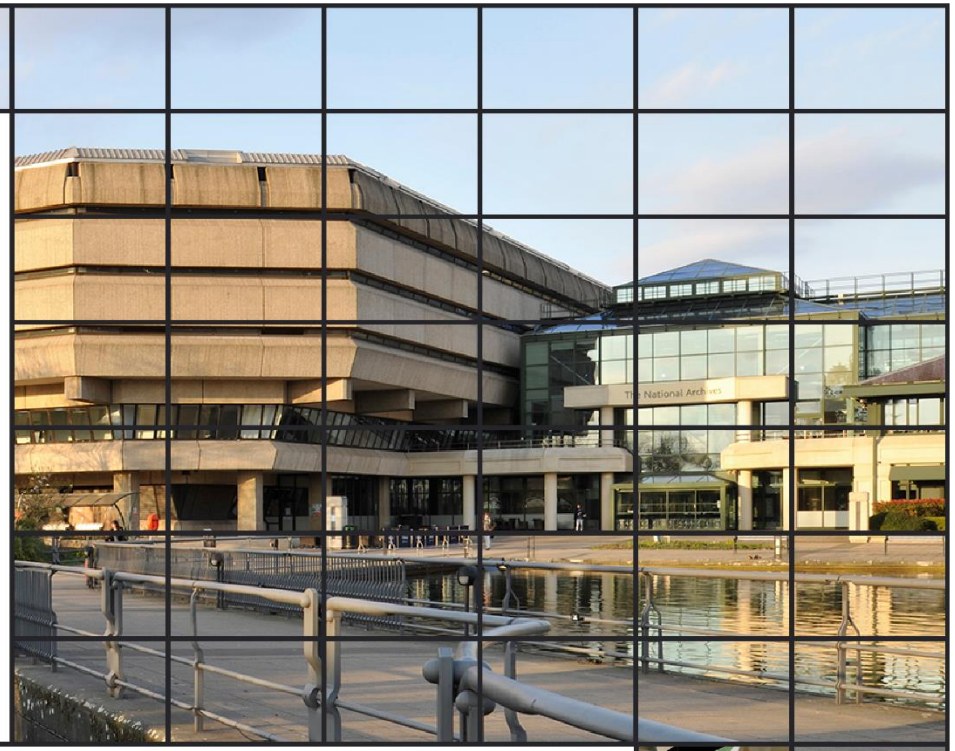
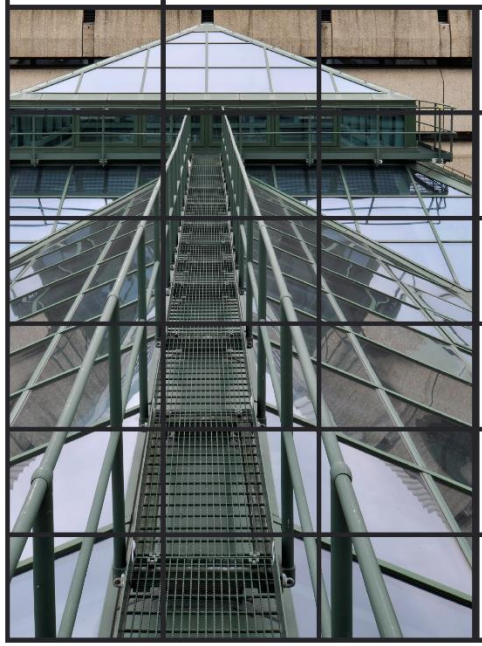
Records in the cloud: The challenge of transferring records from Google Workspace





Paul Young

20 May 2021

THE NATIONAL ARCHIVES

Photo by Mitchell Luo on Unsplash

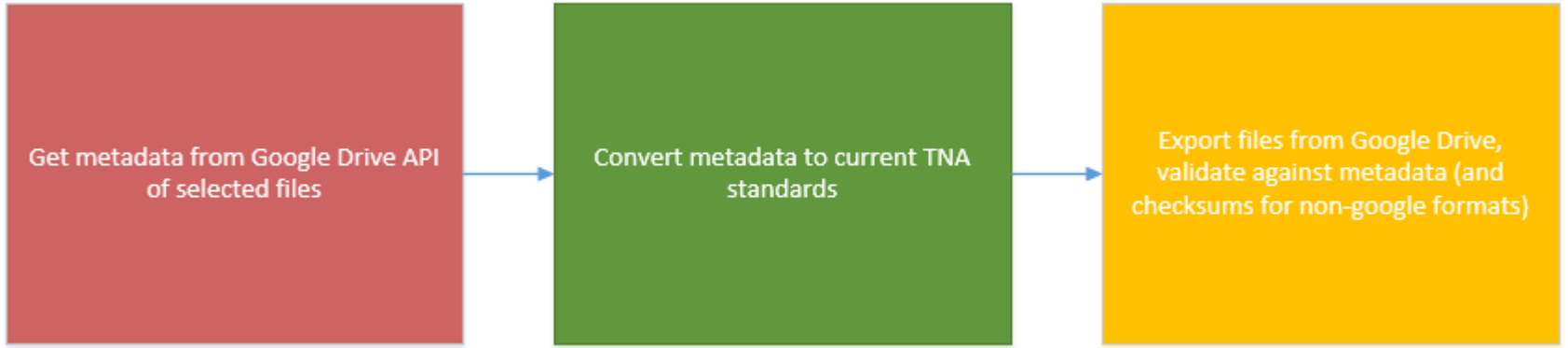
Google Transfer

- Google Workspace. Departments using it as a document management system, with material stored in Google Drive. In some areas similar to EDRMS systems but also comes with new challenges
- How do we get this material to TNA?
- Need to look at methods of export which preserve files and metadata through the transfer process.
- Google Drive holds standard formats (PDF, Docx, JPEG), as well as Google Native Formats (Google Docs, Google Sheets, etc)

THE NATIONAL ARCHIVES

Transfer using Google Drive API

- Test process using Proof of Concept (POC) – GitHub link
- Undertaken to inform development of a cloud to cloud extension of 'Transfer Digital Records'
- Download files and metadata using Google Drive API (including checksum for non-google native files)
- Preserve date and other contextual metadata
- Allows a scalable approach which includes relevant metadata



THE NATIONAL ARCHIVES

Issues

Some issues in common with EDRMS

- Characters not allowed in Windows filesystem **(/<>\?*)**
- Object storage system (different to File systems), arrangement can be reconstructed via Google IDs
- Duplicate filenames are allowed in same folder
- Links between documents are organised via the Google ID, therefore this needs to be captured in order to preserve these links

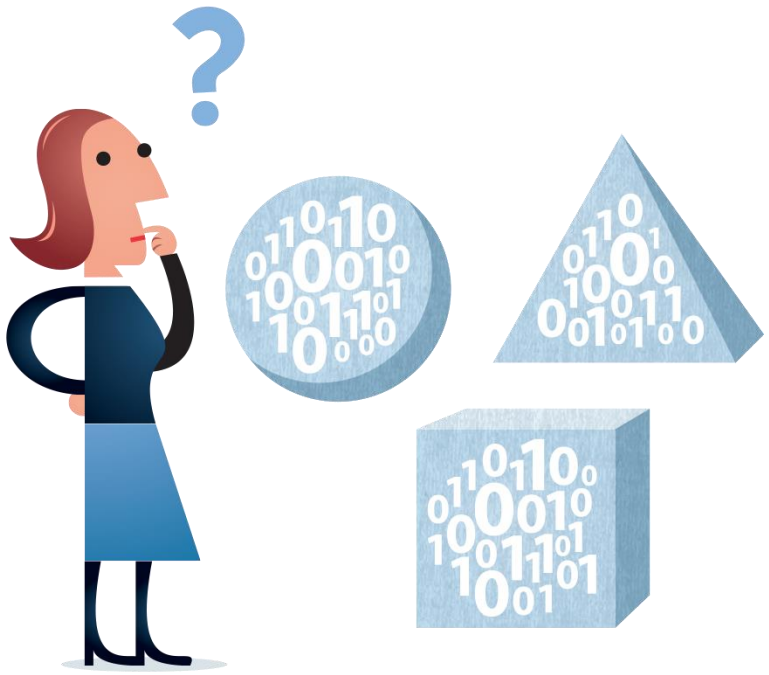
Some new problems

- Third party RM plugins may hold additional metadata to be exported
- **Google Native formats, how to preserve?**

THE NATIONAL ARCHIVES

Google native formats

- Google Docs holds a complete revision log for each file, similar behaviour with other tools
- No checksum! While Google holds an MD5 for non-google native files, there is none for Google native files, as they do not exist as an individual file inside Google

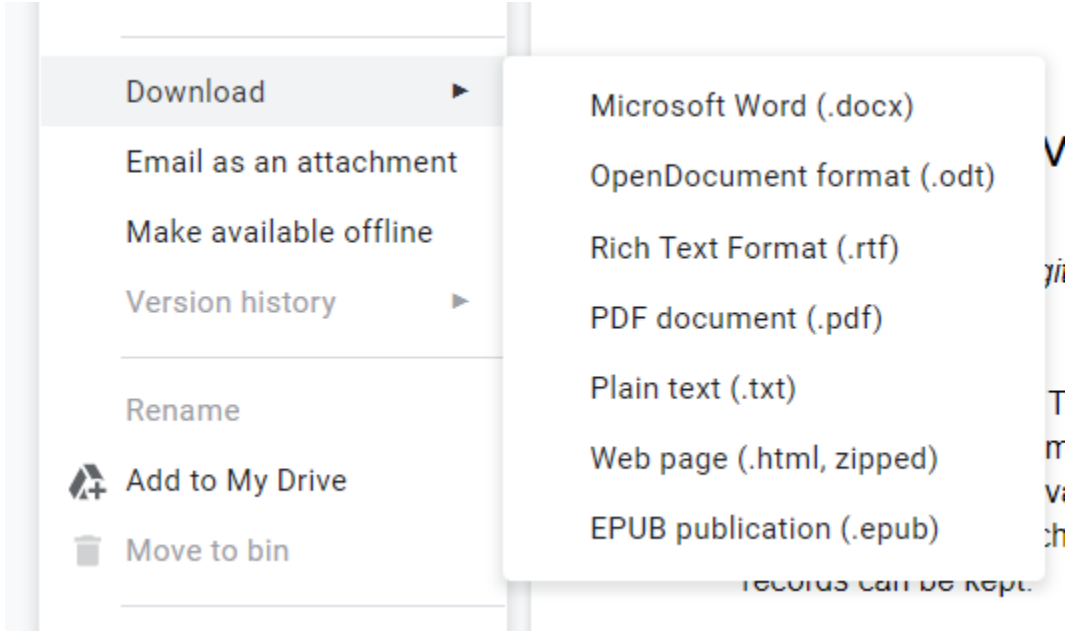


Digitalbevaring.dk

THE NATIONAL ARCHIVES

So what's different?

Google Docs can't be exported as a Google Doc, just as a range of options.



THE NATIONAL ARCHIVES

Export comparison

Created samples of every export option and compared outputs to original format.

Spreadsheet of exports from small sample set – 8 Google Docs, 11 Google Sheets, 2 Google Slides, 2 Google Draw and 2 Google Jamboards.

filename	format	comments	suggestions
15 principles of good service design	Google Docs	yes	yes
15 principles of good service design	PDF	no	no
15 principles of good service design	ODT	yes	yes
15 principles of good service design	txt file	yes	no
15 principles of good service design	rtf	yes	no
15 principles of good service design	docx	yes	yes
15 principles of good service design	html	yes	no
Birthday Wishes	Google Docs	n/a	n/a

THE NATIONAL ARCHIVES

Google – Open Office

- Findings showed that generally the Open Office and Microsoft formats provided the closest functionality and content to the original formats
- With GDS recommending ODF solutions if exporting currently looking at capturing ODF formats for Google native files.
- However some of the more complex files showed where errors could come in – specifically with Google Sheets
  - Google Sheets can use unique formula to Google, this can include API crawls for data or visualisations.
  - For these with ODS and XLSX exports some data or visualisations may be lost on export. Formulas are generally captured but do not work.
  - PDF and HTML formats can capture the data and visualisations but lost some of the functionality of the original format.

THE NATIONAL ARCHIVES

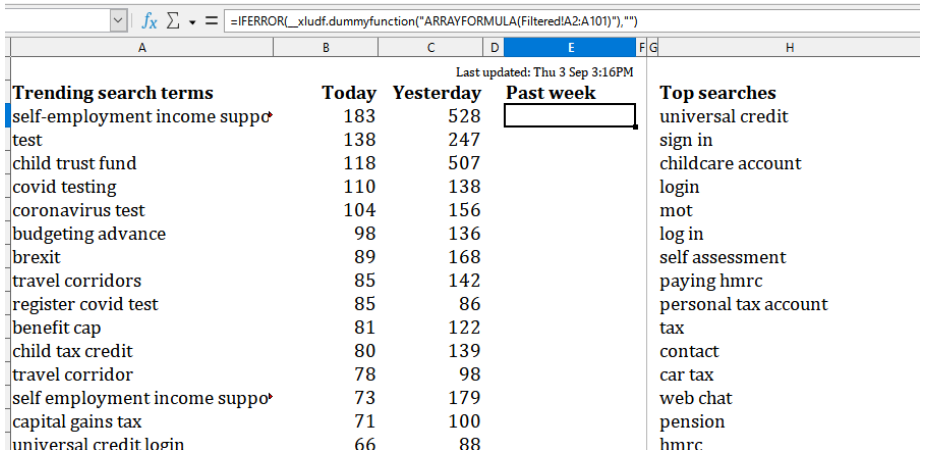
=ARRAYFORMULA(Filtered!A2:A101)

Last updated: Thu 3 Sep 3:16pm

Trending search terms	Today	Yesterday	Past week	Top searches
self-employment income supp	183	528		universal credit
test	138	247		sign in
child trust fund	118	507		childcare account
covid testing	110	138		login
coronavirus test	104	156		mot
budgeting advance	98	136		log in
brexit	89	168		self assessment
travel corridors	85	142		paying hmrc

Dashboard | For Data Studio | Filtered | Combined | Report

THE NATIONAL ARCHIVES



ODS

Last updated: Thu 3 Sep 3:16pm

Trending search terms	Today	Yesterday	Past week	Top searches
self-employment income supp	183	528		universal credit
test	138	247		sign in
child trust fund	118	507		childcare account
covid testing	110	138		login
coronavirus test	104	156		mot
budgeting advance	98	136		log in
brexit	89	168		self assessment
travel corridors	85	142		paying hmrc
register covid test	85	86		personal tax account
benefit cap	81	122		tax
child tax credit	80	139		contact
travel corridor	78	98		car tax
self employment income supp	73	179		web chat
capital gains tax	71	100		pension

HTML

- If using Google specific formula ODS or XLSX may not capture full data or visualisations.
- For Google Sheets a mixture of ODS and HTML exports to ensure greatest capture possible.

THE NATIONAL ARCHIVES

Additional thoughts:

- The complexity of the format will influence how effective the export is.
- When exporting files outside of Google date metadata may be lost. API crawls of dates and other metadata prior to export ensure the most accurate date metadata is captured.
- Web archiving techniques could also be a preservation option, or used in conjunction with an export format. This could preserve some original functionality. Current issues with scalability of this approach.
- Likely to occur in other cloud systems not just Google.
- Things change! These systems can be removed or edited at any point. See - https://workspaceupdates.googleblog.com/2021/05/Google-Docs-Canvas-Based-Rendering-Update.html

THE NATIONAL ARCHIVES

DPC Focus group

- Published a blog on the Digital Preservation Coalition, currently looking for volunteers to join a focus group to collect characterstics of Google Docs which the digital preservation and archival community consider vital for preservation.
  - What features of a native Google format do you need to capture?
  - What features of a native Google format do you believe future users will need to access?

  We can then share these findings with Google.
- If interested in joining a focus group please contact myself or Jenny Mitcham

paul.young@nationalarchives.gov.uk jenny.mitcham@dpconline.org

THE NATIONAL ARCHIVES