Born-digital archives @ Hull: early steps and lessons learnt (so far)

Simon Wilson, Digital Archivist (AIMS Project)

outline

What is the AIMS Project?

Steps taken @ Hull

Steps still to take

Questions

A 2GB pen drive is capable of holding more than 900 photographs

University of Virginia

University of Hull

Stanford University

Yale

AIMS Project

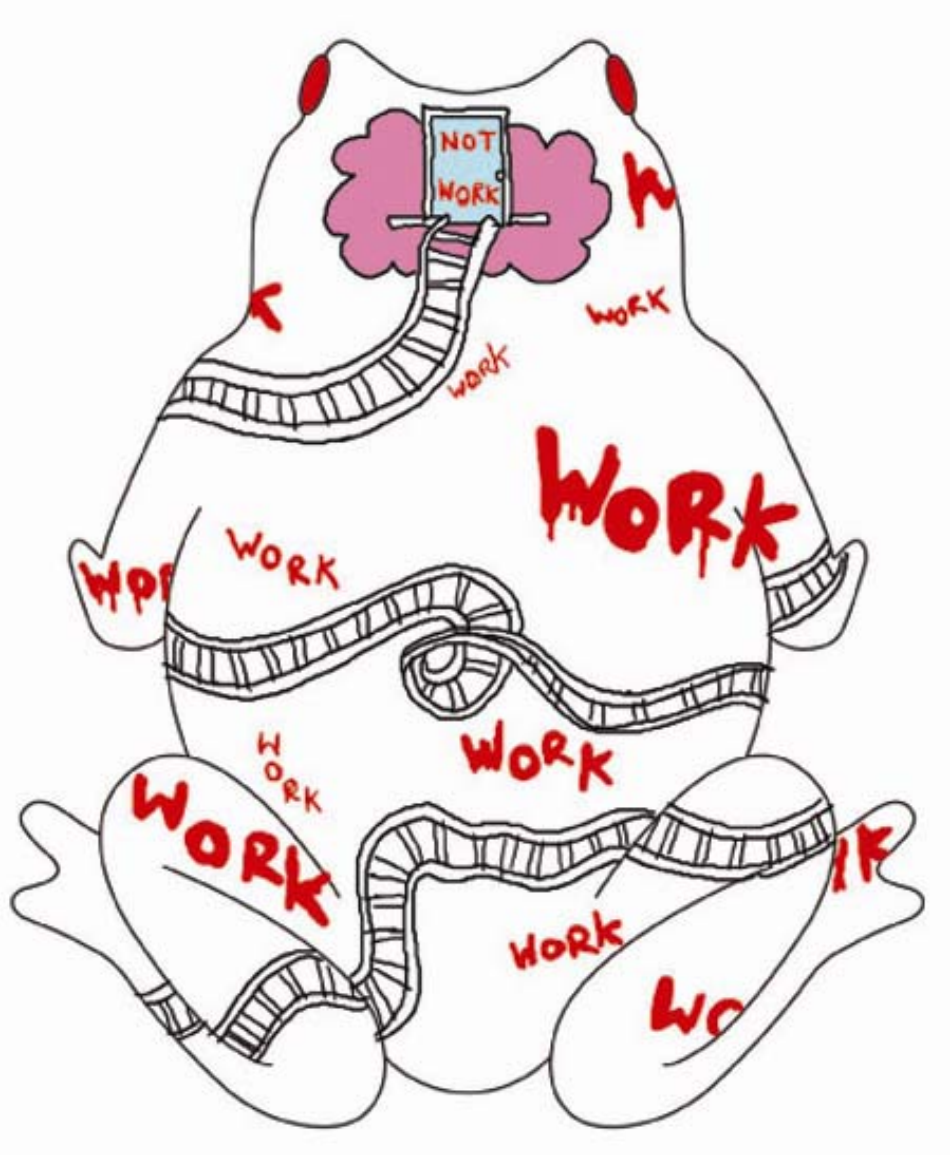
An inter-**I**nstitutional **M**odel for **S**tewardship

Each partner employing a Digital Archivist

Two year project (ends Sept 2011)

To **process** born-digital collections

To use Hydra, a **Fedora** repository-based solution



AIMS Approach

Starting point - traditional **archival theory/principles**

AIMS Model – identify **commonality** across the 4 partners - not looking to create a single solution

To **disseminate** the methods & lessons learnt

Virtual collaboration – weekly **Skype** calls and ***GoogleDocs*** - multiple editors work on same text

but also face to face mtgs

Toad in the Hull, design for a toad that was never built (Larkin with Toads archive)

look in your store

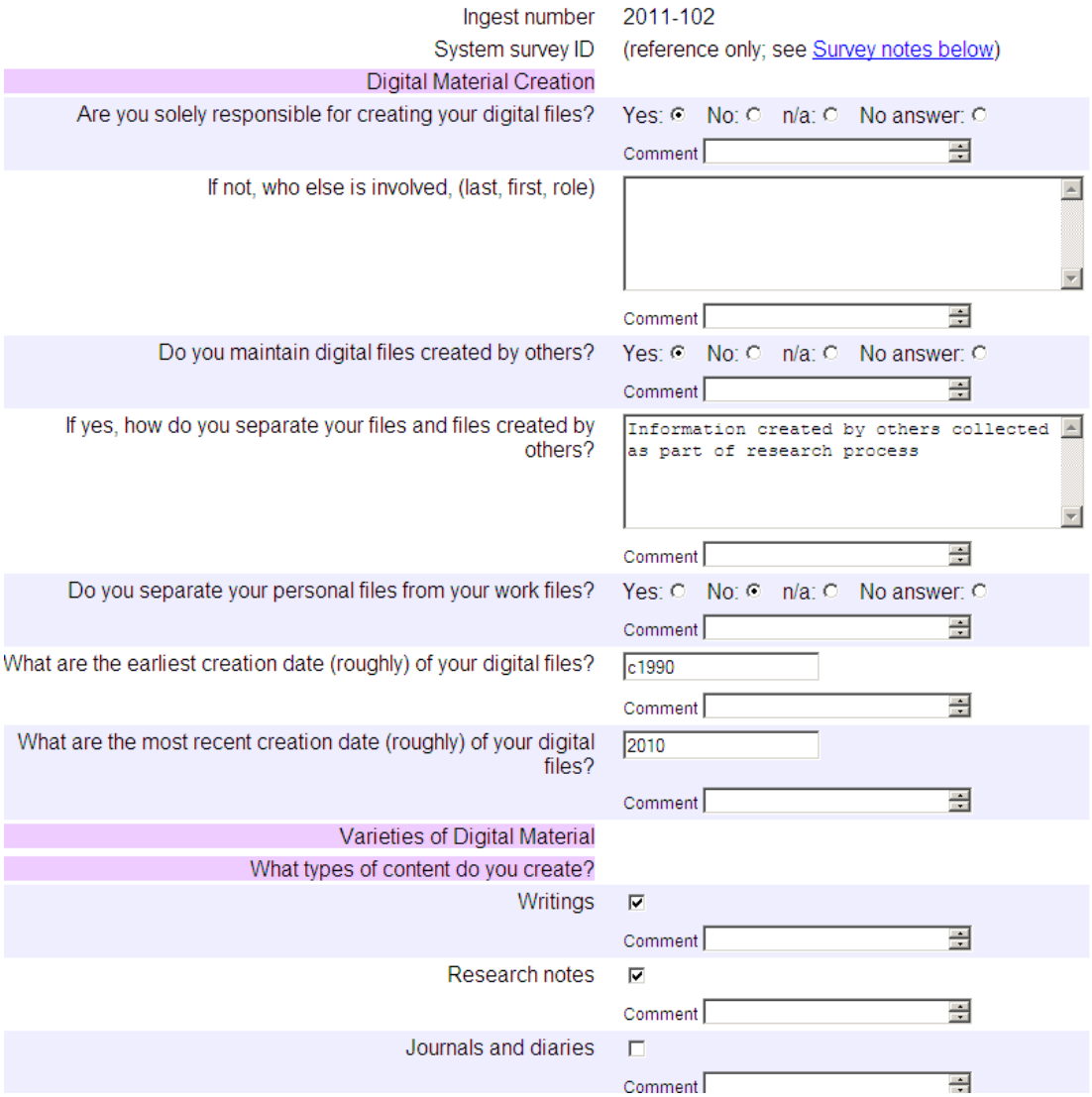
Ref	Collection Name	Item Description	No files	Size (GB)	Media inc. issues	No formats	Format Issues
U DBV (2007/05)	British Union for the Abolition of Vivisection	2 CDs containing recordings of You and Your from Radio 2, 2004.	2	0.004	Audio	1	Audio .cda for Can this be converted (to MP3) post-ing
U DCL/911/1	Liberty	File. Police, Lay Visiting Information Sheet.	[1]	TBC	5.25" floppy disk	1	WordPerfect
U DEN3/149	Papers of Sir William Empson	Original version (b-d) Ts. text of draft book, 'The face of the Buddha', by William Empson	2	0.56	CD same file in RTF & Doc formats	2	RTF file is 500
U DFF/7/3	Fred Fletcher	List of contents of Fred Fletcher's discs relating to local and Quaker history, labelled 'Disc index'			floppy disc (tbc?)		
U DGA (Add) 2010/15	Stephen Gallagher	born digital material including drafts, scripts etc	14,320	13.6	Amstrad discs (x ?);		
U DJE/*/*	Jean Hartley						
U DOJ Add 2010/16	Out of Joint	born digital material including organisational and productions upto 19?	10,283	2.25	extracted from server onto external hard drive		
U DSM Add 2004/13	Socialist Health Association	tbc Not in box - box marked 'return to depositor'	1712	0.16	CD		

Survey your holdings for existing (digitised or) born-digital material - search your catalogue for CD, disc etc

Identify distinct **media** and **content** issues, possible level of cataloguing etc

Knowledge of quantity, media & range of media formats is a step forward

Screenshot of survey spreadsheet using MS Excel



Ingest number	2011-102
System survey ID	(reference only; see Survey notes below)
Digital Material Creation	
Are you solely responsible for creating your digital files?	Yes: ◉ No: ○ n/a: ○ No answer: ○ Comment
If not, who else is involved, (last, first, role)	Comment
Do you maintain digital files created by others?	Yes: ◉ No: ○ n/a: ○ No answer: ○ Comment
If yes, how do you separate your files and files created by others?	Information created by others collected as part of research process Comment
Do you separate your personal files from your work files?	Yes: ○ No: ◉ n/a: ○ No answer: ○ Comment
What are the earliest creation date (roughly) of your digital files?	c1990 Comment
What are the most recent creation date (roughly) of your digital files?	2010 Comment
Varieties of Digital Material	
What types of content do you create?	
Writings	☑ Comment
Research notes	☑ Comment
Journals and diaries	☐ Comment

depositors

Relationship more **critical** than ever before, new **questions** we need to ask – hardware, software, passwords etc

Stephen Gallagher: 2005 - 42 boxes paper archives inc some of his TV work on video/DVD

2010 - born-digital material – 14,320 digital files (13.6GB) via an external hard drive

Screenshot of web –based survey developed by the AIMS Project, currently in beta

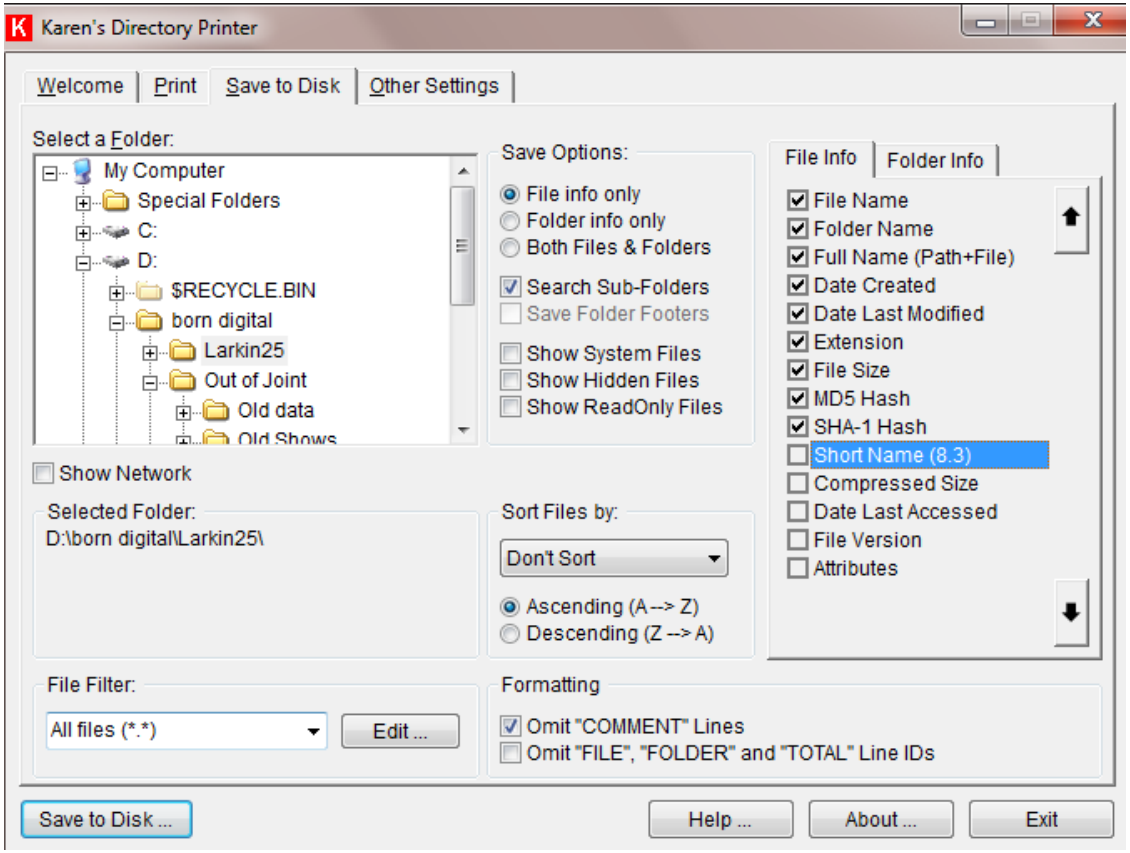
quarantine space

Setting-up a **quarantine PC** - test files for viruses

Purchase new PC with lots of **expansion** space, look to create some **backward compatibility** eg USB floppy / zip drive

Establish **policies** and **procedures** for receiving and managing born-digital archives

Karen's Dir Printer



Useful **free** software tool allows you to retrieve key information about each file/folder inc date created, checksum etc

Archives and/or depositor – create a file **manifest** at the time of transfer

Also use **retrospectively** for material already deposited

Karen's Directory Printer software – see http://www.karenware.com/powertools/ptdirprn.asp

Jump to full list

Category summary for accession ingested files

data	1
other	7
still image	1
textual	3
Total	12

DROID identities for the accession ingested files

PUID (DROID hit count)	Format	Mime-type	File name
(1) x-fmt/265	Tape Archive Format	application/x-tar	accession/arch_1/outer.tar
(1) x-fmt/265	Tape Archive Format	application/x-tar	accession/arch_1/outer/inner.tar
(1) fmt/96	Hypertext Markup Language	text/html	accession/arch_1/outer/inner/icewm_howto_hints.html
(1) x-fmt/263	ZIP Format	application/zip	accession/arch_1/simon_test_files.zip
(1) x-fmt/263	ZIP Format	application/zip	accession/arch_1/simon_test_files/NHS_survey_results.ods
(1) fmt/18	Portable Document Format	application/pdf	accession/arch_1/simon_test_files/clp_SHA.pdf
(1) x-fmt/429	Microsoft Web Archive	text/html	accession/arch_1/simon_test_files/socialisthealth-licensed_for_non-com
(1) x-fmt/263	ZIP Format	application/zip	accession/arch_1/simon_test_files/stafford.odt
(1) x-fmt/265	Tape Archive Format	application/x-tar	accession/arch_1/tmp.tar
(1) fmt/111	OLE2 Compound Document Format	application/msword	accession/arch_1/tmp/AIMS+architecture-v03-100426.doc
(1) x-fmt/263	ZIP Format	application/zip	accession/arch_1/tmp/NHS_survey_results.ods
(2) fmt/62	Binary Interchange File Format (BIFF) Workbook	application/vnd.ms-excel	accession/arch_1/tmp/mascot_tables.xls

Full list of files in the SIP

51699b33-cc67-48b8-a793-aaaeef23cae8

Rubymatica

Tool (developed by AIMS project programmer and currently in beta testing)

Takes the files that are to be **ingested** into Fedora and creates a metadata wrapper

Use DROID, FITS and other established tools

A screenshot from a Rubymatica report

steps still to take

Getting the material into Fedora will be a huge achievement....

... **but** we still have the following on our to do list:

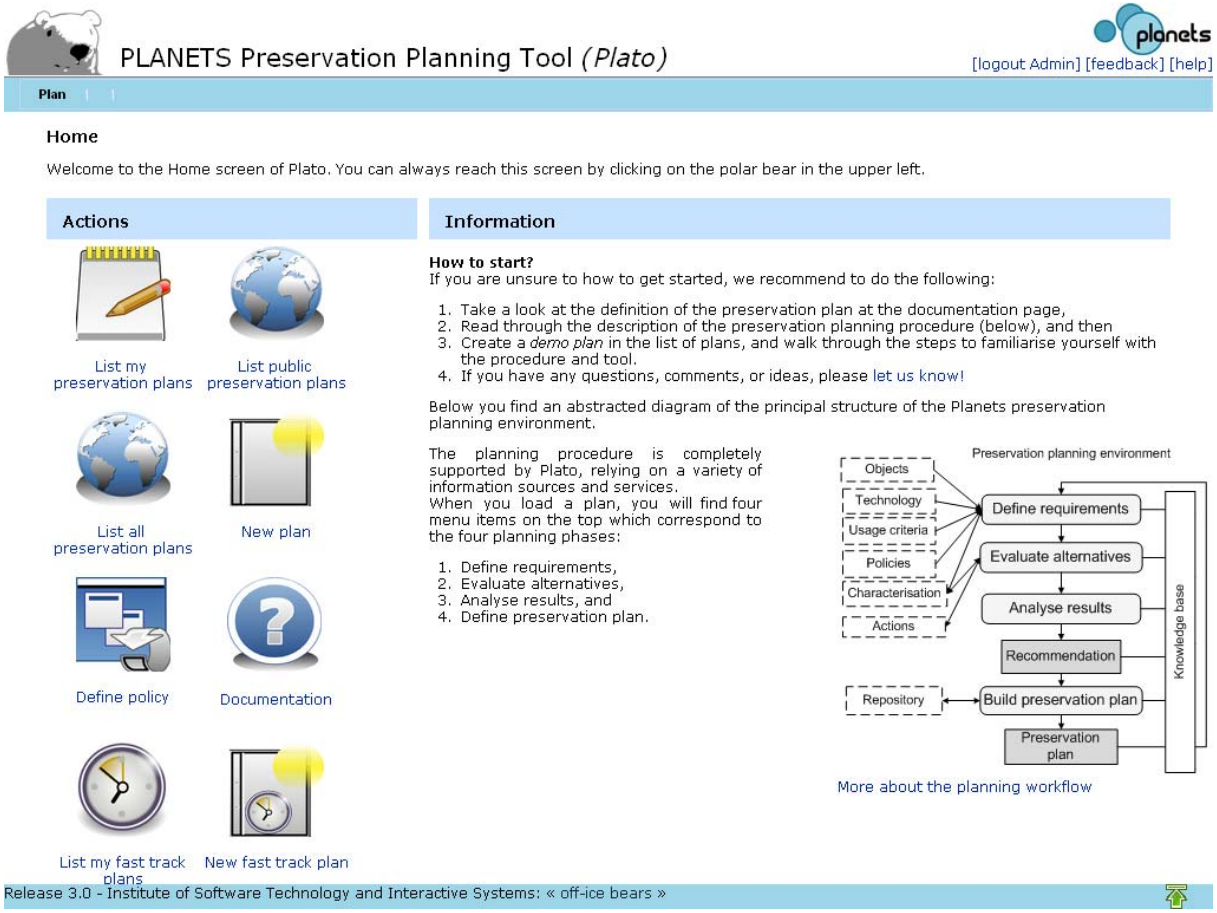
1. Email (watching brief)

2. Preservation actions

3. Arrangement & Description

4. Discovery & Access

Staromestska Metro Station, Prague http://www.flickr.com/photos/dcdead/4678139577/



Plato

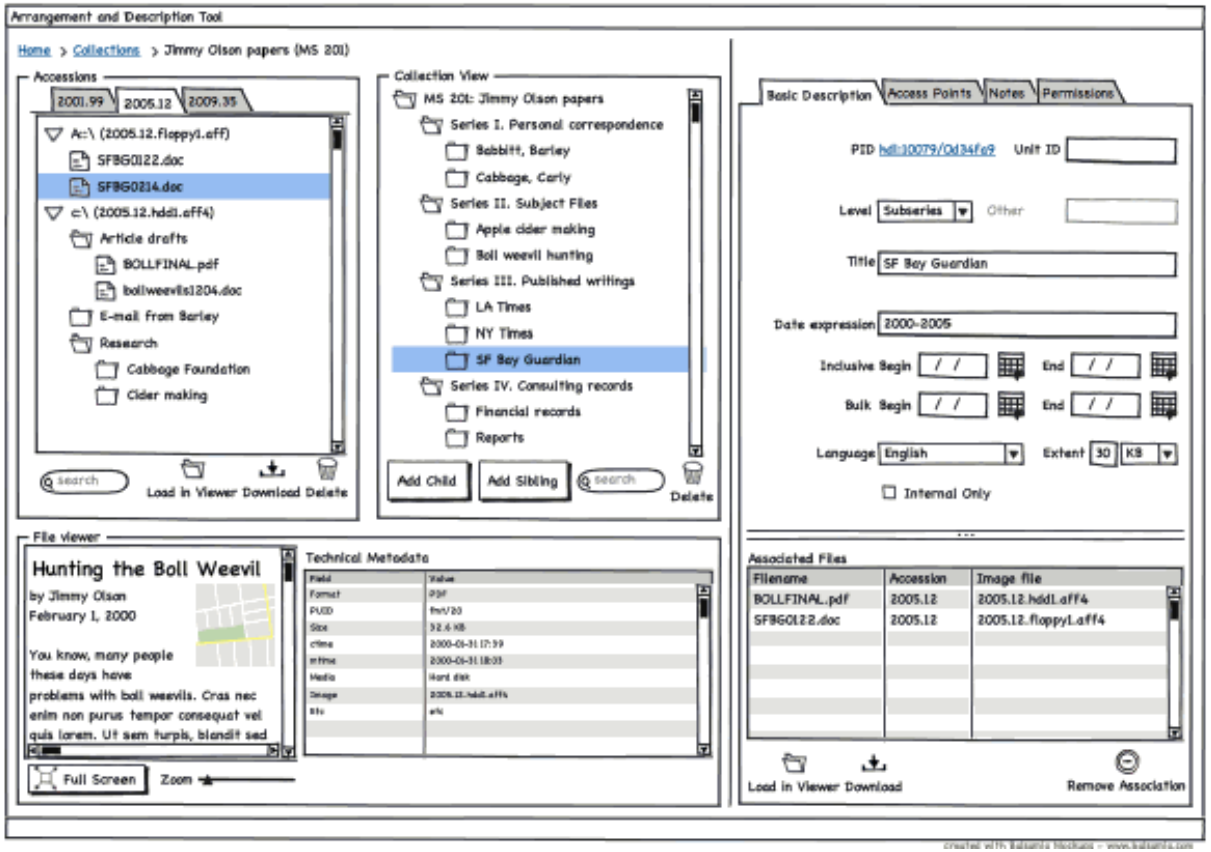
We will be looking to start using the **Plato** tool to test **migration** from one file format to another

Experiments are run on test corpus of files

Establish **policies** and **procedures** for receiving and managing born-digital archives

Plato – see http://www.ifs.tuwien.ac.at/dp/plato/intro.html

Hypatia



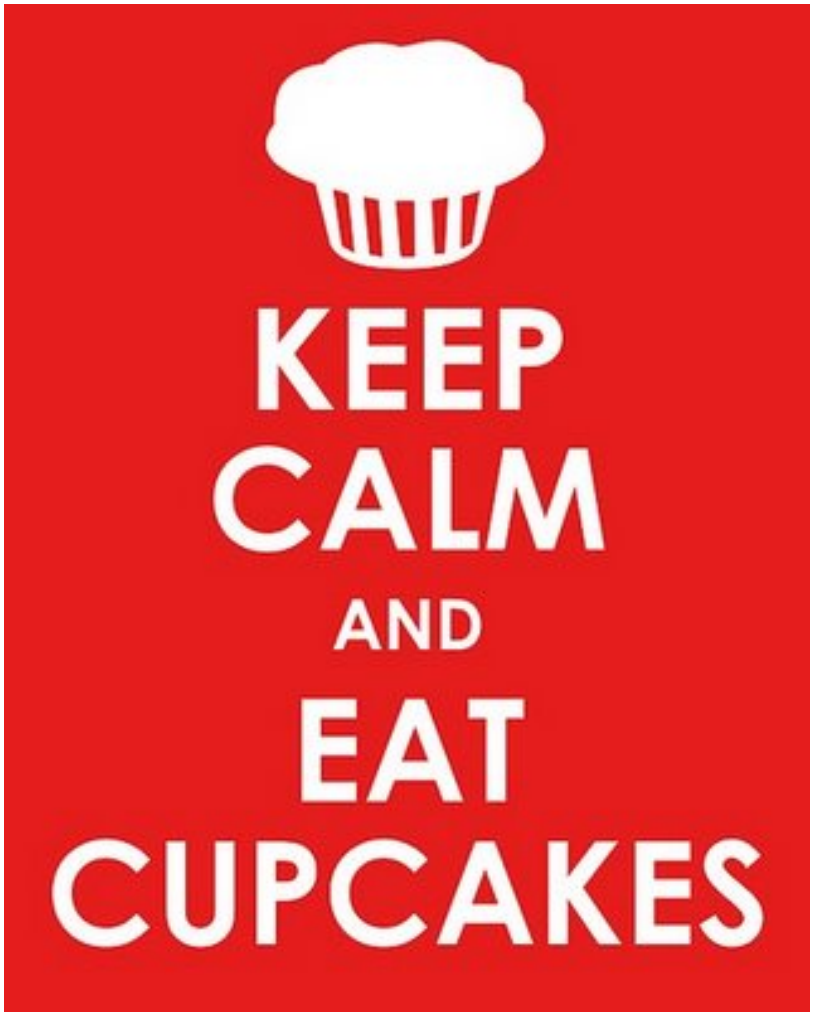
How do we arrange and describe born-digital archives?

Ideally we would like to view the file & integrate it with finding aids etc

Hypatia - new tool defined by archivists for archivists

Proof of concept is currently being built by developers – showcase at SAA in Chicago (tbc)

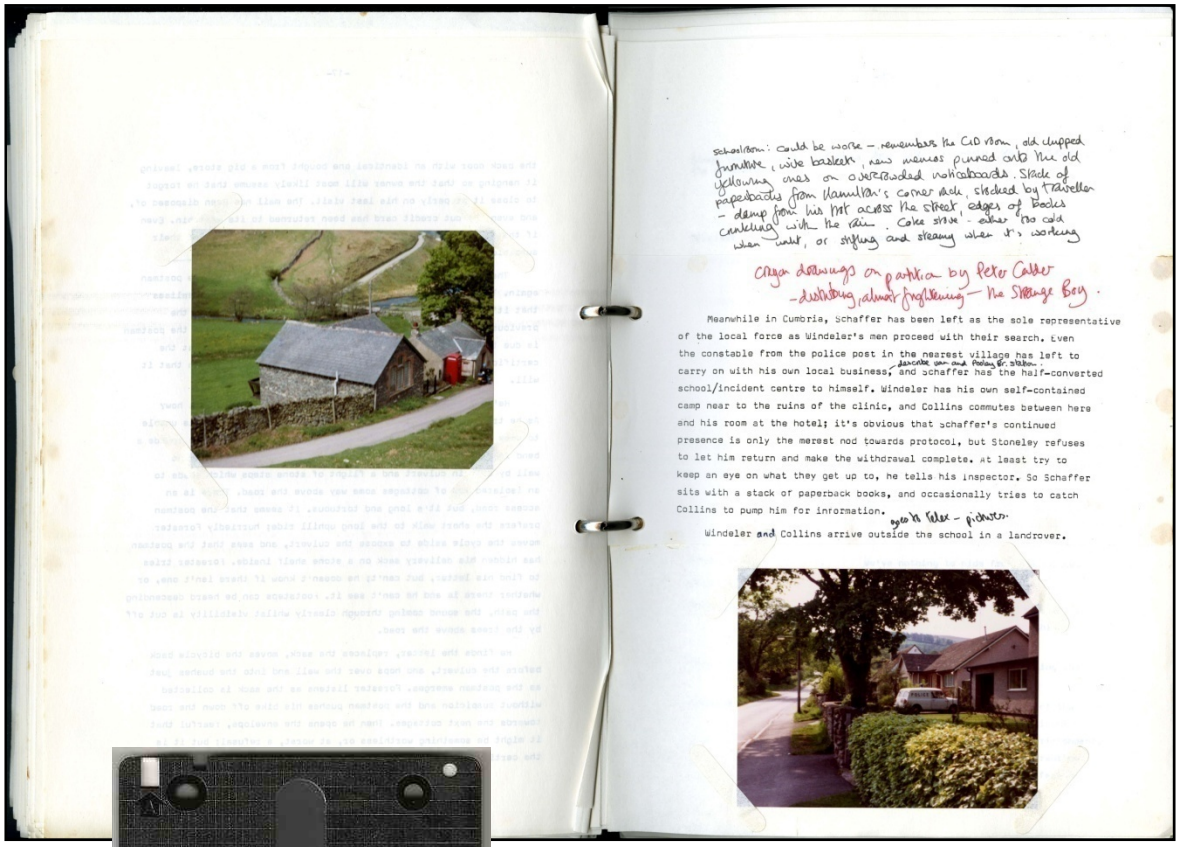
Screen mock-up for the Hypatia tool



Axiell

Working with colleagues at LSE, Parliamentary Archives, Wellcome Library etc

Looking at how CALM can exchange information / data (via API) with digital repository's workflow



Meanwhile in Cumbria, Schaffer has been left as the sole representative of the local force as Windeler's men proceed with their search. Even the constable from the police post in the nearest village has left to carry on with his own local business, and Schaffer has the half-converted school/incident centre to himself. Windeler has his own self-contained camp near to the ruins of the clinic, and Collins commutes between here and his room at the hotel; it's obvious that Schaffer's continued presence is only the merest nod towards protocol, but Stoneley refuses to let him return and make the withdrawal complete. At least try to keep an eye on what they get up to, he tells his inspector. So Schaffer sits with a stack of paperback books, and occasionally tries to catch Collins to pump him for information.

Windeler and Collins arrive outside the school in a landrover.

user experience

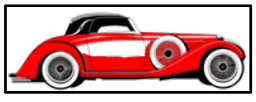
Paper archives offer a **sense of discovery**

Notebooks - plot outlines, notes of meetings with producers, snippets of dialogue etc for different work all **intermingled**

Born-digital material the same information is **dispersed** between multiple digital files

Acc 2008 box 15, Chimera file 2

archives workflow...



Quarantine

Check for mould & infestation etc

Archives Work room

Initial review / appraisal

Donor survey
talk to depositor

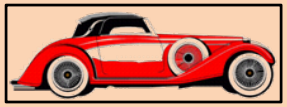
Re-box the material into archive folders / boxes etc

into archive repository

Accession record

REPOSITORY

Produce surrogate copies for user access (m/film)



Cataloguing Room
first serious look at material, appraisal, arrangement & description; identify any legal/access restrictions etc

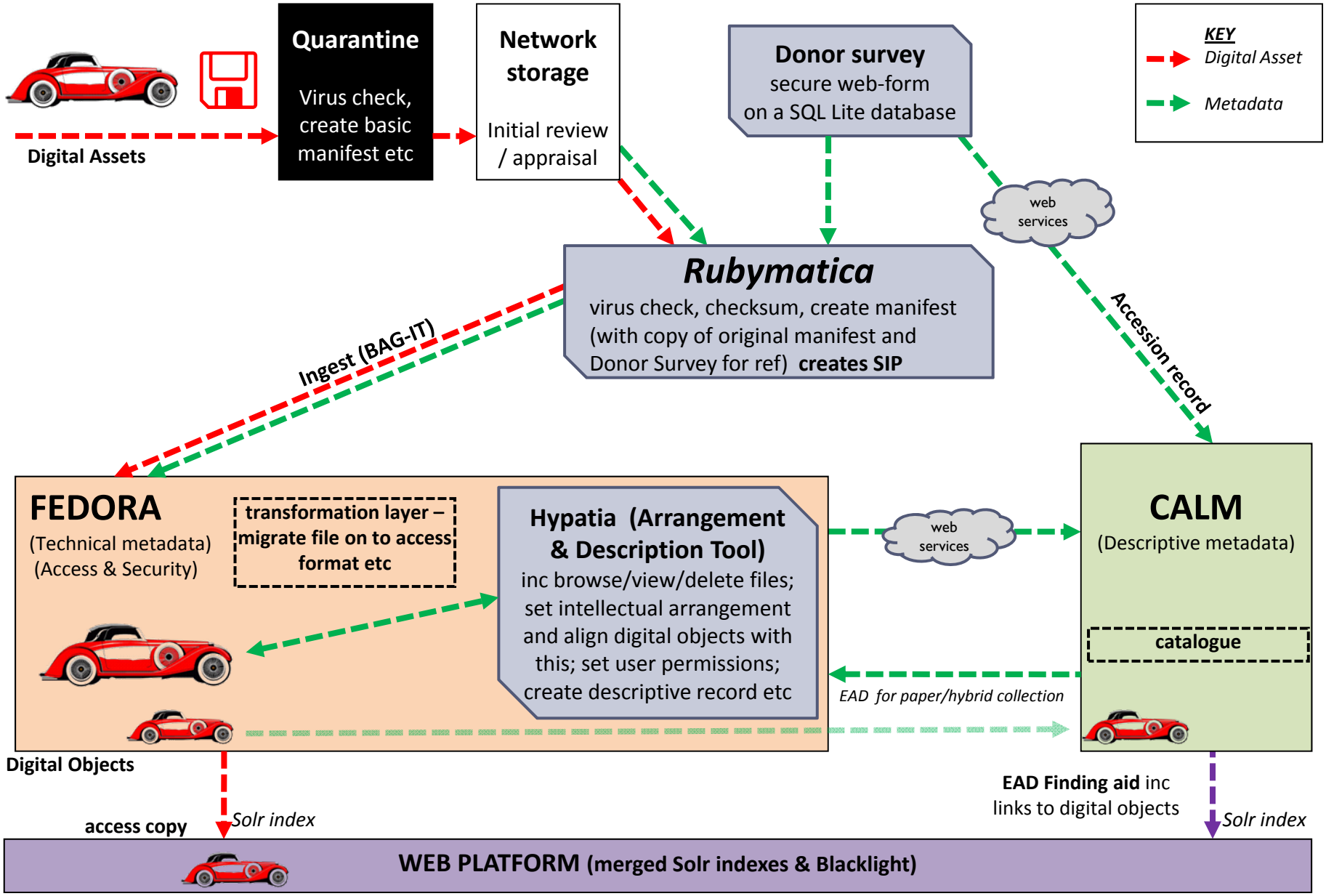
CALM

catalogue

original item produced in search room

WEB PLATFORM – catalogue available online

AIMS, CALM & Fedora workflow...



scale of the task

We don’t have a complete solution in-place yet....

... but we already have **40,000** born-digital files from just **4 collections**

Expect to have **1m+** born-digital files in **5 years** and that **backlog** will be measured in **TB**

Two banks of servers at Facebook by Darren Mckeeman
http://www.flickr.com/photos/tjcrowley/2419036573/in/photostream/

conclusion

The nature (and format) of archives has **changed dramatically** in recent years

It is not for archivists to try and **predict** the future

We **can't** "keep everything" and **hope** *Google* create an algorithm to enable access

Archivists : already have the skills – just need the **confidence** to appreciate they are applicable

on monday

There are some things you can do **on Monday ...**

1. Survey your collections

2. Install Karen's Directory Printer or DROID etc

3. Have a play – hands-on learning does count

You already have the skills – apply to this new domain

Vourvourou ,Greece, http://www.flickr.com/photos/dcdead/4720653919/

any questions?

Contact details

Simon Wilson
Digital Archivist
Hull History Centre

Tel 01482 317506
s.wilson@hull.ac.uk

AIMS Project blog – http://born-digital-archives.blogspot.com

Portrait of Claude-Henri Watalet blogging, after Jean-Baptiste Greuze
http://www.flickr.com/photos/notionscapital/2497196140/in/photostream/