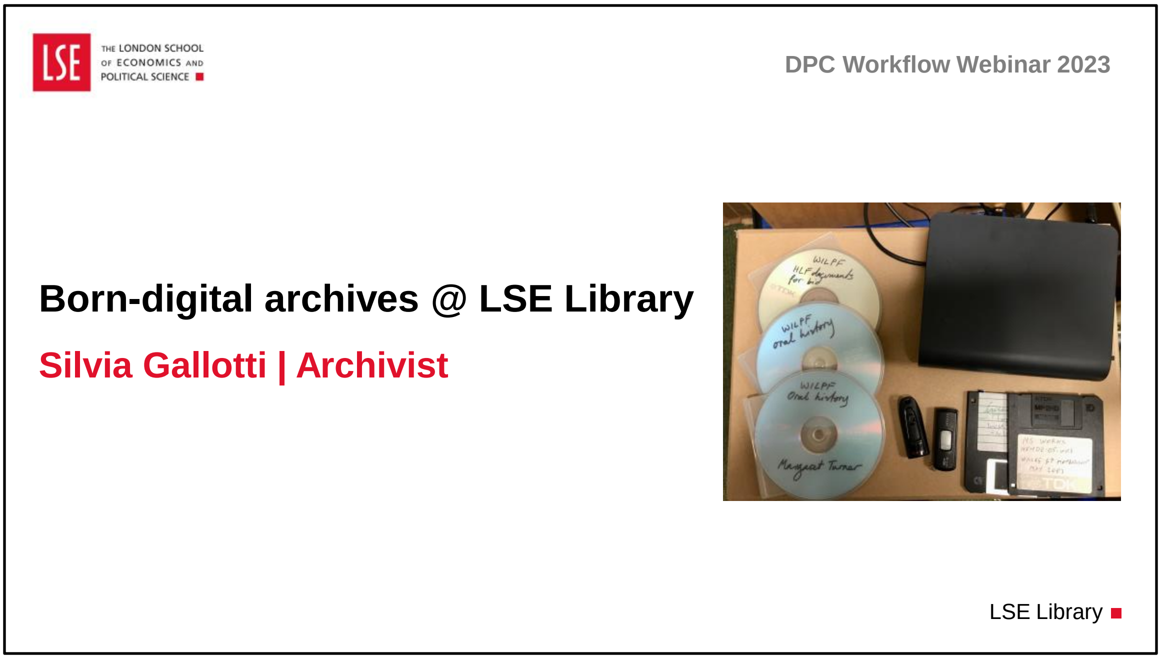


THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DPC Workflow Webinar 2023

Born-digital archives @ LSE Library

Silvia Gallotti | Archivist

LSE Library

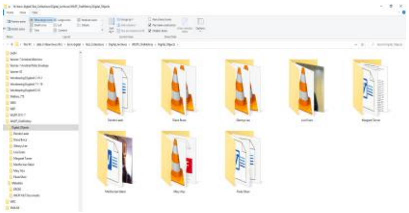
Today I am going to talk to you about our workflow for managing born-digital archives, from acquisition to discoverability.

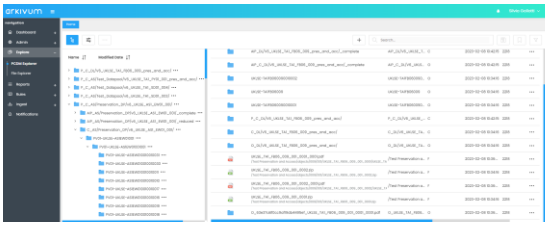
Let’s start with a bit of background to understand how the workflow came about.

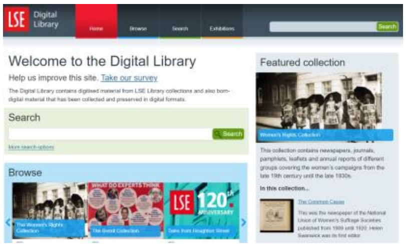
We hold the historical records of LSE itself, donated archives relating to modern British political and economic history and the development of social science, and the Women’s Library collection. And we have been receiving born-digital records for about 15 years.

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DPC Workflow Webinar 2023







Welcome to the Digital Library

Help us improve this site. Take our survey

The Digital Library contains digitised material from LSE Library collections and also born-digital material that has been collected and preserved in digital formats.

Search

More search options

Browse

The Women's Rights Collection

The Brexit Collection

Tales from Houghton Street

Featured collection

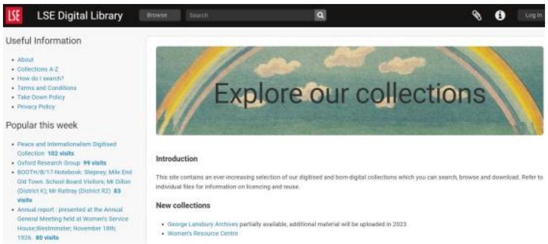
Women's Rights Collection

This collection contains newspapers, journals, pamphlets, leaflets and annual reports of different groups covering the women's campaigns from the late 19th century until the late 1930s.

In this collection...

The Common Cause

This was the newspaper of the National Union of Women's Suffrage Societies published from 1909 until 1920. Helen Swanwick was its first editor.



Useful Information

- About
- Collections A-Z
- How do I search?
- Terms and Conditions
- Take Down Policy
- Privacy Policy

Popular this week

- Peace and Internationalism Digitised Collection 102 visits
- Oxford Research Group 99 visits
- BOOTH/B/17-Notebook: Stepney; Mile End Old Town. School Board Visitors; Mr Dillon (District K); Mr Rattray (District R2) 83 visits
- Annual report : presented at the Annual General Meeting held at Women's Service House,Westminster; November 18th; 1926. 80 visits

Introduction

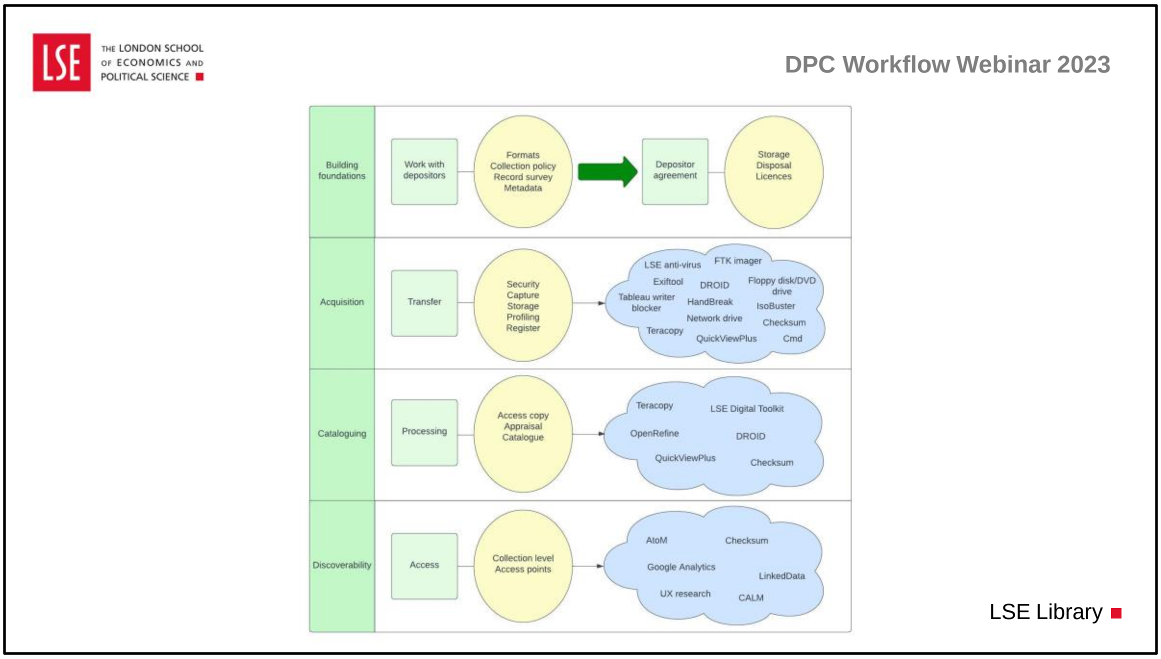
This site contains an ever-increasing selection of our digitised and born-digital collections which you can search, browse and download. Refer to individual files for information on licencing and reuse.

New collections

- George Lansbury Archives partially available, additional material will be uploaded in 2023
- Women's Resource Centre

LSE Library

First steps towards digital preservation were taken back in about 2010. An archivist was appointed with responsibility for born-digital material and a Digital Library was developed using a Fedora repository. All legacy born-digital collections were transferred from their original media onto our network drives and profiled and a digital asset register was created. But then it all stopped because of lack of resources, mainly the loss of said archivist and IT support within the Library. This meant that full-scale preservation was never developed and in 2018 a proprietary system was purchased, Arkivum. Since then we have been creating workflows for ingesting digital assets, both digitised and born-digital, into the system and we are gradually transferring legacy collections from our network drive to Arkivum and from legacy DL to new DL, powered by AtoM.



THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DPC Workflow Webinar 2023

Building foundations

Work with depositors

Formats
Collection policy
Record survey
Metadata

Depositor agreement

Storage
Disposal
Licences

Acquisition

Transfer

Security
Capture
Storage
Profiling
Register

LSE anti-virus
FTK imager
Exiftool
DROID
Floppy disk/DVD drive
Tableau writer blocker
HandBreak
IsoBuster
Network drive
Checksum
Teracopy
QuickViewPlus
Cmd

Cataloguing

Processing

Access copy
Appraisal
Catalogue

Teracopy
LSE Digital Toolkit
OpenRefine
DROID
QuickViewPlus
Checksum

Discoverability

Access

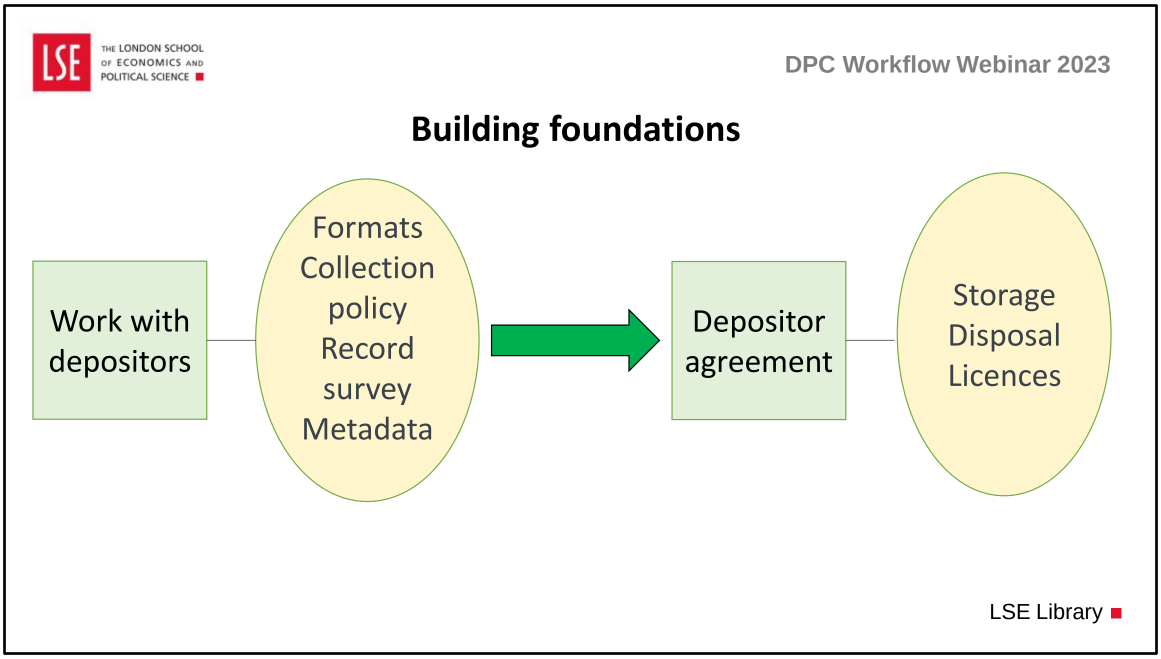
Collection level
Access points

AtoM
Checksum
Google Analytics
LinkedData
UX research
CALM

LSE Library

And this is the workflow we have developed for born-digital archives. The problem was that we had some bits of the workflow already, particularly in phase 2. But we were missing the other 3 phases. So, let’s have a look at them all in more details.

Before I start, I just wanted to mention quickly that this workflow is deliberately general as it needs to be flexible and needs to be adapted to each collection. And this is why I used the blue clouds, mixing in all steps and tools, to give the idea that they can happen at different stages, in a different order, and that specific solutions and workarounds often need to be found for different collections. We could almost build a library of case studies. As we go forward, we will develop more detailed workflows for each phase, possibly also tailored to the type of material to process.



DPC Workflow Webinar 2023

Building foundations

Work with depositors

Formats
Collection policy
Record survey
Metadata

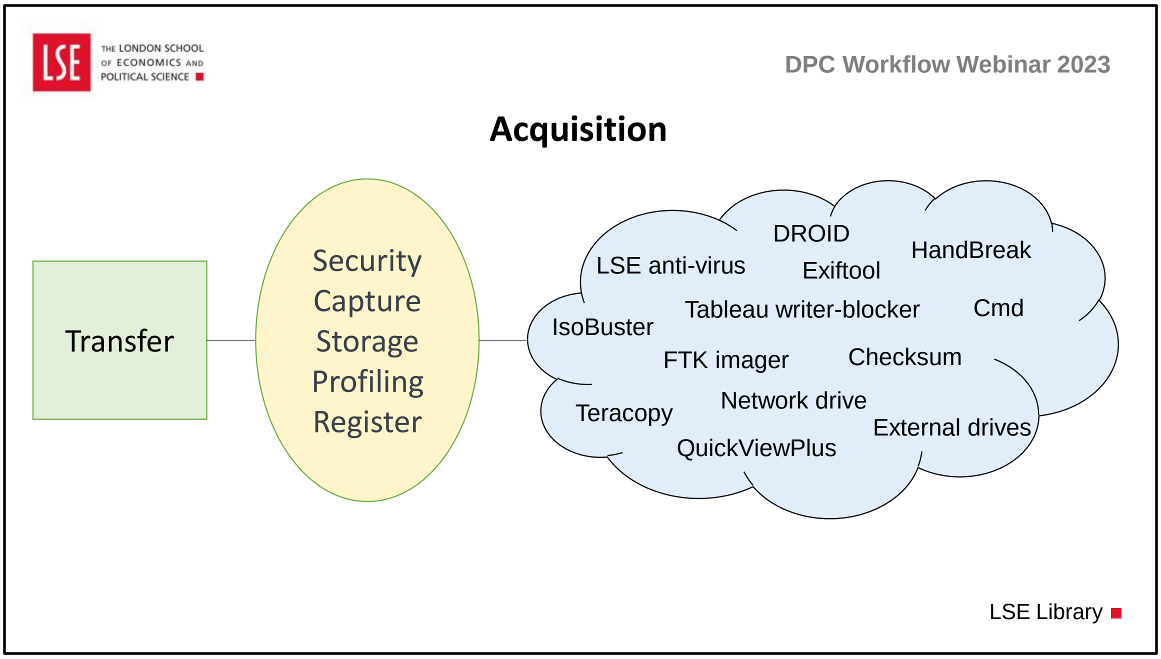
Depositor agreement

Storage
Disposal
Licences

LSE Library

We call the first phase ‘Building foundations’ as the stronger this is, the more stable our workflow will be and the more work we put at this stage, the more we will benefit from it later.
It essentially means working with depositors. In the past we acquired born-digital material almost accidentally, inside a box within the paper archive. Now we want to avoid that as much as possible so we talk to depositors about born-digital when they first approach us about their archive, we ask questions about the material, formats, quantity, where they are stored, names of files, sensitivity issues (we have a growing list of questions, and again every collections will need different questions to be asked). We share with them our guidelines about file naming conventions and accepted formats. We ask depositors to be an active part in the selection process so we share our collection policy where we clearly state the type of material we want and don’t want - if existing depositors, we already know from the paper archive what they produce but there could be new content, e.g. social media). We do a record survey (e.g. can get access to Google drives), and we ask them to ideally provide metadata (we have templates for this). Whether all this will be possible or not really depends on who the depositor is and what resources they have. But we can aim at doing this and as they say, every little helps!

We then sign the depositor agreement, which includes provisions for digital material (it explains what we do with them, where they are stored, what we won’t keep) and details on licences (very important for access purposes).



THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DPC Workflow Webinar 2023

Acquisition

Transfer

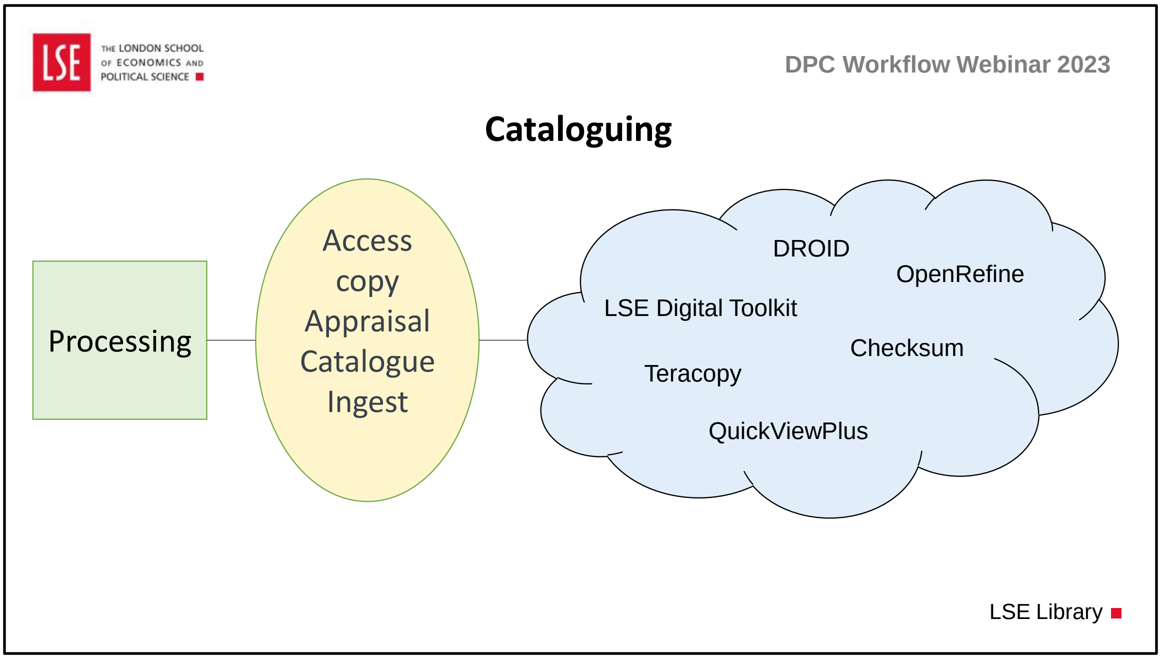
Security
Capture
Storage
Profiling
Register

LSE anti-virus
DROID
Exiftool
HandBreak
Tableau writer-blocker
Cmd
IsoBuster
FTK imager
Checksum
Teracopy
Network drive
External drives
QuickViewPlus

LSE Library

The second phase is the acquisition phase, when we get the material. This is where the workflow really needs to be flexible and changes every time with different collections. When depositors have been an active part in the previous stage, collections are transferred and catalogued and very swiftly ingested. We know the content because we've done work with depositors, so for example it might not be necessary to use forensic tools. When depositors haven't been an active part, and this is the case with most of our legacy collections, we may need to use tools to extract the content, for example FTK imager. Other collections need to be investigated more before we even transfer them onto our drive. With a recent collection I worked on, I used tools such as QuickViewPlus , IsoBuster, HandBreak to investigate content on media pre-transfer and to see hidden material, and my colleagues and I appraised the material at this stage – this is because we knew we wouldn't have wanted to keep most of it and I didn't think it would be worth spending a lot of time transferring it to our share drive – it would have taken weeks to do it.

We initially save things on LSE network drive – DTS have asked us to investigate SharePoint but this is not a possibility as it changes the digital objects somehow. We use this sometimes for initial appraisal using access copies of the material. We use external drives and writer blocker for material on portable media, or FTP softwares (Filezilla – WinSCP). Everything is automatically checked for viruses by the anti-virus software LSE provides and automatically backed-up daily once it is on the network. And we use DROID to profile collection and cmd lines to create file directories or Exiftool to extract embedded metadata on images.



DPC Workflow Webinar 2023

Cataloguing

Processing

Access copy
Appraisal
Catalogue
Ingest

DROID

OpenRefine

LSE Digital Toolkit

Checksum

Teracopy

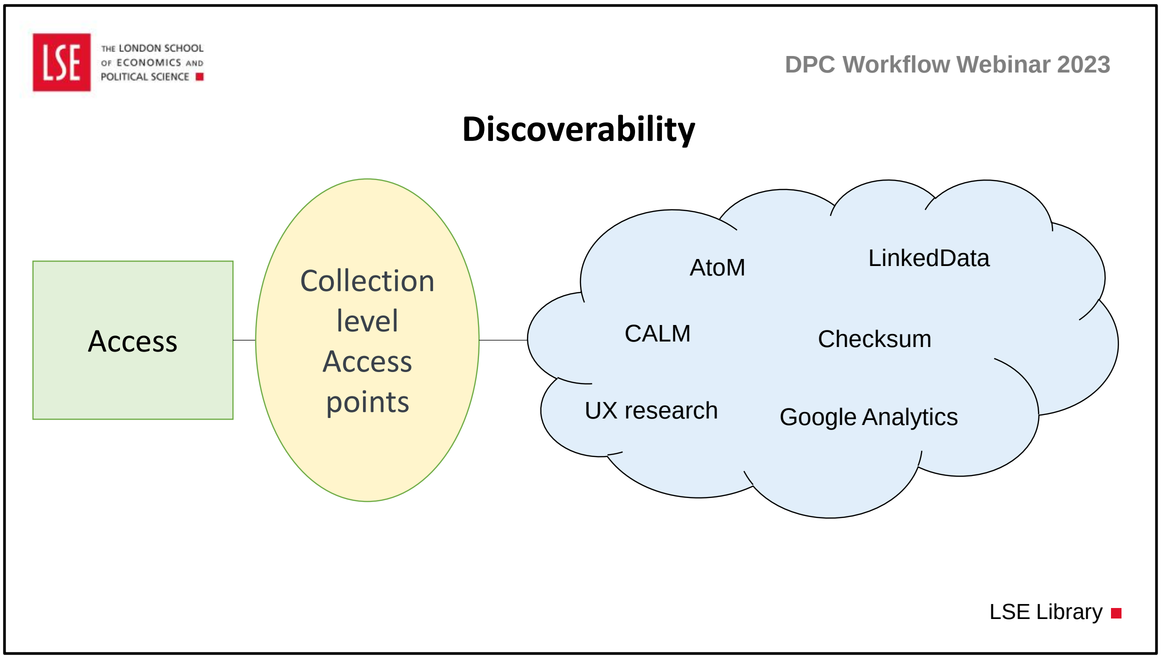
QuickViewPlus

LSE Library

The cataloguing phase is mainly guided by the LSE Digital Toolkit, which was developed by our developer Nick Bywell, who you are going to hear from shortly. I am not going to talk about the toolkit as our colleague Fabi Barticioti gave a demo back in 2021 (share the link after) – although she focused on digitised material, the process is very similar. Main differences are some configurations, b-d files keep original file names and cataloguing metadata fields are slightly different. I'll be happy to share details of these with anyone – I am currently refining the guide for this too which will be publishing on GitHub with the existing one. Also, you will hear a bit about it from Nick shortly.

We create an access copy using Teracopy, to process material to avoid risk of corruption of master copy. We catalogue in CSV file, with the mandatory ISADG fields as per Arkivum requirements (to show in AtoM which is based on ISADG). The lowest level at which we catalogue is file, and digital objects are attached to the file (so one file can have multiple digital objects inside it).

Once the collection is catalogued in the csv, a script creates the folder structure to mirror the hierarchical structure of the collection. We then have to manually move digital objects to the relevant folders (using Teracopy) – hopefully this part of the workflow will be automated at some stage. Then our toolkit allows us to validate the csv file, to check content of folders, and to create the upload package, with digital objects and metadata. We ingest this via WinSCP and it appears in Atom if we can make the material available. If not, we only ingest in the preservation datapool.



DPC Workflow Webinar 2023

Discoverability

Access

Collection level Access points

AtoM

LinkedData

CALM

Checksum

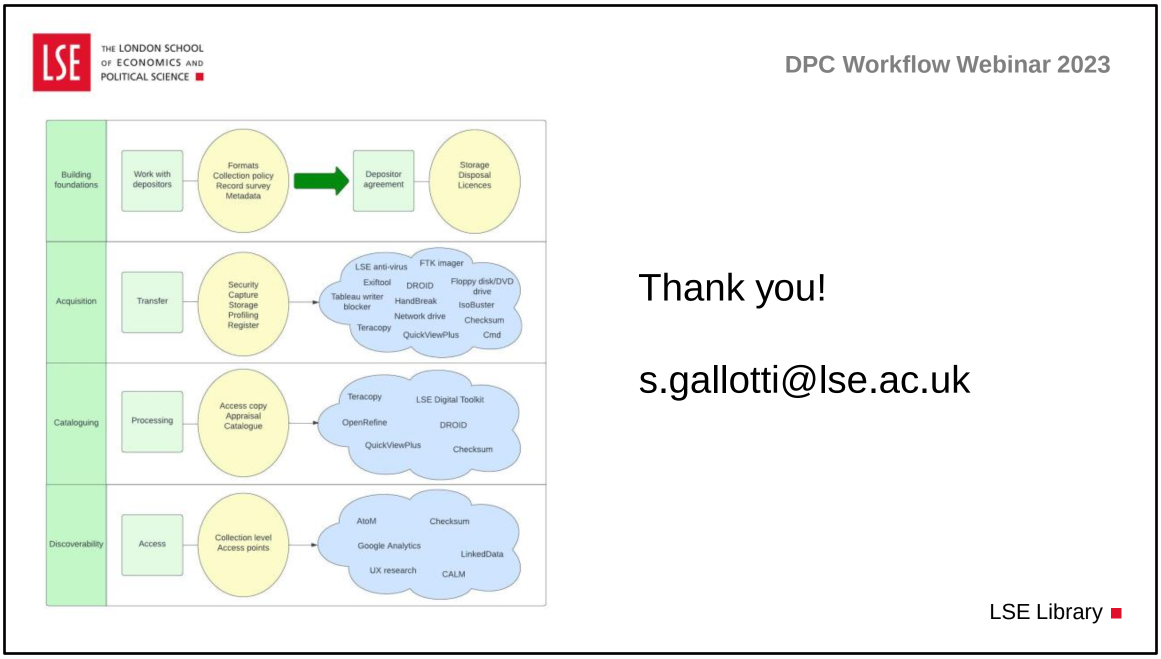
UX research

Google Analytics

LSE Library

Once ingested, Arkivum takes care of all preservation actions for us, and creates DIP which feeds through to AtoM and shows our records there. In AtoM we create high level descriptions and add access points for subjects/people and organisations, using VIAF and Wikimedia terms in the hope we will be able to use LinkedData eventually. We use Google analytics and UX research to work on users needs – this is an area that we will develop further going forward.

We are currently investigating ingesting collections and provide users with temporary logins for accessing restricted material. We should be doing a pilot project with LSE committee minutes.



THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DPC Workflow Webinar 2023

Thank you!

s.gallotti@lse.ac.uk

LSE Library

To conclude, How well I feel our workflow solve the problem that it was designed to solve?

It’s worked very well for the collections we have ingested so far. The first phase is the strongest so far. The second phase will keep expanding, with many more tools to be explored for different material.

The main issue will be to discover how scalable the third phase is, so far we’ve had fairly small collections. And we will be looking to develop further discoverability of our assets.