University of Sussex Library

Web Capture Workflows

Duncan Harrison (Research Data & Digital Preservation Officer)
duncan.harrison@sussex.ac.uk

UNIVERSITY OF SUSSEX

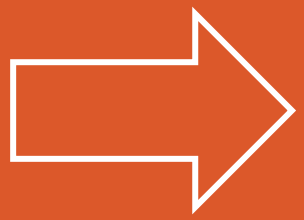
Web archive activities at UoS Library

- Bulletins and newsletters
- Project specific websites
- Sections of the main website
- Harvesting of documents hosted online
- Teaching and advocacy

UNIVERSITY OF SUSSEX

Who?

- Researchers
- Academics
- University depts.
- Project leads
- Library staff
- Internal comms



- Library Special Collections archivist

- Library Digital Development & Systems team

UNIVERSITY OF SUSSEX

Why?

- Content has moved from paper to digital
- Projects end and their websites need to be preserved as outputs
- Content is migrated or removed
- Record rapid response
- Actively preserving institutional memory

UNIVERSITY OF SUSSEX

How?

- Research into methods & standards
- Testing open-source tools
- Engagement with community of practice
- Lots and LOTS of errors, failure and going back to the drawing board!

UNIVERSITY OF SUSSEX

Summary of tools

Conifer by Rhizome.
Manual capture tool.

Archiveweb.page by webrecorder project.
Manual capture tool.

GNU Wget. Command line web crawler.

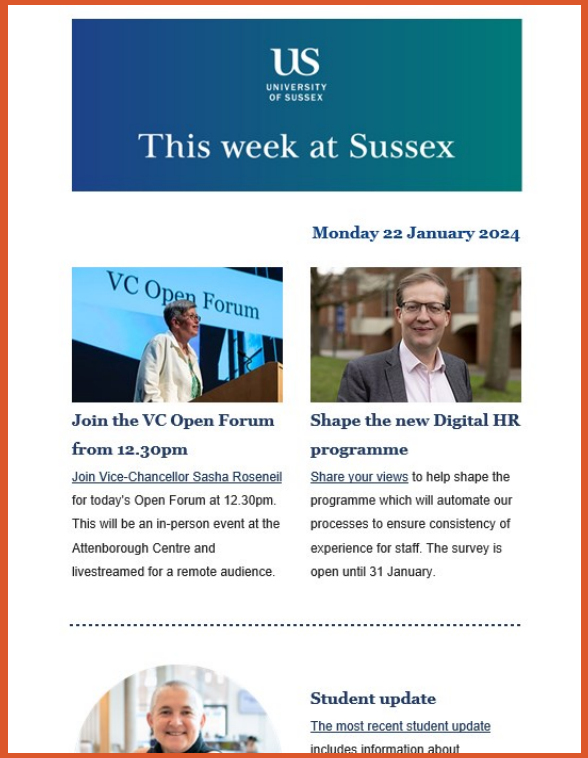
Google Drive. Can host html files. University managed accounts.

Drive to web (www.drv.tw)
Publishes html from Google Drive & OneDrive.

Web Archive file (WARC).
Contains linked assets, instructions & metadata which comprise websites.
Can be 'played' to replicate original browsing experience.

UNIVERSITY OF SUSSEX

Bulletins & Newsletters



US
UNIVERSITY OF SUSSEX

This week at Sussex

Monday 22 January 2024

Join the VC Open Forum from 12.30pm

Join Vice-Chancellor Sasha Roseneil for today's Open Forum at 12.30pm. This will be an in-person event at the Attenborough Centre and livestreamed for a remote audience.

Shape the new Digital HR programme

Share your views to help shape the programme which will automate our processes to ensure consistency of experience for staff. The survey is open until 31 January.

Student update

The most recent student update includes information about

Chris van Tulleken: how ultra-processed food took over the world

Jubilee Lecture Theatre | Thursday 25 January | 12.30pm-1.15pm

Hear from Chris, an award-winning broadcaster, practicing NHS doctor and leading academic, who will talk about ultra-processed food and its impact on our health and weight.

Building trust and psychological safety in teams

Library Meeting Room 1 | Tuesday 23 January | 10am-11.45am

Book now for this workshop giving an introduction to psychological safety and exploring how to create a safe and supportive working environment that fosters trust, inclusion, collaboration and wellbeing within teams.

Shape how we listen at Sussex

Online | Tuesday 23 January | 10.30am-11.30am

Join the final focus group aiming to identify the best ways for staff to share feedback about working at Sussex between our staff surveys.

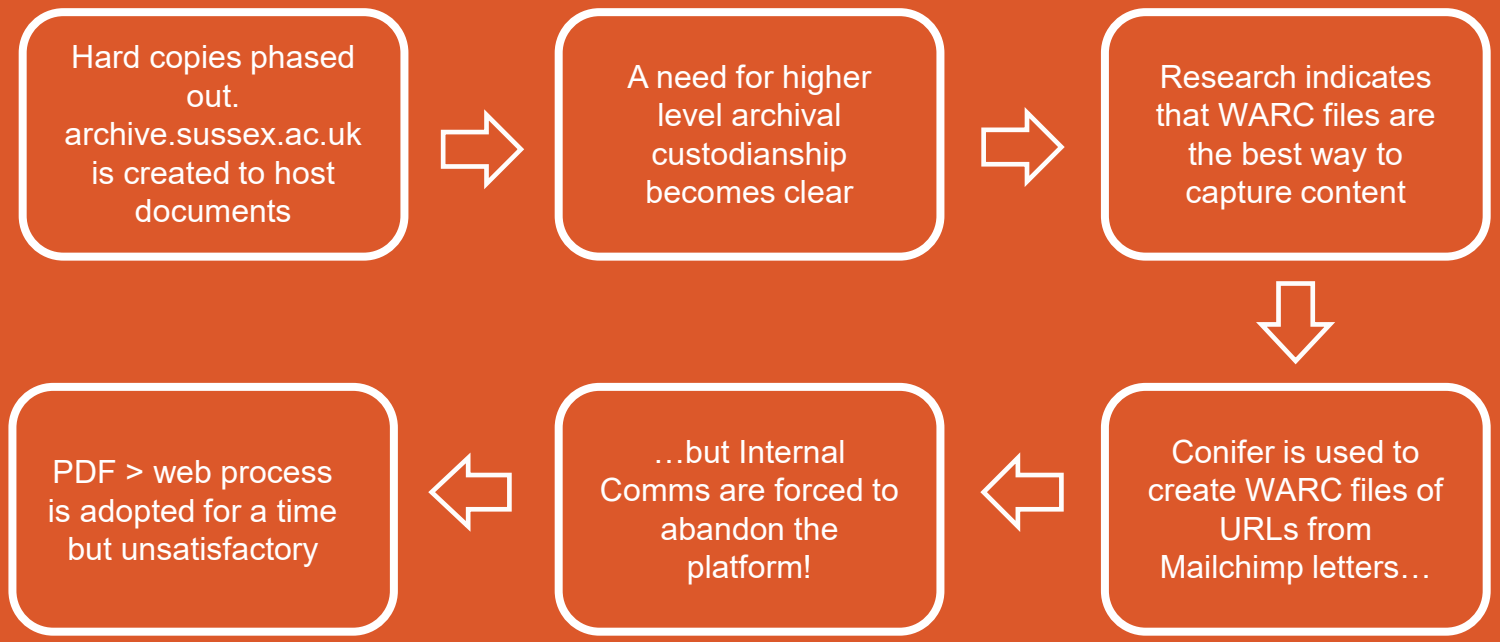
Staff apprenticeships drop-in

Online | Wednesday 24 January | 10am-12 noon

Drop in and chat with Chris Hamilton, Staff Apprenticeships Officer, about how an apprenticeship could help you grow your career.

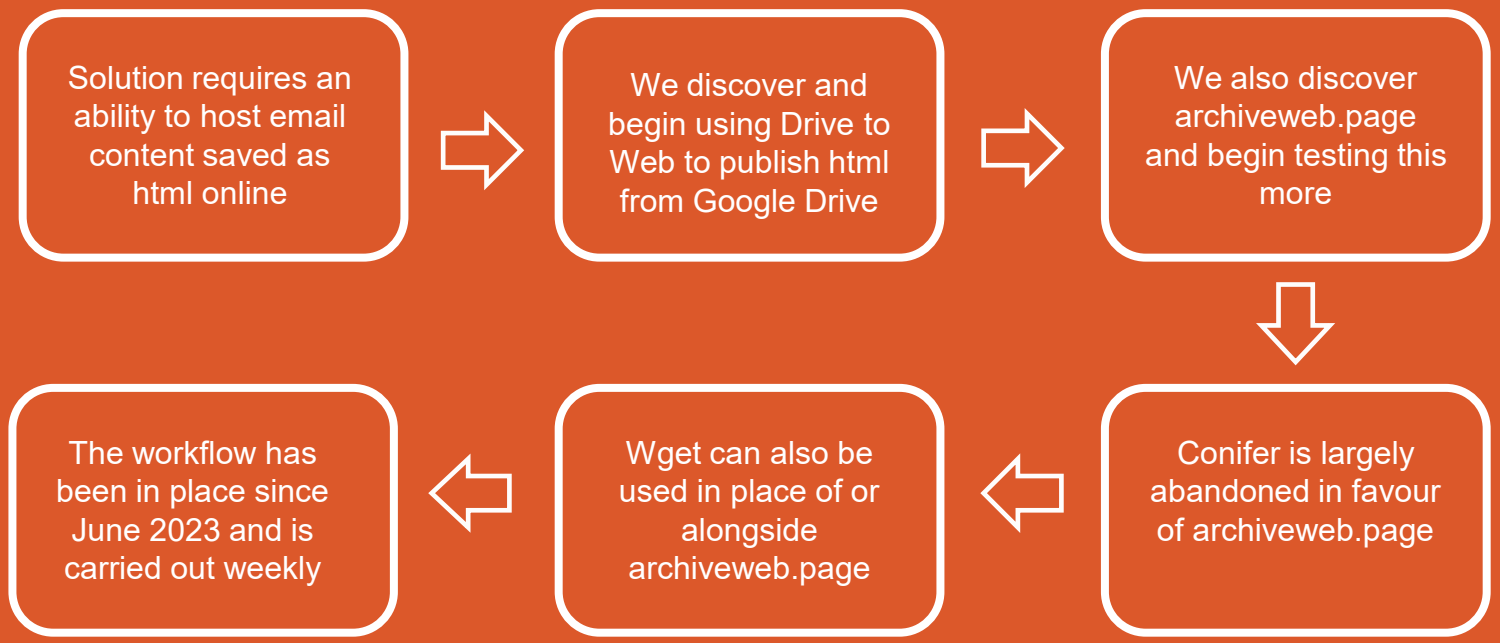
US
UNIVERSITY
OF SUSSEX

Development of processes



UNIVERSITY OF SUSSEX

Development of processes (cont.)



UNIVERSITY OF SUSSEX

Bulletin capture

Email bulletins are saved as html.

Folder of html contents is uploaded to Google Drive

Drive To Web finds and publishes html content in Google Drive

Making the bulletin content available as a website with URL

Captured using free tools then saved as WARC/WACZ

UNIVERSITY OF SUSSEX

Larger website capture

- Multiple pages
- Complex architecture
- Manual crawl would be time consuming and prone to error
- We may also want to harvest individual assets hosted within the website (documents, audio, images etc) after the crawl

UNIVERSITY OF SUSSEX

Larger website capture

- Define depth (i.e. # of clicks)
- Restrict to domain (i.e. sussex.ac.uk)
- Restrict parents (i.e. crawl no higher than sussex.ac.uk/example-page
- Create directories for files
- Define what will be saved/deleted

UNIVERSITY OF SUSSEX

Larger website capture

wget “sussex.ac.uk/library/blog”
--mirror (-r –l3)
--domains sussex.ac.uk
--no-parent
--page-requisites
--span-hosts
--convert-links
--warc-file=“Library Blog 2024”

All assets & WARC file

UNIVERSITY OF SUSSEX

Storage, cataloguing and ingest

- WARC and/or WACZ files are stored in institutional Box locations accessible by key staff
- Catalogue records are updated via CALM
- Items are named via a fixed schema containing reference number, date of publication and date of capture.
- MD5 checksum files updated following each new file
- Collections are ingested on a per year basis

UNIVERSITY OF SUSSEX

Pros & cons

- ✓ The workflow is flexible and can be used for large and small projects.
- ✓ The workflow results in outputs that accord with archival standards.
- ✓ Processes can be interoperative.
- ✓ We are able to train staff and researchers in these processes.

- × Reliant upon third party tools and services.
- × Online content changes rapidly. Success is never guaranteed.
- × Following process is simple, troubleshooting failure is not.

UNIVERSITY OF SUSSEX

The future

- Continued engagement with web archiving community
- Trial new tools
- Explore possibilities of AI and automation
- Promote the archives among students and researchers
- Continue to develop teaching offer

UNIVERSITY
OF SUSSEX

Thank you!

Questions? Clarifications? Advice?

Duncan.Harrison@sussex.ac.uk

UNIVERSITY
OF SUSSEX