fingerprint", any changes to the data would change its "fingerprint". One of the issues we encountered when transferring data was the difficulty in authenticating data in this way against that hosted in a cloud-based environment.

Whilst M365 uses a hashing algorithm, designed by Microsoft, this does not offer the level of data integrity assurance needed. Additionally, this algorithm is not supported by our digital repository which uses DROID for integrity and fixity checking. For our pilot we chose SHA256 – a commonly accepted standard in the digital preservation sector and supported by DROID – as the hashing algorithm to validate our files. However, it is not possible to run executables in the SharePoint environment, necessitating the extraction of SharePoint information 'intact' to allow for this vital validation step to work.

Extracting Information from SharePoint

After deciding our criteria for validating information during the SharePoint transfer process, we ran repeatable tests on two extraction methods to ascertain if our metadata, hash value, and visual check criteria could be successfully fulfilled.

Method 1: Download

Initially we tested the regular process of downloading files from SharePoint onto a local device. We ran DROID reports on them in our staging area and downloaded them a second time from SharePoint but found that their hash values did not match upon subsequent downloads. We also compared their metadata for creation date and last modified against what was shown in the SharePoint environment and found that they also did not match.

Method 2: OneDrive Sync Client

We then tested the OneDrive Sync Client. With administrative access we were able to sync the target site library with a local device. We synced the target site library to a local device, ran a DROID report, desynced and resynced it again. We found that each time we did so the hash values matched. In order to ensure this wasn't because of local caching, we also synced the same site library to different local devices and compared hash values generated by DROID reports across the different devices which also matched. For our metadata validation criteria, we found metadata matched that in SharePoint. The visual check validated identical results for layout and formatting. Our last check was to import the files back into the SharePoint cloud environment and see if they retained their user applied metadata labels, views, and retention instructions. These were imported back in successfully and we could be assured that the files' contextual metadata was saved as part of the file.

Developing a Workflow

After this testing and 'proof of concept', we established an end-to-end transfer workflow. This included the whole process end-to-end from identification of information with the business, to enabling access to that information within the archive. Collaboration between IRMS, Digital Preservation and ROs and IAOs across Parliament is central to this workflow and will remain a focus in future iterations of the workflow. The transfer workflow comprises three main stages: identification and authorisation; transfer, description, and storage; and disposal and access management.

Two pieces of documentation are key to ensuring trust and transparency throughout the process. The Transfer Authorisation Report is an agreement between ROs, IAOs and the Parliamentary Archives marking the start of the process. This then forms the basis of metadata mapping and cataloguing work undertaken by Digital Preservation. At the end of the process, source deletion is recorded here too, providing oversight and transparency for all three parties. Secondly, the Transfer Tracker, which is updated throughout the process. This will prove increasingly important as we scale up SharePoint transfer and allows us to regularly report on progress.

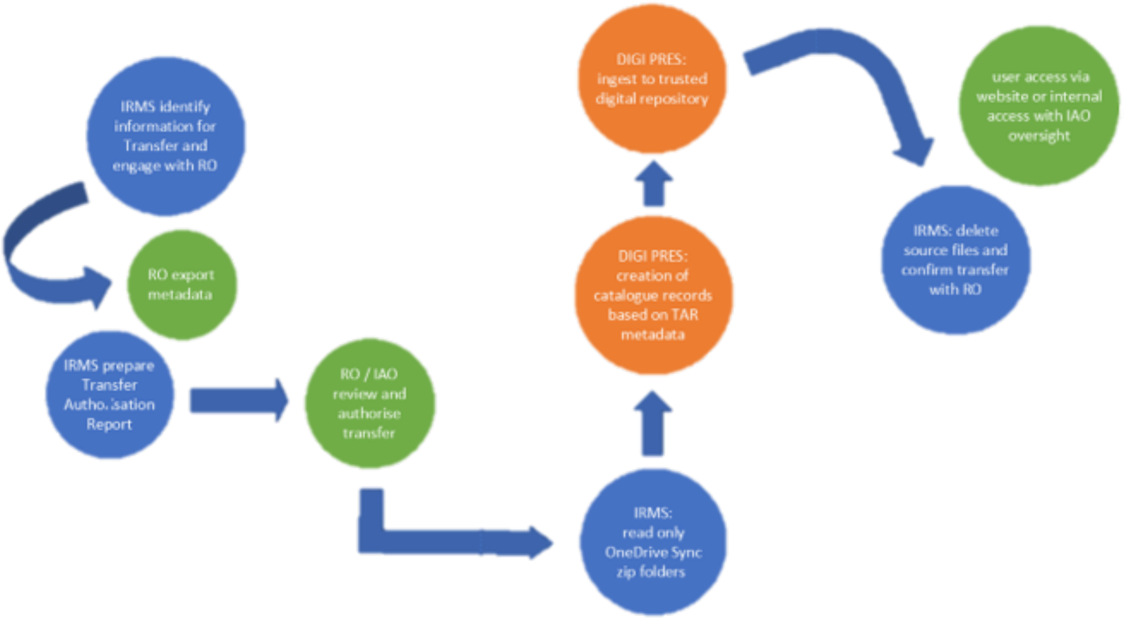
IRMS identify information for Transfer and engage with RO
RO export metadata
IRMS prepare Transfer Autho.:sation Report
RO / IAO review and authorise transfer
IRMS: read only OneDrive Sync zip folders
DIGI PRES: creation of catalogue records based on TAR metadata
DIGI PRES: ingest to trusted digital repository
IRMS: delete source files and confirm transfer with RO
user access via website or internal access with IAO oversight

Figure 1. SharePoint Transfer workflow diagram, showing first iteration of the process.

Piloting the Process

To test our process, we developed a pilot with the aim of focussing more on stakeholder engagement, communication and building out an end-to-end process. We also aimed to test our process on various file formats and a larger number of files. Our test data for this pilot was circa 800 digital items in the 'Control and Disposal' library on the Parliamentary Archives SharePoint site. This information was organised in a logical folder structure with limited user-added metadata. Our key findings from the pilot were that exporting information from SharePoint was much quicker than with the previous EDRMS and that cataloguing and data preparation before ingest, undertaken by Digital Preservation, was the most time-intensive part of the process. This and other areas could be automated in line with resource limitations, which was identified as something to scope and research in future.

Looking Ahead

After the pilot we still have a number of questions to answer. On the technical side, we need to understand how we need to adapt our workflow for information organised with metadata labels, or held in restricted libraries, and undertake further testing on the flexibility and portability of the workflow. We also want to ensure our engagement and messaging is clear in what we want colleagues to 'think', 'feel', and 'do' in line with the SharePoint Disposal project. Finally, scaling up remains a key question, and the resource and processes needed to establish regular transfers from teams across Parliament – exploring automation and a focus on engagement is key here.

For next steps, the focus will be on establishing a formal SharePoint Transfer project for which the testing, workflow and pilot outlined in this brief report will provide the foundation. In the first phase of this project, we aim to formalise our workflows and processes, undertake further piloting on unclaimed data from the decommissioned EDRMS and undertake some discovery work into automation using available tools such as the Microsoft Power Platform, and also into communications and engagement strategies. This work will prepare us to work with 'live data' in phase two of the project, where we hope to engage with Committees in the House of Commons and begin to tackle our questions around metadata labels.

Acknowledgements

The authors would like to thank Adèle Tamar, Head of Preservation and Access (P&A) and Shona Robertson, Head of the Information and Records Management Service (IRMS) for their support of this project from the outset. This work would not have been possible without the efforts of Caitlin Wilson, Digital Archive Assistant, to ingest SharePoint information into our digital repository, and the work of Gemma Lofthouse, Sophie Hurford and other members of IRMS who supported the selection and export information across SharePoint sites at UK Parliament.