MAKING SENSE OF A COLLECTION

Gareth Knight

London School of Hygiene & Tropical Medicine

gareth.knight@lshtm.ac.uk

Getting Started in Digital Preservation

The Information Technologists, London

23rd April 2015

Case Studies

ahds arts and humanities data service

National service that preserved research, teaching and learning resources in arts & humanities between 1996 - 2008

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Institutional RDM service that helps LSHTM researchers to curate & preserved research data in public health & tropical medicine

Need for Digital Preservation

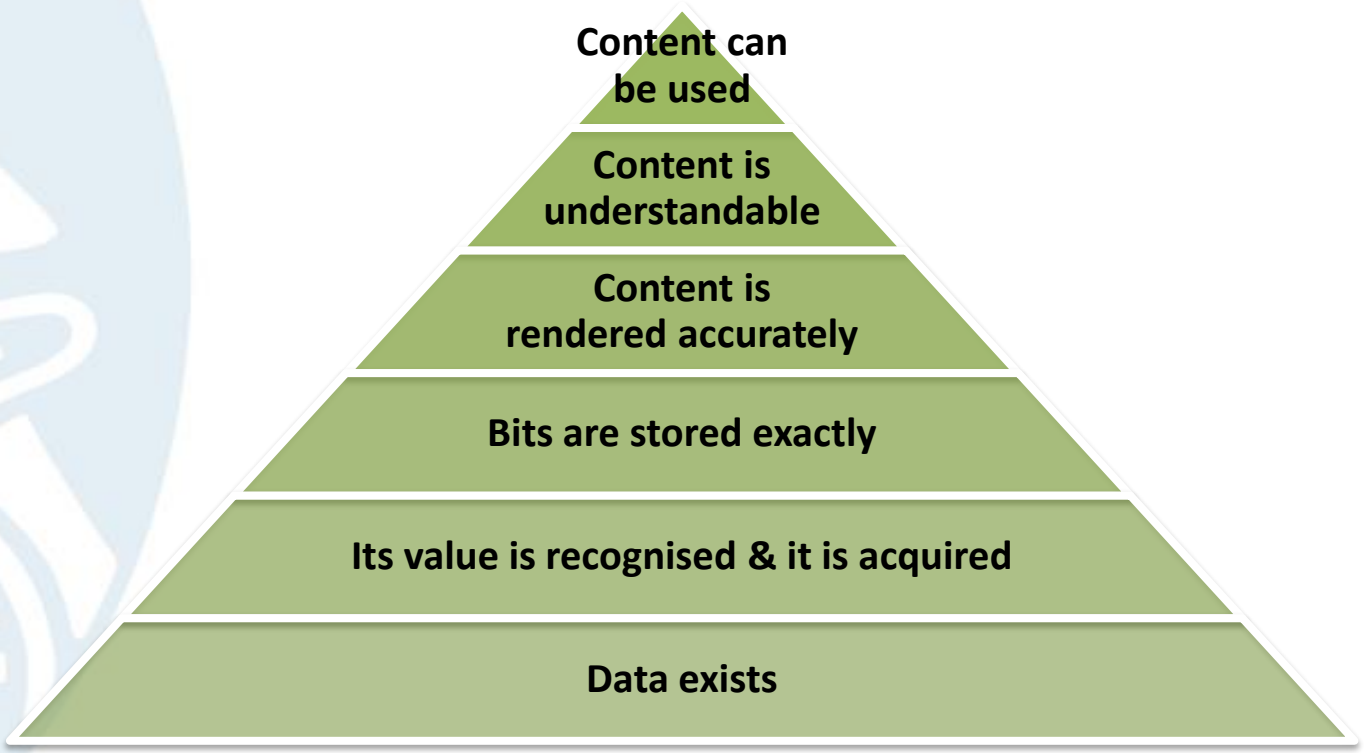
“Digital information lasts forever – or five years, whichever comes first”

Jeff Rothenberg, 1997

Digital Preservation

"the series of managed activities necessary to ensure continued access to digital materials for as long as necessary."

Neil Beagrie and Maggie Jones (2008)



Modified version of Caplan's Preservation Pyramid

Beagrie & Jones: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts
Caplan: http://journals.ala.org/ltr/article/view/4224/4809

Digital Detectives

- Digital preservation often a process of investigation & deduction
- Resource intensive
  - Time
  - Physical space
  - Hardware/software costs
- How much effort are you willing to make? What is good enough?

https://www.flickr.com/photos/ollieolarte/3028314931

Acquire data

Acquisition depends upon object to be preserved & how stored

- Media: Floppy disk, CD/DVD, ZIP/Jaz disk, hard disk, solid state devices, etc.
- Electronic: Email, cloud services

Invest in infrastructure to support preservation process

- Standalone computers
- Digital repositories
- 3rd party services can provide advice and hardware rental where needed

https://www.flickr.com/photos/adactio/13127134455

Case Study: AHDS History dataset

Deposited by children of noted researcher in 2006 & processed by GK

Documentation:

Accompanying notes in researcher's handwriting described a history DB they were working on in 1988.

Challenges:

- 5.25" disk drive was available
- Disk was failing, but managed to create a complete copy on 5th attempt
- Disk analysis revealed text content…

The author's short stories, not a dataset!

Result:

Not accessioned, but children were pleased

History database created on a Shelton Instruments Sig-Net, running CP/M 2.2.operating system in 1988 & saved to 5.25" disk

http://www.old-computers.com/museum/computer.asp?st=1&c=810

Check completeness

What does the creator intend to provide?

- Data
- Documentation
- Research instruments

What have they actually provided?

- Some data
- Creation software & random files
- Personal music collection?
- **Request a file manifest:**
  - Filename
  - Description
  - Format

https://www.flickr.com/photos/kyngpao/14455832915

Case Study: Early English Books Online

Collection of 125,000 early printed books deposited for preservation:

- XML files, scanned TIFFs & PDFs for each page
- Well structured & labelled
- Print-out of file listing provided

Problems:

- Hard disk was failing
- XML output from Content Management system - incomplete header & missing schema
- 30% of files referenced in XML were missing

Solution:

- Obtained schema & missing files (but took a long, long time)

Render Data's content

File formats

Reflect tools & software available at point of creation:

- Information content
- Contextual information (documentation/metadata)

Organisation structure

Reflects intrinsic relationships:

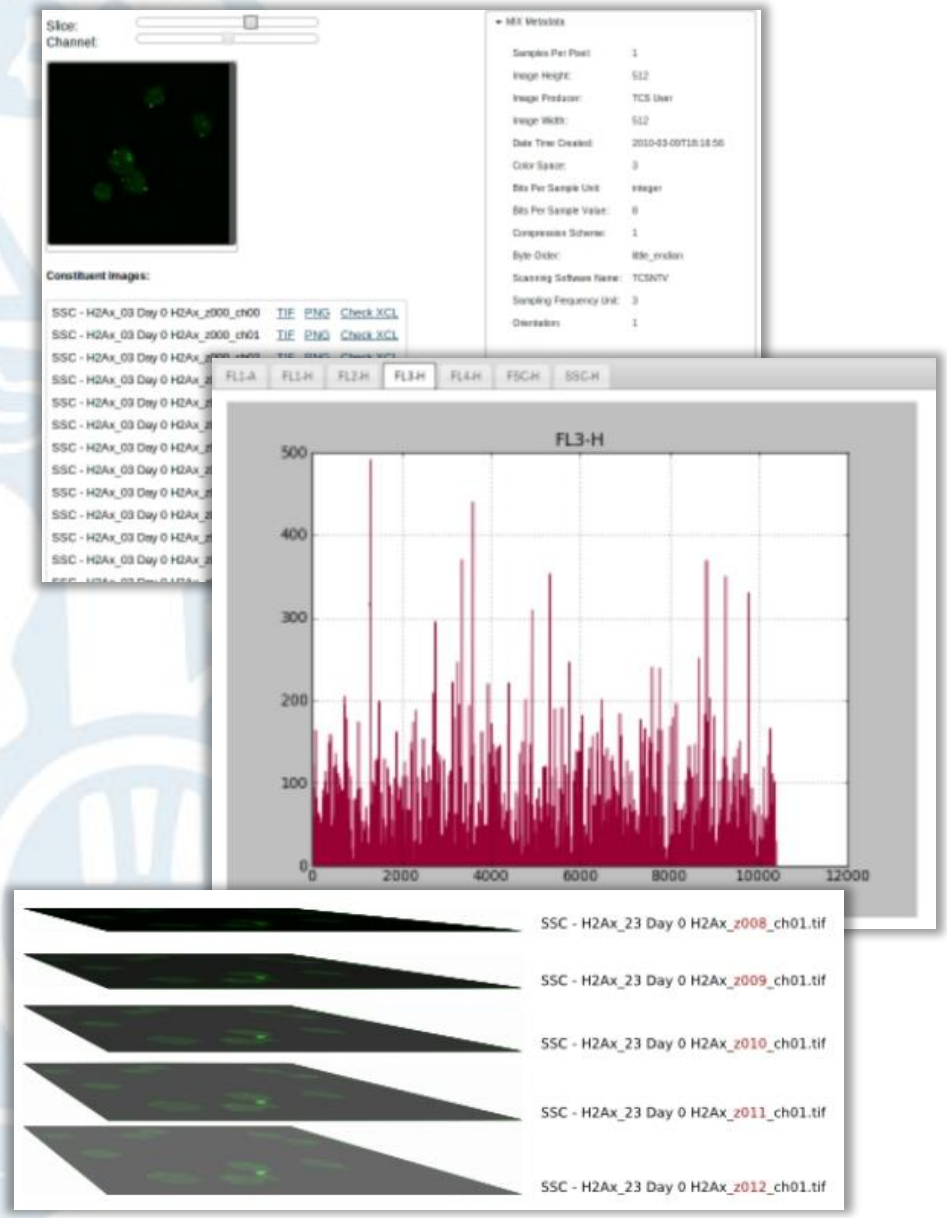
- Filenames
- Directory structure

Solution

- Specialist software may be required to access
- Liaise with data creators

https://www.flickr.com/photos/hawksanddoves/83818392

Case Study: Scientific dataset



USB stick of LSHTM dataset containing:

- *FCS2.0* - tabular data outlining experiments to count cells, sort them & identify biomarkers
- *Leica Experiment Collection* - .lei library file & associated images with embedded metadata

Challenges:

- Domain & proprietary formats
  - FITS (file) provides limited info on .lei
  - FCS not recognised
- Complex relationship in Leica experiment - recorded in filename & internal manifest

(partial) Solution

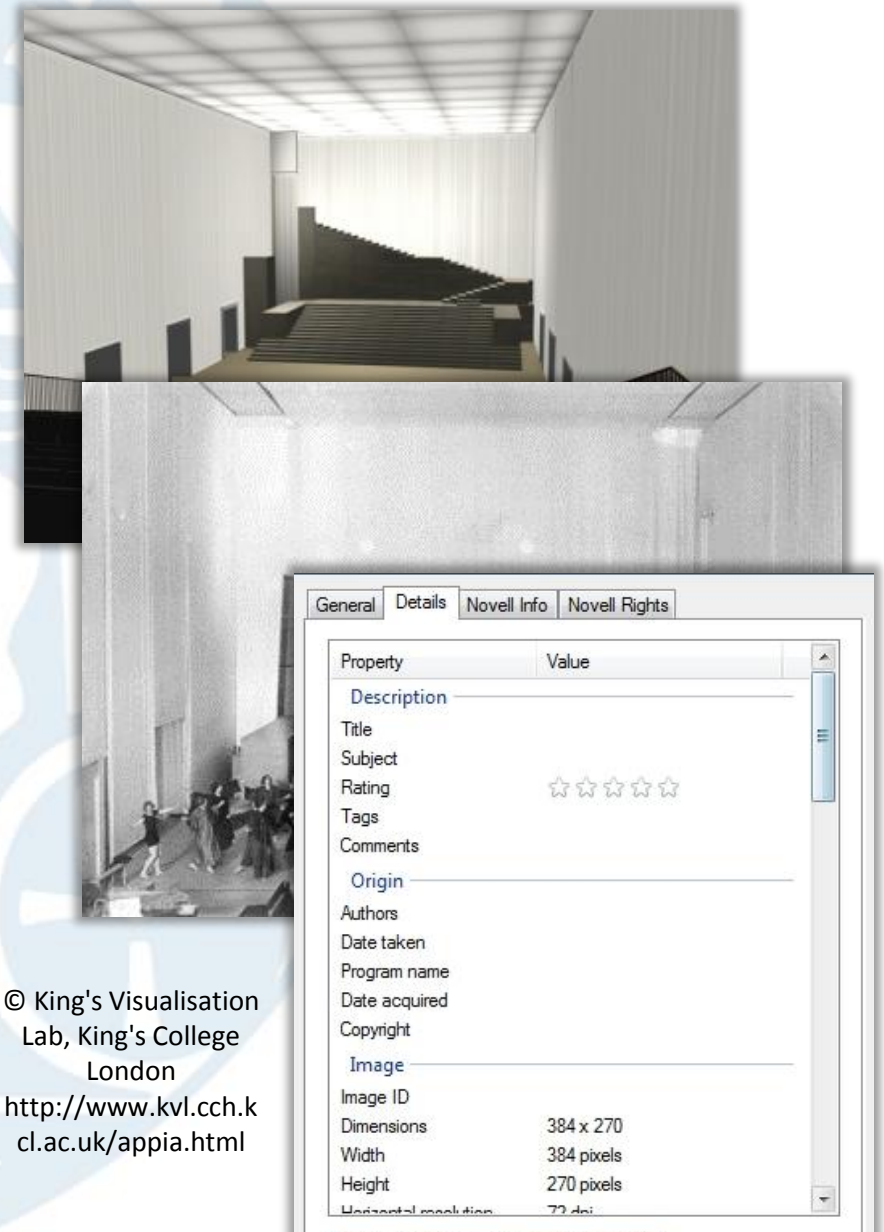
- Store files as-is
- Obtain text output of FCS files
- Analyse using open source tools

Understand data

- 17th-18th Century Enlightenment built on information sharing
- Openness & transparency essential for academic research
  - Evidence of activity
  - Open to scrutiny & replication
- Can you establish who, what, where, when & how?
- *How much documentation can only be found in the data creator's head?*

https://www.flickr.com/photos/domiriel/5234590796

Case Study: Adolphe Appia



http://www.kvl.cch.kcl.ac.uk/appia.html

Warwick Uni. School of Theatre Studies modelled performance space of Appia's Festspielhaus at Hellerau.

Collection deposited on several CDs:

- Digitised photographs of 1991 performance
- VRML 3D models of performance space
- Videos of 3D models in .mov format
- Documentation & Metadata

Problem

- Data well structured & extensive documentation, but descriptive metadata for images was missing

Solution:

- Descriptions added to file attributes, which were being removed when written to disc
- Output file attributes to text file
- Compressed files and copied to disk

Final thoughts

1. Analyse your needs & capabilities

- What can you do with resources?
- Where is investment needed?

2. Inform users of your expectations from the outset

- File formats
- Documentation
- File structure & naming conventions
- Permissions

3. Help them to fulfil expectations

- Advice and guidance

http://www.keepcalm-o-matic.co.uk/p/keep-calm-and-curate-41/