Collecting Email Archives at the British Library

Jonathan Pledge

Born Digital Archives at the British Library

- Born Digital research started at the BL in 2000.
- Pioneers in the born digital field, in particular forensic capture.
- Since April 2015 we have focused on curator-led processing of born digital archives.
- Work with BL Digital Preservation

Contemporary British Born Digital Archives

- Usually hybrid archives; personal archives, primarily of writers and scientists.
- Smallest born digital collection is 71MB and the largest is 380GB (including system files).
- Many archives contain email particularly where we have acquired a hard drive with operating system.
- Largest email archive is of the poet Wendy Cope consisting of 40,000 emails

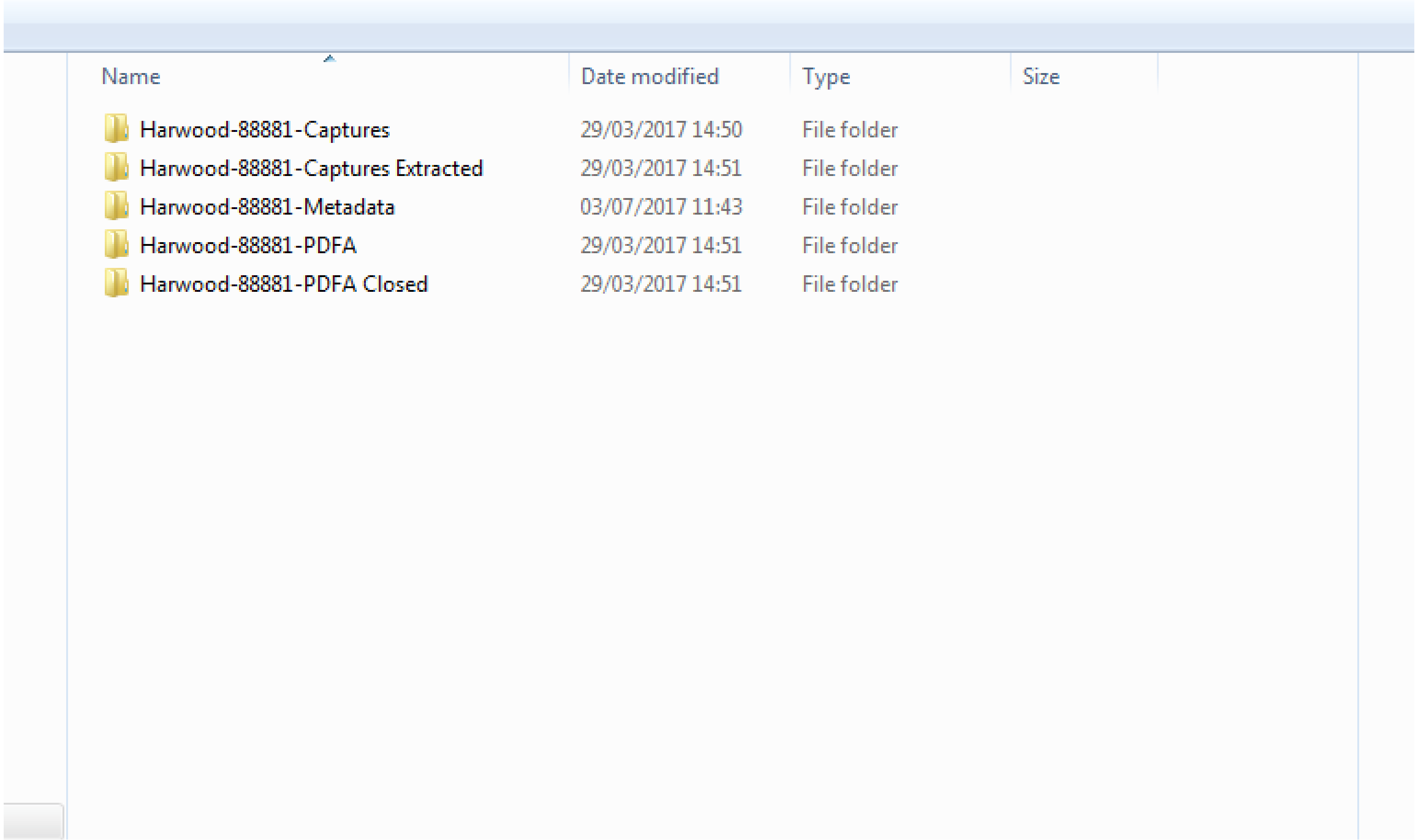
Workflow: Forensic capture

- Hard drives/Flash drives acquired as .E01 files
- CDs/DVDs acquired as iso or .E01 files exported as raw files (.001 or .img)
- Floppy disks (3.5 inch) acquired via Kryoflux as .img or raw files.

Workflow: Processing

- Extract folders and digital objects from captured forensic image.
- Create master copy of extracted material (with a minimum of two back up copies).
- Export master capture to create ‘Extracted Captures’ (replicating original structures and titles)

Workflow: Processing

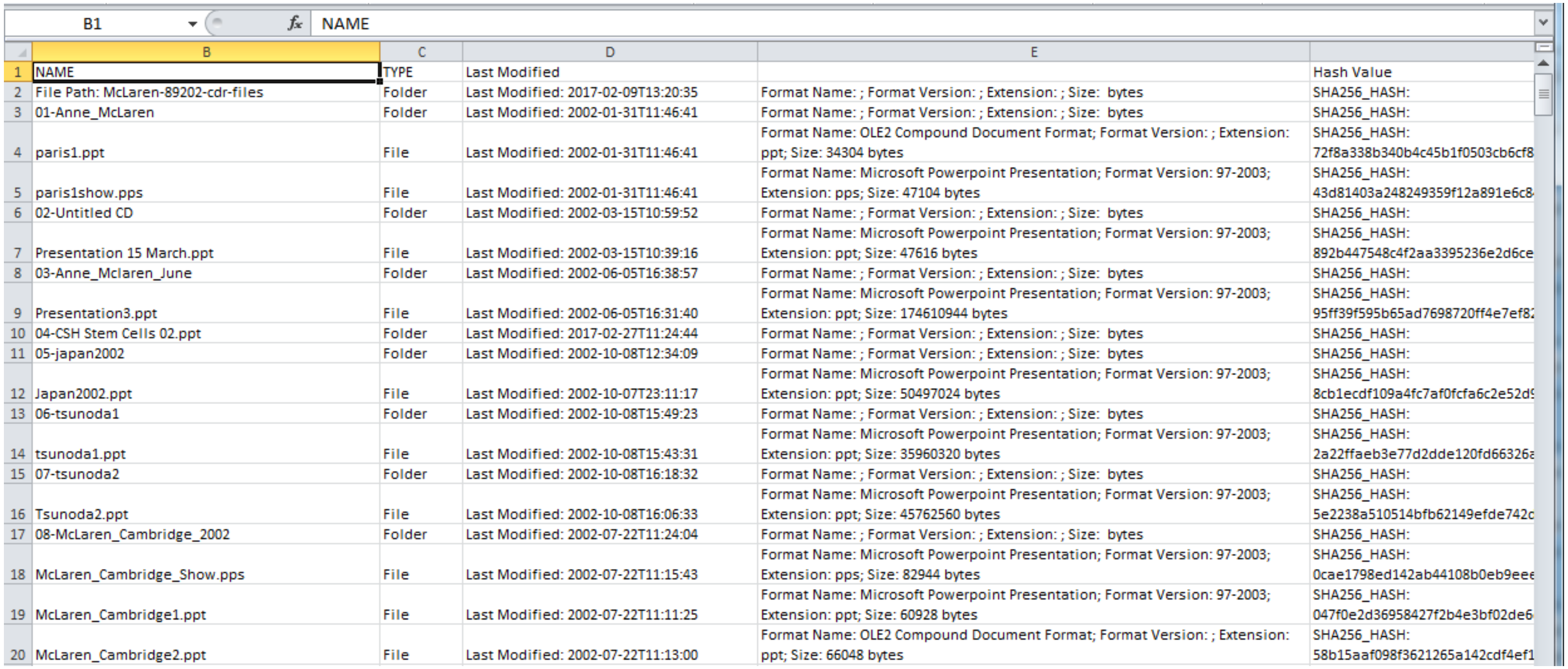


Name	Date modified	Type	Size
Harwood-88881-Captures	29/03/2017 14:50	File folder	
Harwood-88881-Captures Extracted	29/03/2017 14:51	File folder	
Harwood-88881-Metadata	03/07/2017 11:43	File folder	
Harwood-88881-PDFA	29/03/2017 14:51	File folder	
Harwood-88881-PDFA Closed	29/03/2017 14:51	File folder	

Workflow: Processing

- DROID is used to extract metadata from the ‘Extracted Captures’. The .csv metadata file is then saved to Excel, sorted and arranged. This information is then copied to a catalogue migration spreadsheet and published via an uploader to our Explore Archives and Manuscripts catalogue.
- PDF/As (PDF/A-b) are created from the ‘Extracted Captures’ and are renamed in sequential order with archive reference numbers (e.g. Add MS)

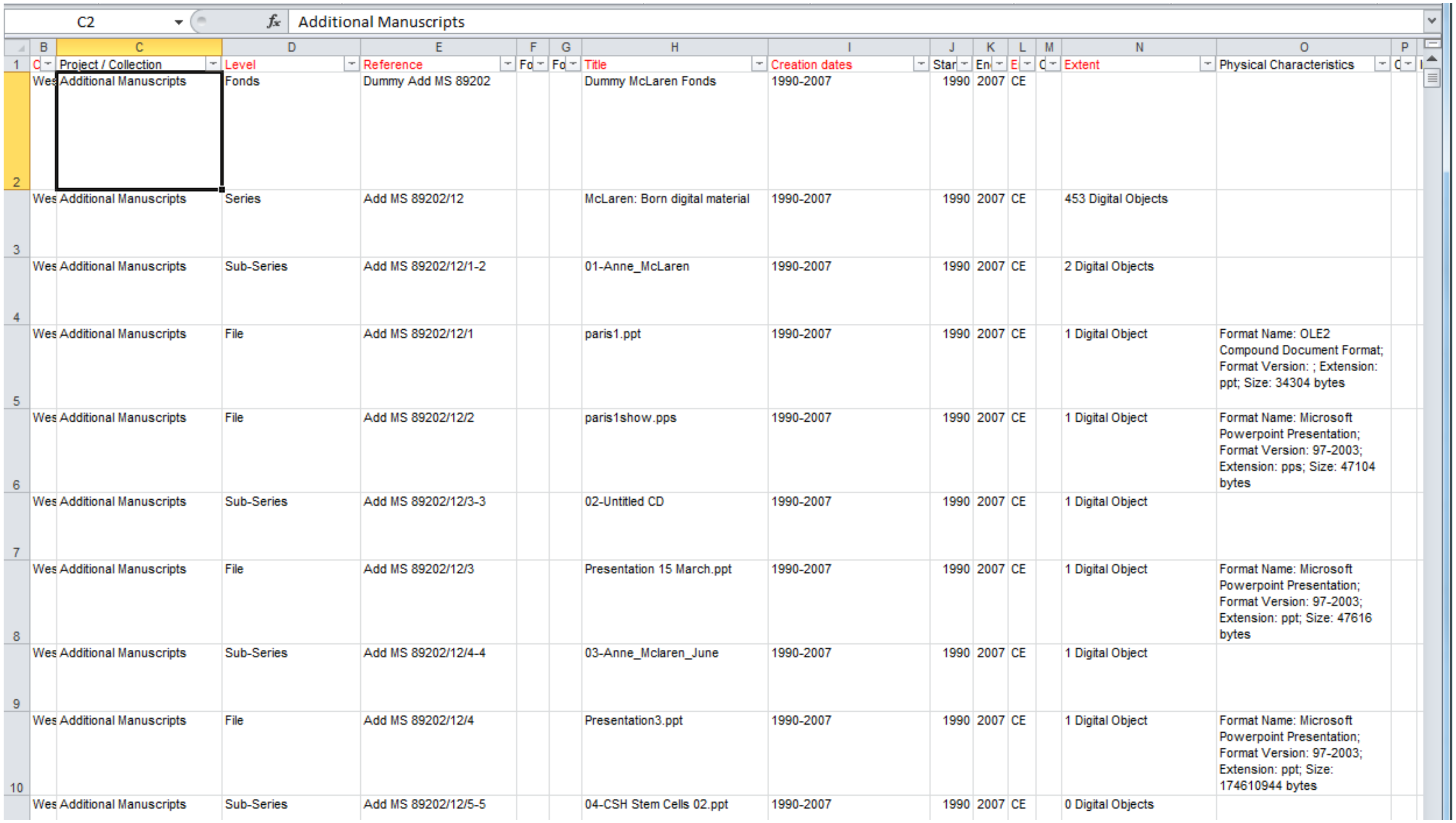
Workflow: Processing



B1 fx NAME

	B	C	D	E	
1	NAME	TYPE	Last Modified		Hash Value
2	File Path: McLaren-89202-cdr-files	Folder	Last Modified: 2017-02-09T13:20:35	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
3	01-Anne_McLaren	Folder	Last Modified: 2002-01-31T11:46:41	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
4	paris1.ppt	File	Last Modified: 2002-01-31T11:46:41	Format Name: OLE2 Compound Document Format; Format Version: ; Extension: ppt; Size: 34304 bytes	SHA256_HASH: 72f8a338b340b4c45b1f0503cb6cf8
5	paris1show.pps	File	Last Modified: 2002-01-31T11:46:41	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: pps; Size: 47104 bytes	SHA256_HASH: 43d81403a248249359f12a891e6c8
6	02-Untitled CD	Folder	Last Modified: 2002-03-15T10:59:52	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
7	Presentation 15 March.ppt	File	Last Modified: 2002-03-15T10:39:16	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 47616 bytes	SHA256_HASH: 892b447548c4f2aa3395236e2d6ce
8	03-Anne_Mclaren_June	Folder	Last Modified: 2002-06-05T16:38:57	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
9	Presentation3.ppt	File	Last Modified: 2002-06-05T16:31:40	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 174610944 bytes	SHA256_HASH: 95ff39f595b65ad7698720ff4e7ef8
10	04-CSH Stem Cells 02.ppt	Folder	Last Modified: 2017-02-27T11:24:44	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
11	05-japan2002	Folder	Last Modified: 2002-10-08T12:34:09	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
12	Japan2002.ppt	File	Last Modified: 2002-10-07T23:11:17	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 50497024 bytes	SHA256_HASH: 8cb1ecdf109a4fc7af0fcfa6c2e52d
13	06-tsunoda1	Folder	Last Modified: 2002-10-08T15:49:23	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
14	tsunoda1.ppt	File	Last Modified: 2002-10-08T15:43:31	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 35960320 bytes	SHA256_HASH: 2a22ffaeb3e77d2dde120fd66326
15	07-tsunoda2	Folder	Last Modified: 2002-10-08T16:18:32	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
16	Tsunoda2.ppt	File	Last Modified: 2002-10-08T16:06:33	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 45762560 bytes	SHA256_HASH: 5e2238a510514bfb62149efde742
17	08-McLaren_Cambridge_2002	Folder	Last Modified: 2002-07-22T11:24:04	Format Name: ; Format Version: ; Extension: ; Size: bytes	SHA256_HASH:
18	McLaren_Cambridge_Show.pps	File	Last Modified: 2002-07-22T11:15:43	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: pps; Size: 82944 bytes	SHA256_HASH: 0cae1798ed142ab44108b0eb9eee
19	McLaren_Cambridge1.ppt	File	Last Modified: 2002-07-22T11:11:25	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 60928 bytes	SHA256_HASH: 047f0e2d36958427f2b4e3bf02de6
20	McLaren_Cambridge2.ppt	File	Last Modified: 2002-07-22T11:13:00	Format Name: OLE2 Compound Document Format; Format Version: ; Extension: ppt; Size: 66048 bytes	SHA256_HASH: 58b15aaf098f3621265a142cdf4ef1

Workflow: Processing



C2 Additional Manuscripts

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	C	Project / Collection	Level	Reference	Fo	Fo	Title	Creation dates	Star	En	E	C	Extent	Physical Characteristics	C
2	Wes	Additional Manuscripts	Fonds	Dummy Add MS 89202			Dummy McLaren Fonds	1990-2007	1990	2007	CE				
3	Wes	Additional Manuscripts	Series	Add MS 89202/12			McLaren: Born digital material	1990-2007	1990	2007	CE		453 Digital Objects		
4	Wes	Additional Manuscripts	Sub-Series	Add MS 89202/12/1-2			01-Anne_McLaren	1990-2007	1990	2007	CE		2 Digital Objects		
5	Wes	Additional Manuscripts	File	Add MS 89202/12/1			paris1.ppt	1990-2007	1990	2007	CE		1 Digital Object	Format Name: OLE2 Compound Document Format; Format Version: ; Extension: ppt; Size: 34304 bytes	
6	Wes	Additional Manuscripts	File	Add MS 89202/12/2			paris1show.pps	1990-2007	1990	2007	CE		1 Digital Object	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: pps; Size: 47104 bytes	
7	Wes	Additional Manuscripts	Sub-Series	Add MS 89202/12/3-3			02-Untitled CD	1990-2007	1990	2007	CE		1 Digital Object		
8	Wes	Additional Manuscripts	File	Add MS 89202/12/3			Presentation 15 March.ppt	1990-2007	1990	2007	CE		1 Digital Object	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 47616 bytes	
9	Wes	Additional Manuscripts	Sub-Series	Add MS 89202/12/4-4			03-Anne_Mclaren_June	1990-2007	1990	2007	CE		1 Digital Object		
10	Wes	Additional Manuscripts	File	Add MS 89202/12/4			Presentation3.ppt	1990-2007	1990	2007	CE		1 Digital Object	Format Name: Microsoft Powerpoint Presentation; Format Version: 97-2003; Extension: ppt; Size: 174610944 bytes	
	Wes	Additional Manuscripts	Sub-Series	Add MS 89202/12/5-5			04-CSH Stem Cells 02.ppt	1990-2007	1990	2007	CE		0 Digital Objects		

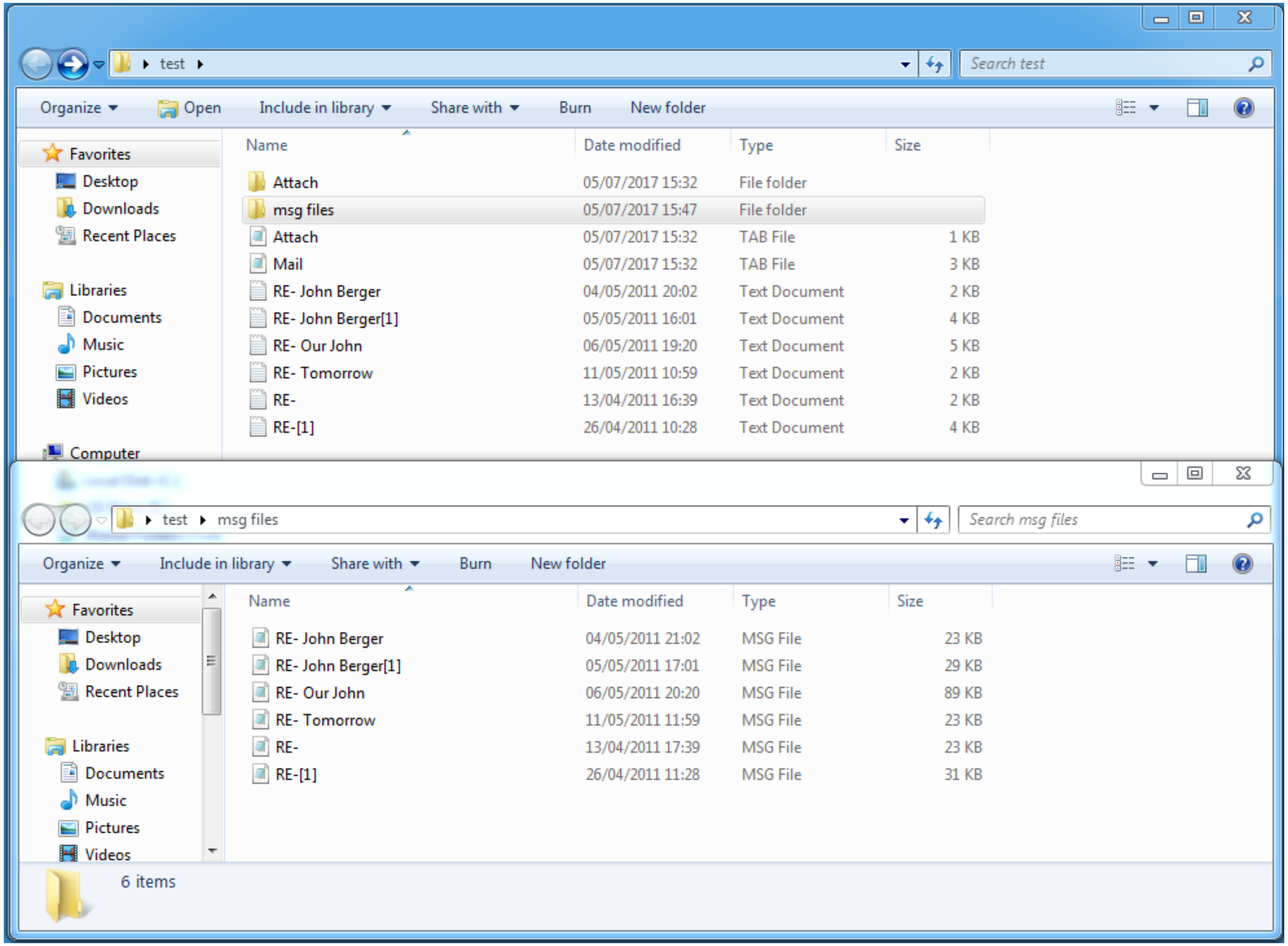
Workflow: Delivery

- PDF/As are uploaded to a BL FTP Server in labelled directories. The server is hosted in the BL Manuscripts Reading Room.
- Born Digital Material is accessed, as PDF/As, through dedicated terminals in the Manuscripts Reading Room via a web browser.
- Digital objects can only be accessed in the Manuscripts Reading Room and cannot be edited, printed or saved.

Email processing

- Available resources: Aid4Mail Forensic (Fookes Holding, Switzerland) and ePADD (Stanford University)
- Aid4Mail Forensic allows for processing and delivery as per our current workflow
- ePADD would seem to offer a holistic approach to email preservation – to be investigated

Email processing



Name	Date modified	Type	Size
Attach	05/07/2017 15:32	File folder	
msg files	05/07/2017 15:47	File folder	
Attach	05/07/2017 15:32	TAB File	1 KB
Mail	05/07/2017 15:32	TAB File	3 KB
RE- John Berger	04/05/2011 20:02	Text Document	2 KB
RE- John Berger[1]	05/05/2011 16:01	Text Document	4 KB
RE- Our John	06/05/2011 19:20	Text Document	5 KB
RE- Tomorrow	11/05/2011 10:59	Text Document	2 KB
RE-	13/04/2011 16:39	Text Document	2 KB
RE-[1]	26/04/2011 10:28	Text Document	4 KB

Name	Date modified	Type	Size
RE- John Berger	04/05/2011 21:02	MSG File	23 KB
RE- John Berger[1]	05/05/2011 17:01	MSG File	29 KB
RE- Our John	06/05/2011 20:20	MSG File	89 KB
RE- Tomorrow	11/05/2011 11:59	MSG File	23 KB
RE-	13/04/2011 17:39	MSG File	23 KB
RE-[1]	26/04/2011 11:28	MSG File	31 KB

Data Protection Act 1998

- UK Act of Parliament.
- In response to 1995 EU Data Protection Directive.
- Defines the law on the processing of data as ‘Personal information’ on identifiable living people.
- ‘Personal information’ means any information about an identifiable living individual in any format. It applies to paper, emails, photographs and sound recordings.
- Exemptions for journalism, literature and art.

Challenges

- Email is a unique entity which often contains ‘personal information’. The nature of the medium.
- Processing becomes a time consuming process due to DP checking of hundreds/thousands of emails.
- Complications arise where emails are saved as threads meaning that in the current BL DP model a single email can mean that the entire thread will be removed.
- Solution may be redaction of offending material at the risk of a loss of meaning.

Thank you

jonathan.pledge@bl.uk
eleanor.dickens@bl.uk