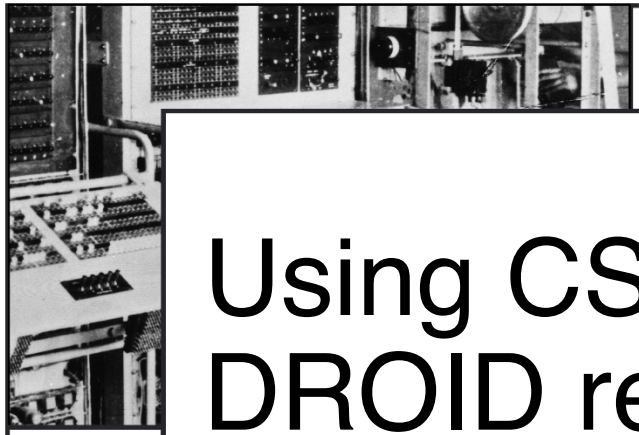
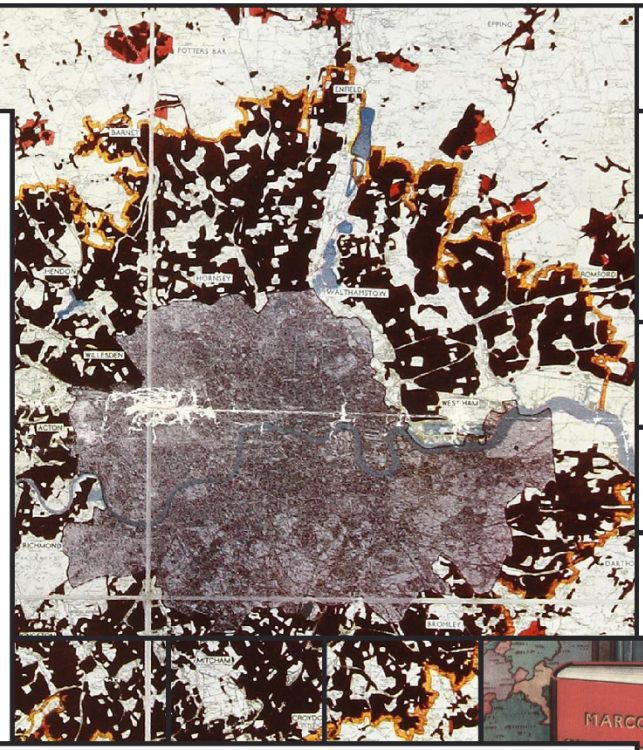
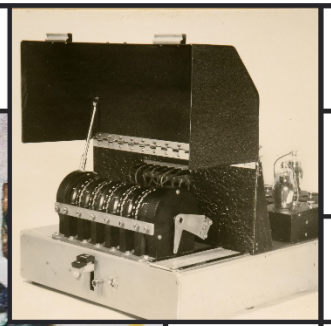


Using CSV Schema and DROID reports for digital preservation

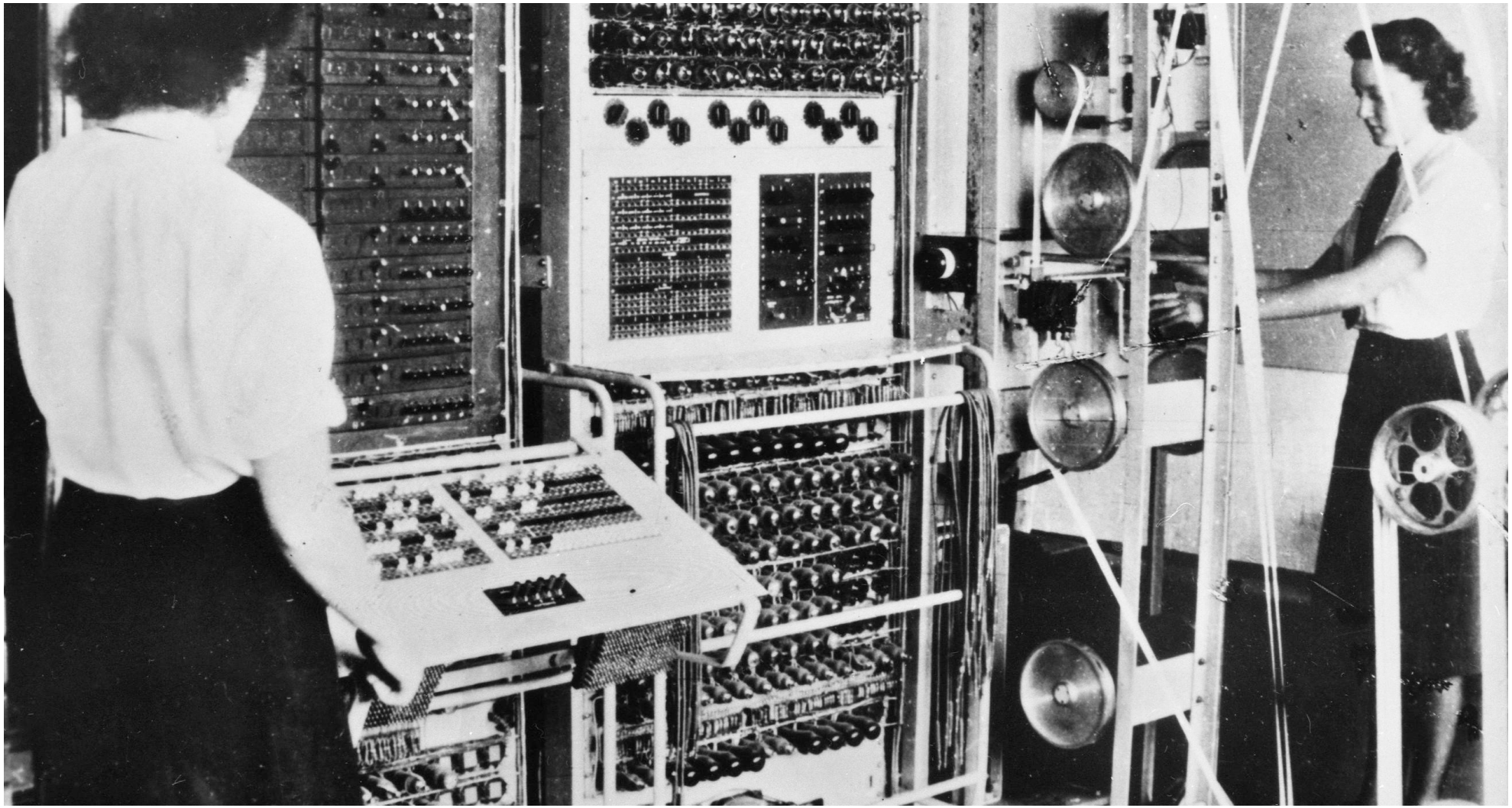




David Underdown

21 January 2020

THE NATIONAL ARCHIVES



Colossus electronic digital computer

THE NATIONAL ARCHIVES

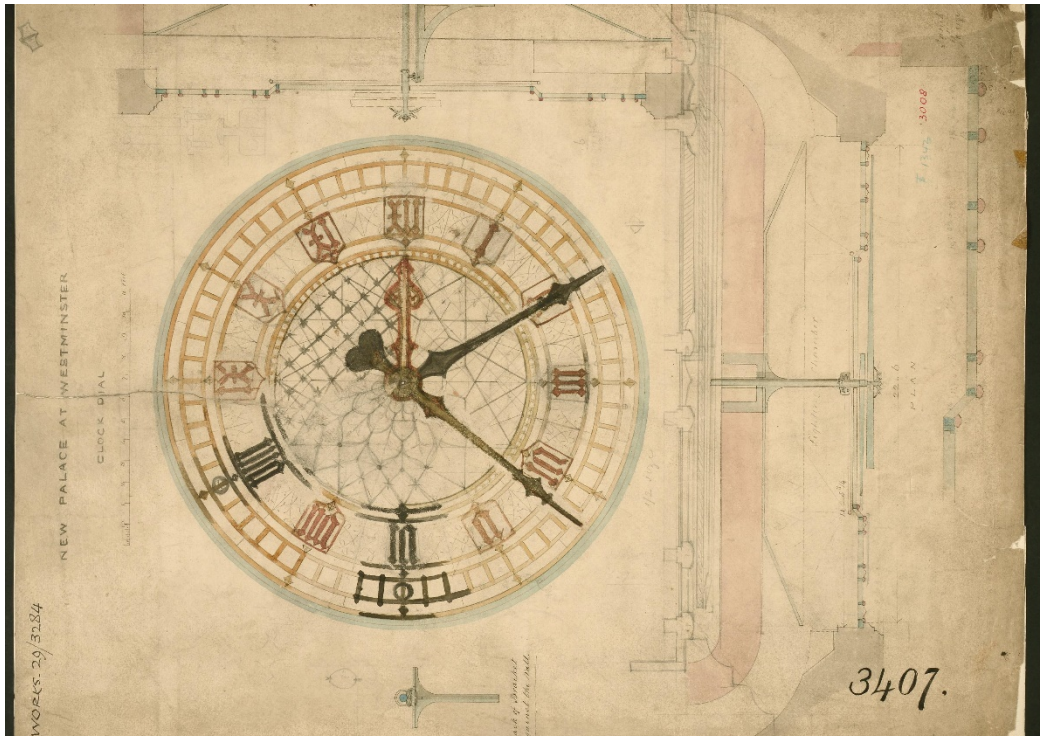
Fixity and integrity checking

- We want to demonstrate over time that file content is fixed (unchanged)
- Part of the provenance of a digital file
  - We can demonstrate that the file is the same thing that we received
  - Contributes to authenticity of the file and trust in the archive
- Typically use a cryptographic hash function (CHF) to generate a checksum (also known as a hash or message digest) eg
  - MD5
  - SHA-1
  - SHA-2
    - SHA-256
    - SHA-512

THE NATIONAL ARCHIVES

One-way functions and hash collisions

- CHFs (and digital encryption more generally) rely on one-way (trapdoor) functions
- For example, it's easy to multiple two numbers together, but to retrieve the original factors is time consuming.
- Only a finite (though large) set of values for each type of checksum, so always a risk of "collision", two different files producing same checksum

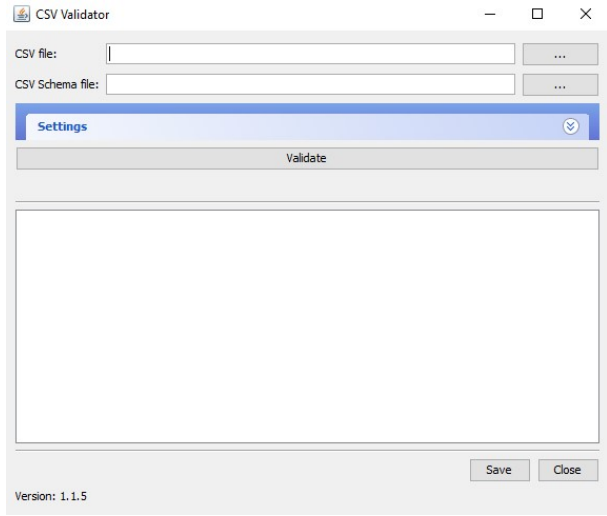


Elevation for clock dial for Big Ben tower

THE NATIONAL ARCHIVES

CSV Validator and CSV Schema Language

- Developed by The National Archives for handling archival metadata in CSV format
- Result of our experiences with inconsistent metadata in Home Guard digitisation pilot
- Can specify that a column is date/time (various formats) and falls in a particular range, numeric, matches a regular expression
- Also conditionals, so appropriate check can depend on the value of another column
- Check a file exists and checksum matches value
- Check that no files present within folders that are **not** listed in your metadata file (integrityCheck)



CSV Validator

EXAMPLE 21

```
a_column: uri   //the value of a_column must be a valid URI
```

An example from the CSV Schema Language, declaring that a column called "a_column" must contain a valid URI. See documentation

THE NATIONAL ARCHIVES

Working with DROID exports

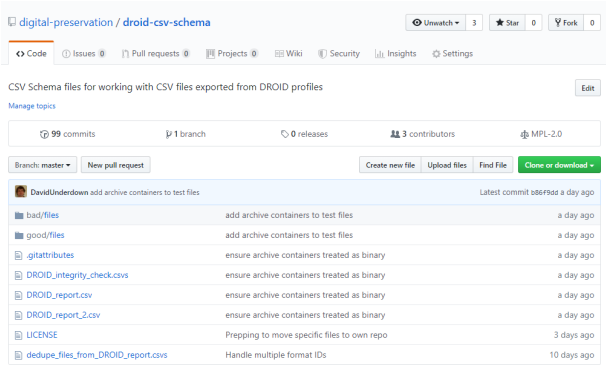
- As Rachel has already demonstrated, having run DROID over a collection of files, you can then export to CSV to do further analysis
- When checking for duplicate files, we were not using the feature of verifying checksums against the stored files, so it was desirable to use the DROID feature that can read files inside ZIPs and other “archive” formats
- Unfortunately, for the use case of integrity checking this feature causes problems, as CSV Validator cannot read inside “archives” so having rows for such files in the export causes problems with the integrityCheck feature which allows us to verify that there are no additional, undocumented, files within collections
- Two choices:
  - New DROID run not looking in “archives”
  - Copy existing DROID export and remove these rows

THE NATIONAL ARCHIVES

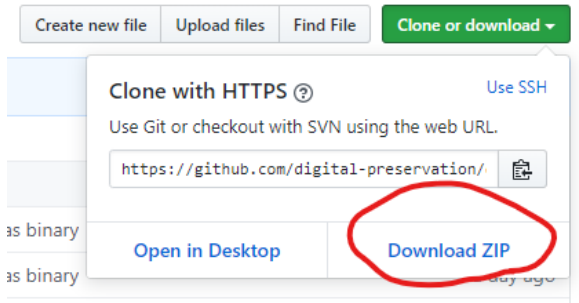
Example files and CSV Schemas

All available at https://github.com/digital-preservation/droid-csv-schema

- Example DROID export CSVs
- CSV Schema
- “Good” and “Bad” file sets
- Just click on green “Clone or download” button and then “Download ZIP”



droid-csv-schema repo



Click the green button “Download zip”

THE NATIONAL ARCHIVES

DROID integrity check schema

version 1.1

//several columns omitted here for clarity

URI: fileExists integrityCheck("","files","includeFolder")

//replace "files" in integrityCheck with name of top level folder within your DROID report (in quotes)

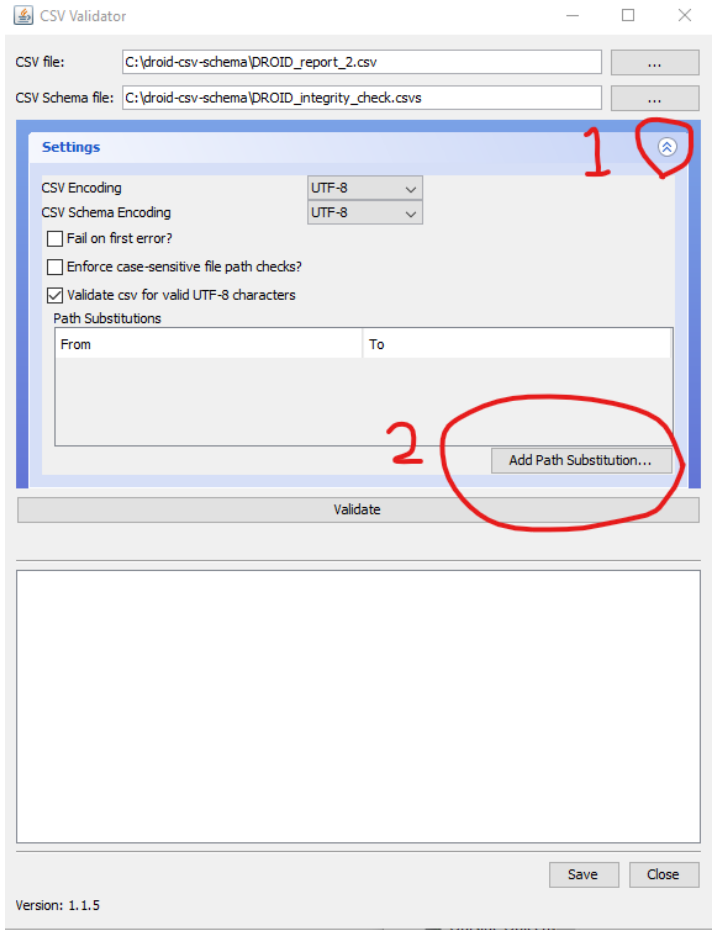
SHA256_HASH: if($URI/ends("/"),empty,if($URI/starts("file:"),checksum(file($URI),"SHA-256"),notEmpty))

//folders do not have a checksum

THE NATIONAL ARCHIVES

Running the integrity check

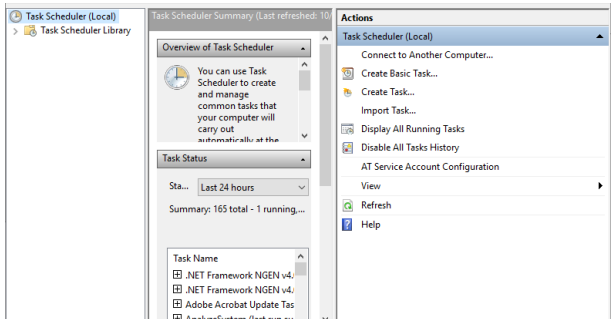
- DROID report probably not created where the files will be stored
- Use “path substitution” to allow us to replace parts of the path as recorded in the DROID report so that CSV Validator can still find the files, for example:
  - Originally C:\files\
  - Now S:\digital archive\files\
- Represented in DROID CSV as URI, so:
  - from file:/C:/files/
  - to file:/S:/digital%20archive/files/



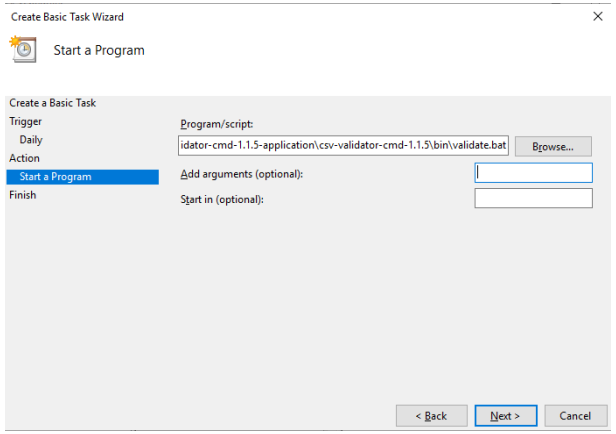
THE NATIONAL ARCHIVES

Automating the integrity check

- Need to use the command line CSV Validator
- Windows Task Scheduler
  - Create Basic Task…
  - Choose a frequency and when you want it to run, eg 1st day of every month, 10pm
  - “Start a program”
  - Program/script: C:\Users\dunderdown\csv-validator-cmd-1.1.5-application\csv-validator-cmd-1.1.5\bin\validate.bat
  - Add arguments (optional): -p file:/C:/=file:/C:/droid-csv-schema/good/ C:/droid-csv-schema/DROID_report_2.csv C:/droid-csv-schema/DROID_integrity_check.csvs > C:\reports\integrity_check.out



Initial view of Task Scheduler



Screen for inputting “Start a program” parameters

THE NATIONAL ARCHIVES

Automation considerations

- Calculating a checksum means reading the whole file
- So runtime is largely dependent on the volume of data you check and the read speed of your storage
  - On cloud storage you could also rack up data access costs
- Think hard about how often you want to check
  - What's the impact on your network, or other people trying to read or write files?
- Check part of the collection at a time (ideally you'll have a DROID report for each accession, so you can group things together)
- For a less intensive check you could just check fileExists, plus the integrityCheck, and schedule this more frequently
  - This only has to interrogate the file system metadata rather than reading back the files themselves

Other possible validation checks

- CSV Schema could also be used to implement a format blacklist for material that you don't want in your archive
    - To exclude thumbs.db (Windows) and .ds_store (Mac) files

NAME: if(is("thumbs.db"),$PUID/not("fmt/111")) @ignoreCase

PUID: not("fmt/682") and not("fmt/394")

- https://github.com/digital-preservation/csv-schema/tree/master/example-schemas has various other examples of schemas we have used at The National Archives

THE NATIONAL ARCHIVES

Questions and follow up

- Time for questions now
- Will share slides
- Comment on the original blog post https://openpreservation.org/blog/2019/05/28/droid-report-as-basis-for-collection-integrity-checks/
- Twitter: @davidunderdown9
- Email: digitalpreservation@nationalarchives.gov.uk
- GitHub: https://github.com/digital-preservation/droid-csv-schema
  https://github.com/digital-preservation/csv-validator
  https://github.com/digital-preservation/csv-schema

THE NATIONAL ARCHIVES