Archives School: Access

Sharing what you have

Hello!

I am Dr Jo Pugh

I am Digital Development Manager at The National Archives.

You can find me at @mentionthewar

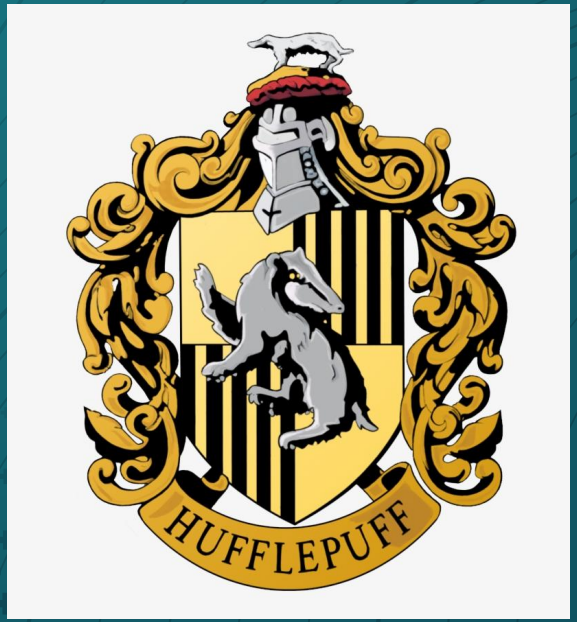
Webinar housekeeping

- Please mute your mic when you're not speaking
- I'll stop frequently for voice questions but feel free to ask a question in the chat at any time
- Files for this session can be found at: https://github.com/mentionthewar/Archives-School-Access

1. Introduction

What is this access we're giving?

Archivists: Hufflepuff or Ravenclaw?



Archivists: Hufflepuff or Ravenclaw?

Hilary Davidson
@FourRedShoes

At dinner last night I somehow found myself assigning Hogwarts houses to the GLAM sector.

I broke it down as:

Galleries - Slytherin
Libraries - Ravenclaw
Archives - Hufflepuff
Museums - Gryffindor.

Discuss.

5:44 AM · Oct 12, 2019 · Twitter Web App

65 Retweets **356** Likes

Archivists: Hufflepuff or Ravenclaw?

Hilary Davidson
@FourRedShoes

At dinner last night I somehow found myself assigning Hogwarts houses to the GLAM sector.

I broke it down as:

Galleries - Slytherin
Libraries - Ravenclaw
Archives - Hufflepuff
Museums - Gryffindor.

Discuss.

5:44 AM · Oct 12, 2019 · Twitter Web App

65 Retweets **356** Likes

Sarah Hegge @SarahLClothes · Oct 12, 2019
Replying to @FourRedShoes
I would very much switch libraries and archives. The librarians in my masters program were all the most helpful, kind people ever. The archivists (of which I was one) were hard working and smart, but in a more analytical way.

1 1 25

Colleen Theisen @LibColleen · Oct 12, 2019
Agree.

6

What is access?

Access is an interaction between:

- Your institution
- Your data
- Your system/interface
- Your users


Digital preservation is nothing more than long term digital access.

Some access principles

- We cannot access what we have not preserved
- Access is the eventual goal of preservation
- Access must never be a barrier to preservation ('Get it in a system')
- Access controls are easy to implement (assessing where to apply them may be more difficult…)
- Considering born digital access can supercharge access systems in a way that will benefit paper records and all researchers
- Cataloguing is now cataloguing for the web – end of.

"But look, you found the notice, didn't you?"

"Yes," said Arthur, "yes I did. It was on display in the bottom of a locked filing cabinet stuck in a disused lavatory with a sign on the door saying Beware of the Leopard."

Douglas Adams, 'The Hitchhiker's Guide to the Galaxy', 1979

MR ‘ACCESS RIGHT’

VS

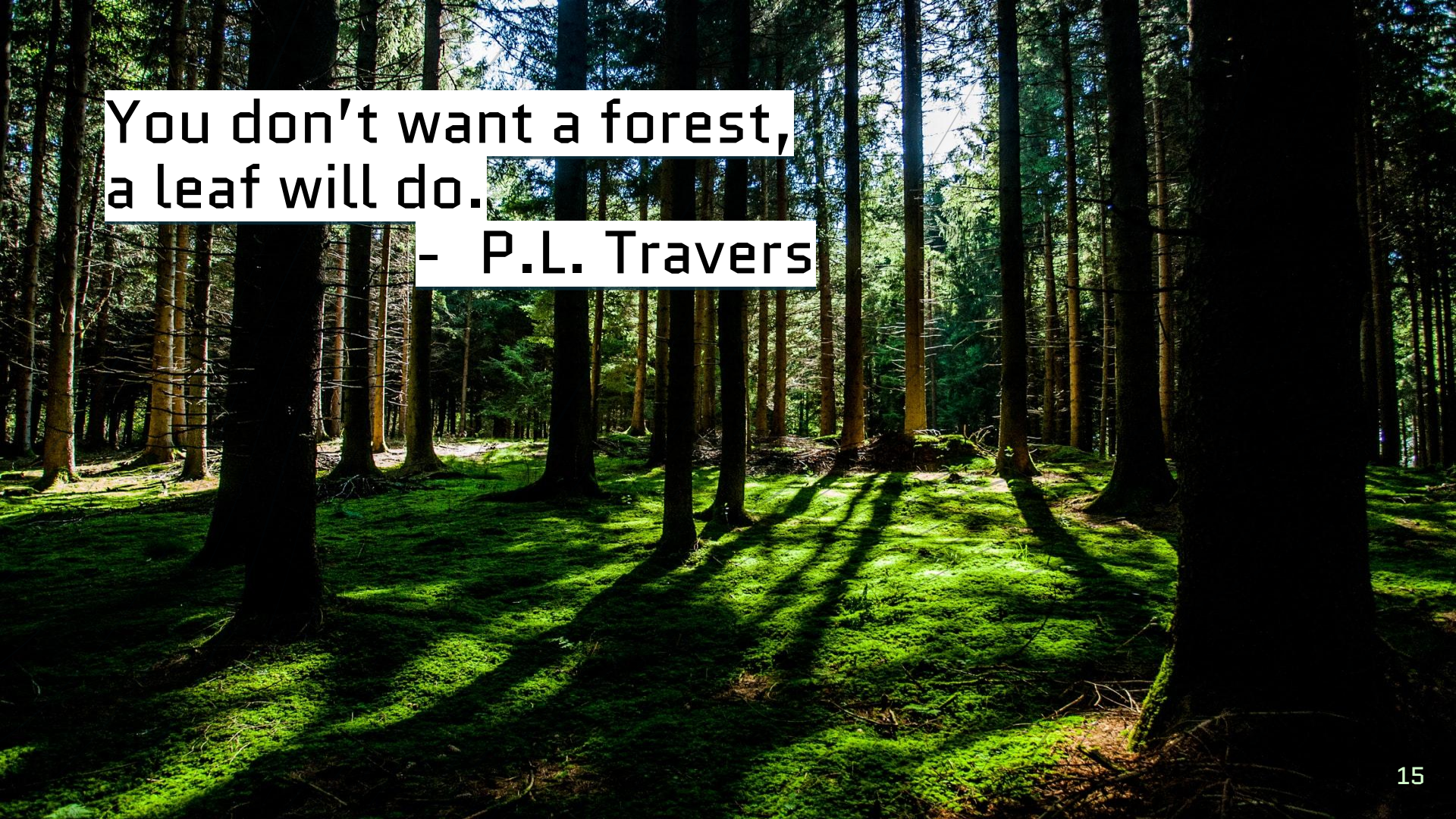
MR ‘ACCESS RIGHT NOW’

The access vicious cycle

- You don’t provide access (or access is hard)
- So you don’t really talk about accessing your digital records
- So no one looks at them
- So you cannot make a case to improve access
- Rinse repeat…

2. Access today

What's the problem with our current systems?


You don't want a forest,
a leaf will do.
- P.L. Travers

We used to want more information, faster...

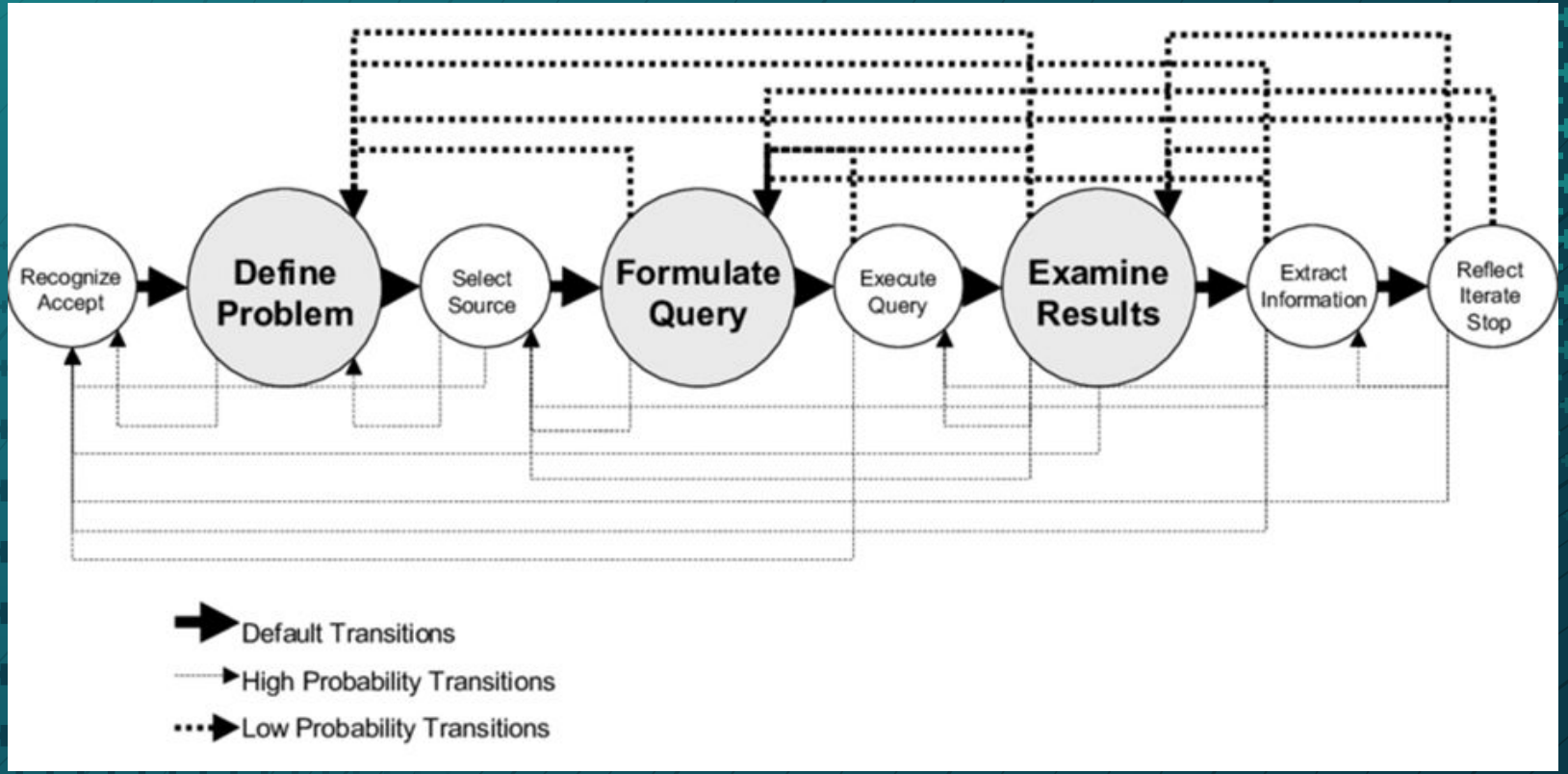
But successful interfaces today are systems that either show us less or help us make sense or construct meaning from large datasets.

Catalogues do neither.

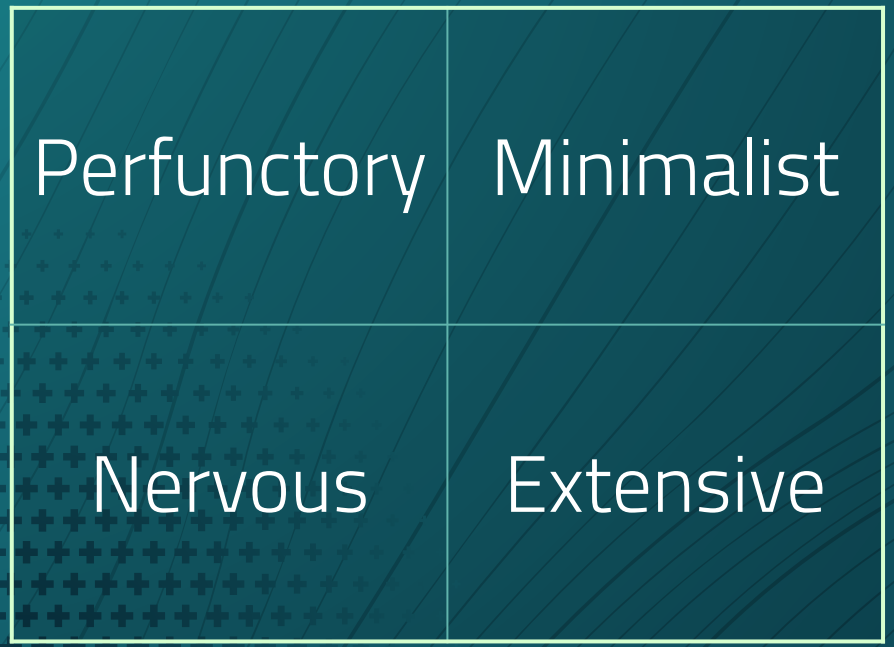
Browsing (Bates, 2007)

1. Glimpsing
2. Selecting/Sampling
3. Examining
4. Acquiring OR Abandoning

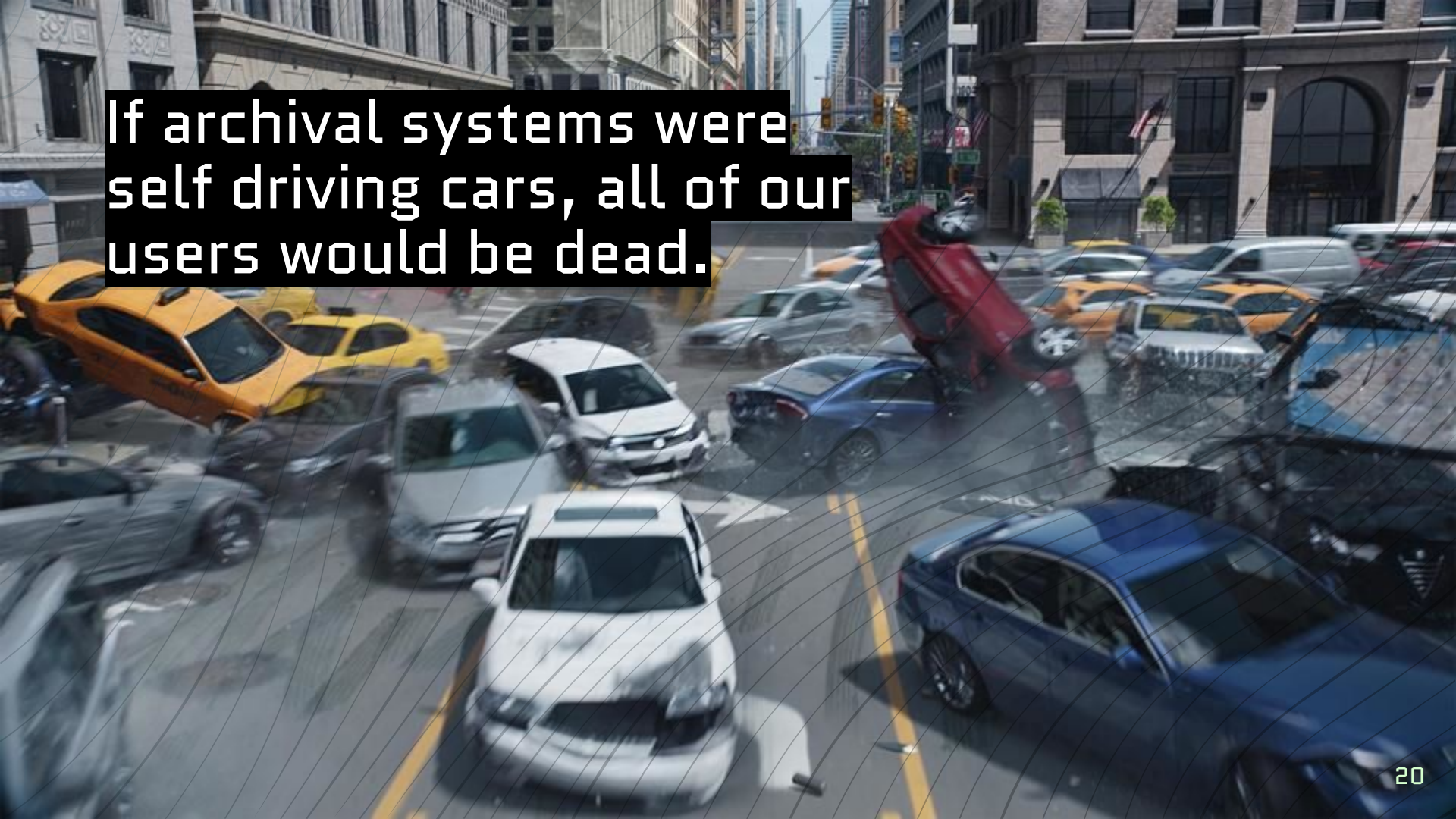
Search (Marchionini, 1995)



Satisficing



Matrix of search depth
(Mansourian and Ford, 2006)


If archival systems were self driving cars, all of our users would be dead.

Disorientation (Pugh, 2017)

1. I often felt I didn't know where to go next.
2. I felt I needed help.
3. The search results were difficult to understand.
4. I found myself going round in circles.
5. I found the search system confusing.
6. I often felt lost during the session.
7. At the end of the search session I felt uncertain.
8. The document descriptions didn't tell me enough to know if what I was seeing was really relevant.
9. I wasn't sure whether what I was looking for was in the collection or not.
10. I found it difficult to keep track of what I was finding.
11. By the end, I was running out of ideas for new queries.
12. I was frustrated because I knew what I wanted but I couldn't get to it.
13. I found it difficult to cope with the sheer volume of material I was looking through.
14. I knew what I wanted but I couldn't see how to get there.

What causes this?

- Too much information (look at the hot mess at: https://catalog.archives.gov/search?q=spoons)
- The hierarchy doesn't help
- Neither does the vocabulary
- There's no other hand-holding
- Linkages between records is poor
- Google indexing is a mixed blessing
- Calls to action to access records are often weak

Cataloguing for the web

- **Do** repeat elements
- Focus on entities (not document forms)
- Avoid complex structures
- Use persistent, linkable and human readable identifiers (especially for creators)
- Structured (or semi-structured) data is better than unstructured data

Schemas: structures in common

PREMIS

PREservation Metadata: Implementation Strategies. Data dictionary to turn OAIS into “implementable semantic units”. Developed by OCLC/RLG.

v.3.0 , 2015 (226 pages)

METS

Metadata Encoding and Transmission Standard for capturing the hierarchy of digital library objects. Developed by Library of Congress / Digital Library Federation.

v.1.6, 2010 (148 pages)

3. Emerging access

What is future access likely to look like?

Discussion:
DLF Levels of Born Digital Access

Level	Accessibility	Description	Researcher Support & Discovery	Security	Tools
1	Researchers are provided with information on the accessibility of born-digital materials	Provide required descriptive elements for a collection-level record and at least one descriptive note about the processed digital materials	Support basic access to and duplication of content. Have a knowledgeable staff member to provide assistance	Provide access to open, authentic, virus-free content on a dedicated on-site public access computer with security measures implemented based on local policies	Provide local access via an on-site public access computer with open and common software to render widely used file formats

The #1 person needing access…is you

Get yourself a tool which will unlock legacy formats

- Quick View Plus
- Zamzar *(web based)*
- Autodesk Viewer (3D files, *web based*)
- VLC (Video formats)

Shneiderman (1996)

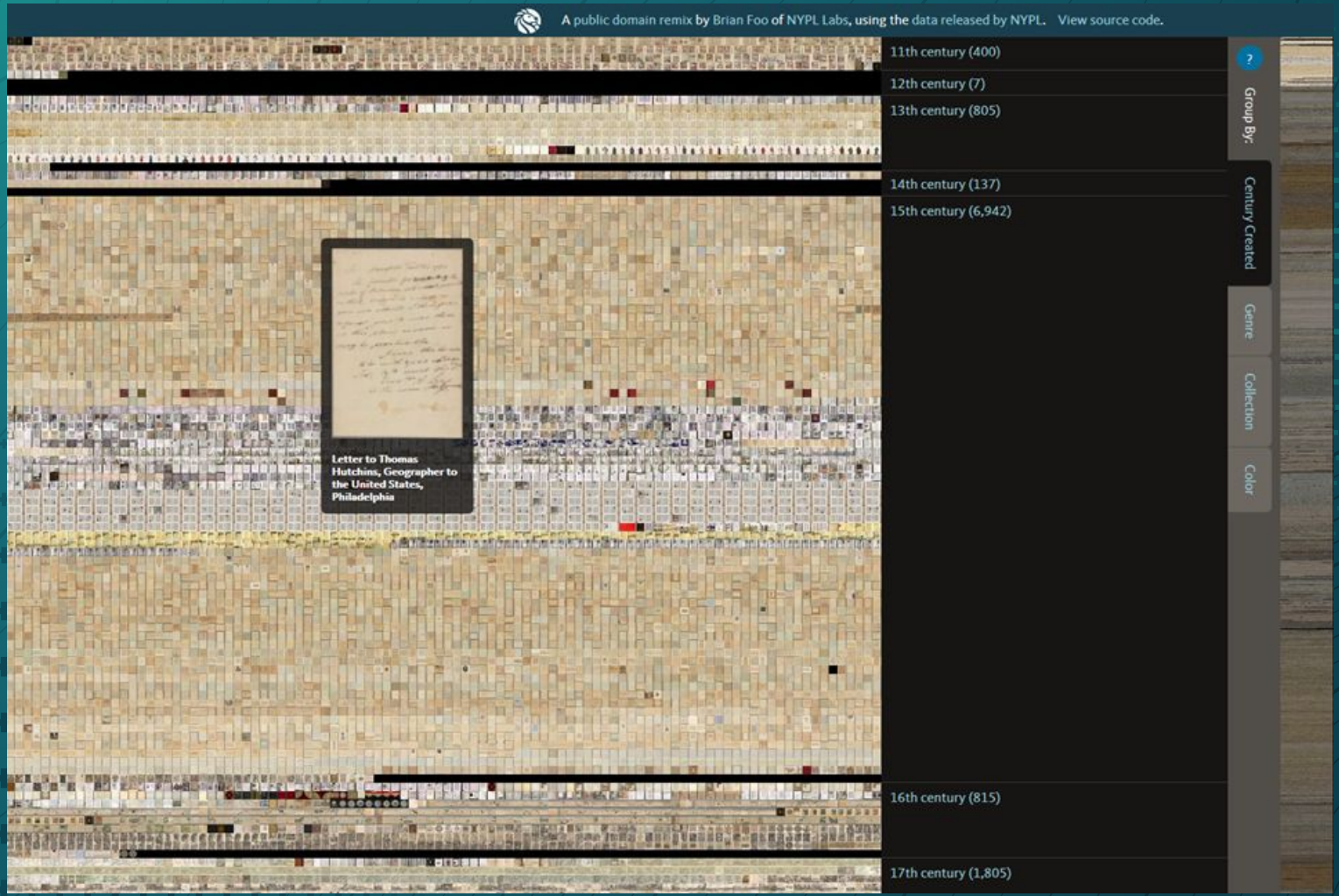
“The [visual] information seeking mantra”

1. Overview
2. Zoom and filter
3. Details on demand

Prospect

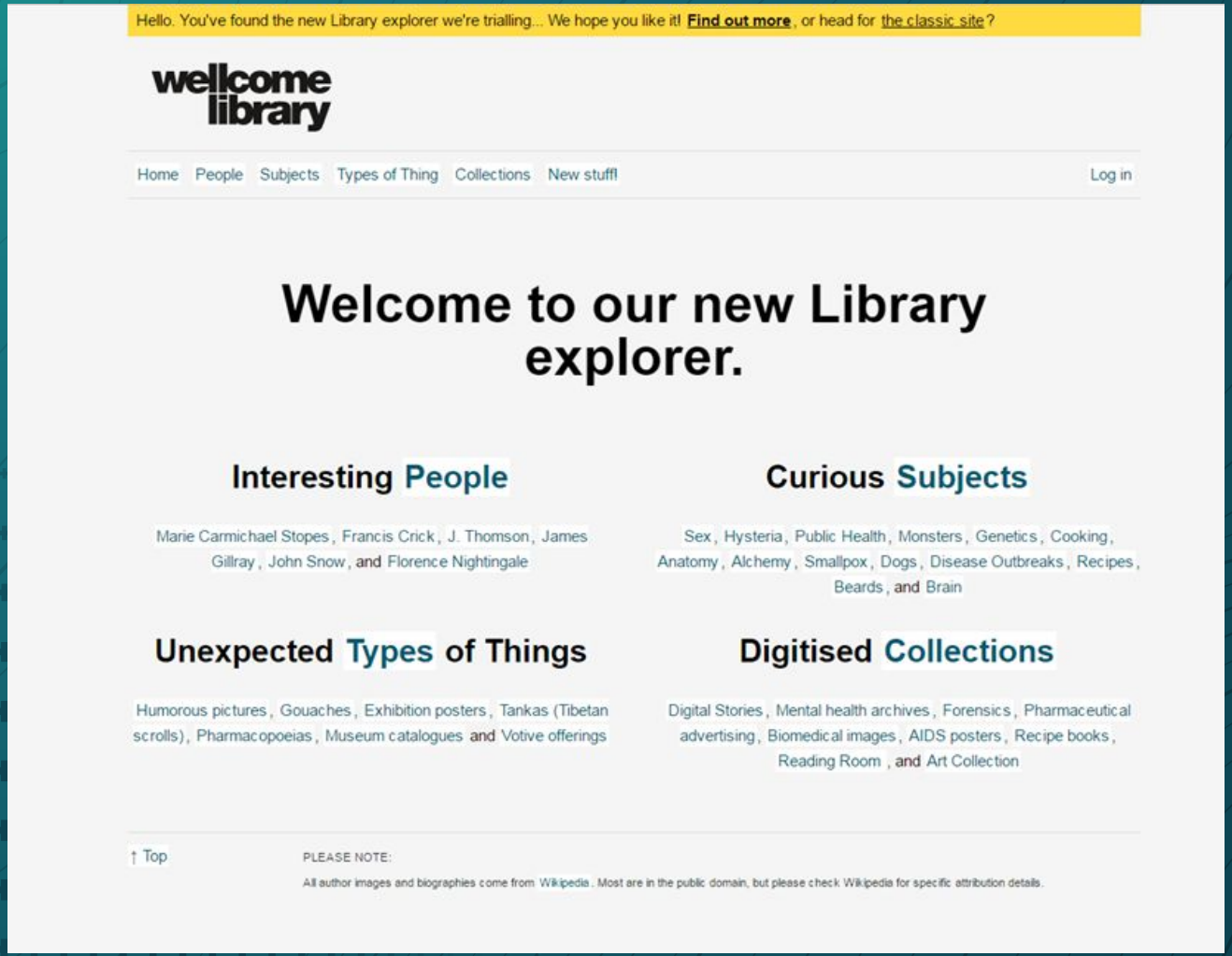
"A view of the world where enough information is available for the perceiver to understand the terrain and have a sense of what it affords, without seeing all the details."

- Ruecker et al, 2011



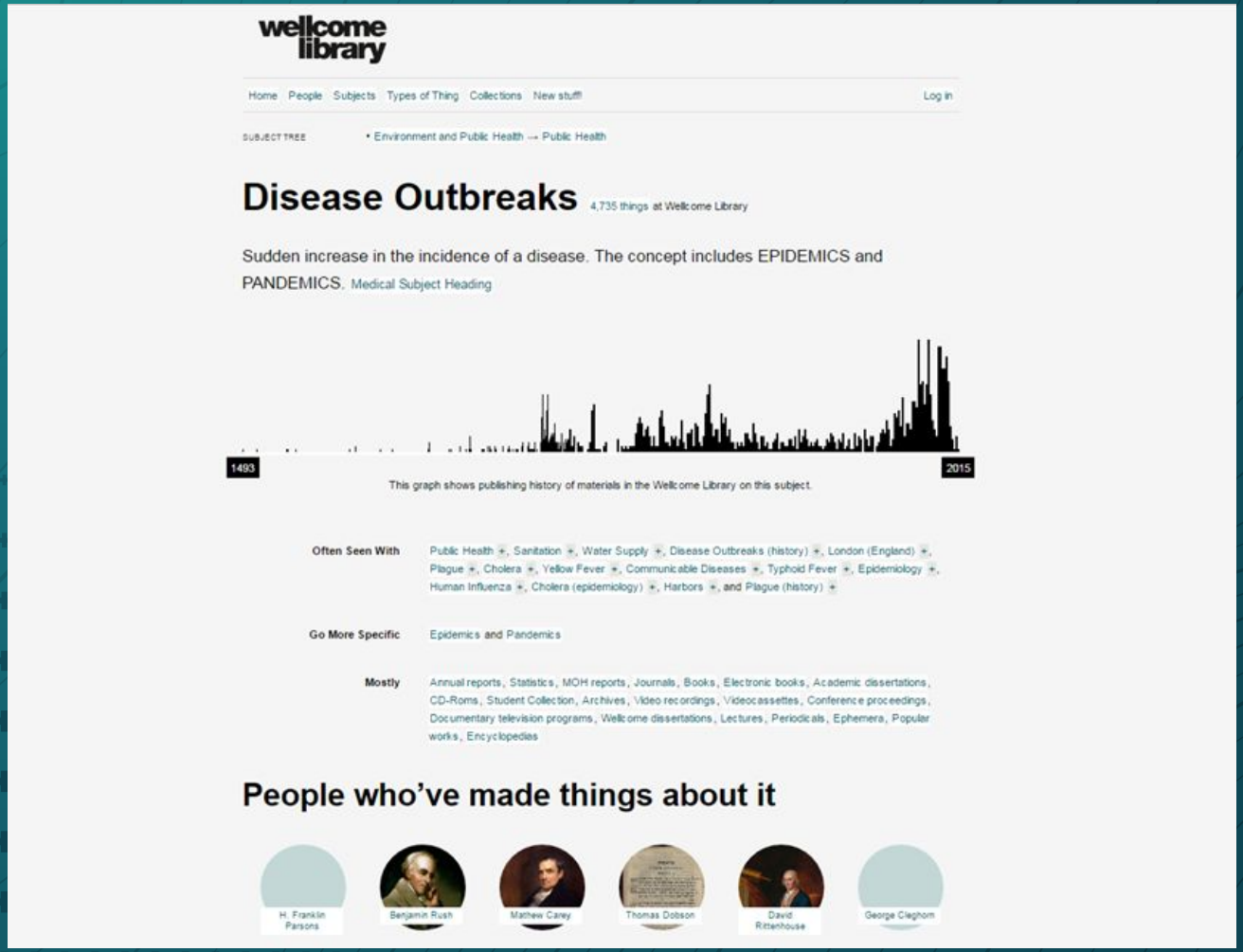
New York Public Library Labs

http://publicdomain.nypl.org/pd-visualization



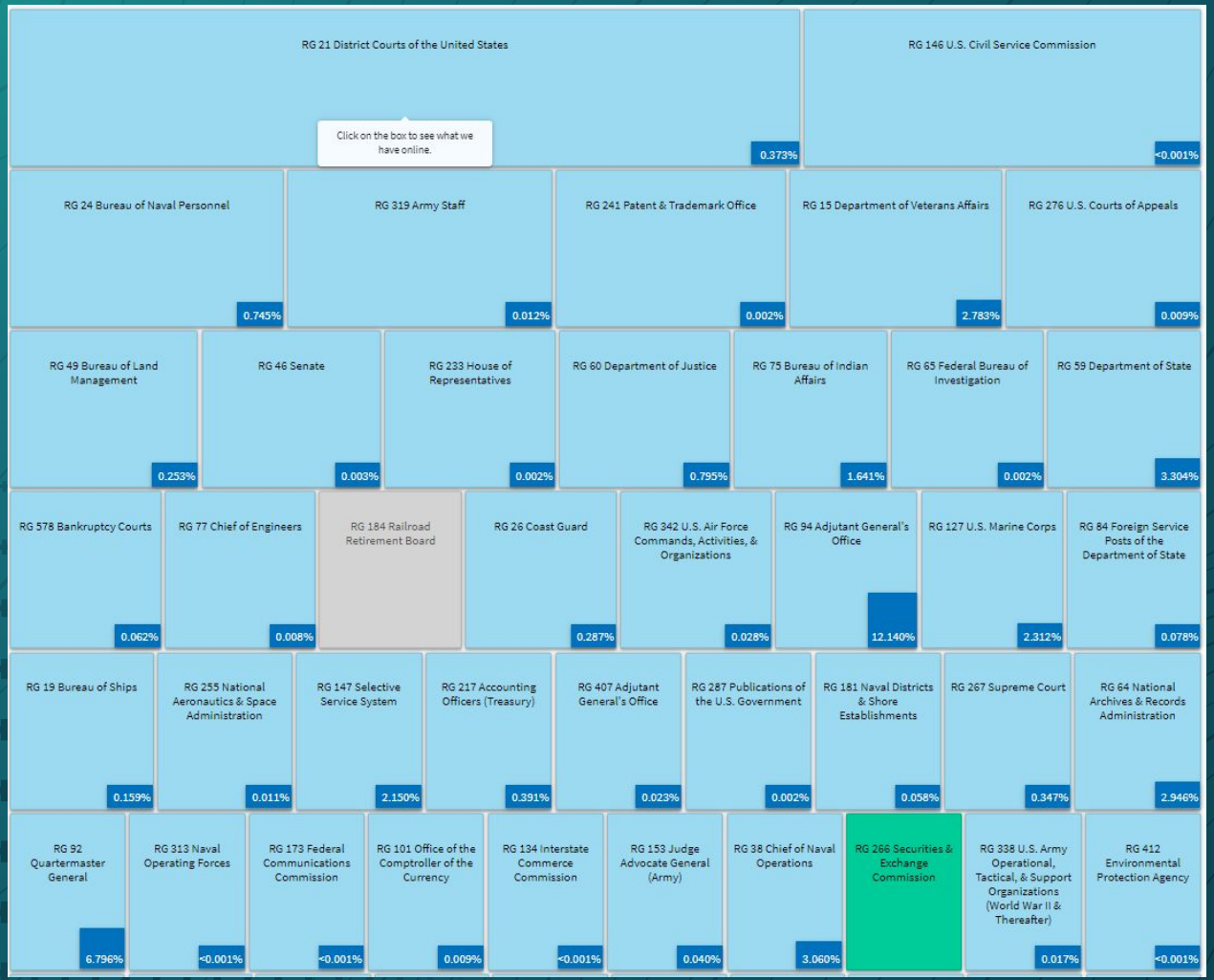
Wellcome / Good, Form & Spectacle

http://alpha.wellcomelibrary.org



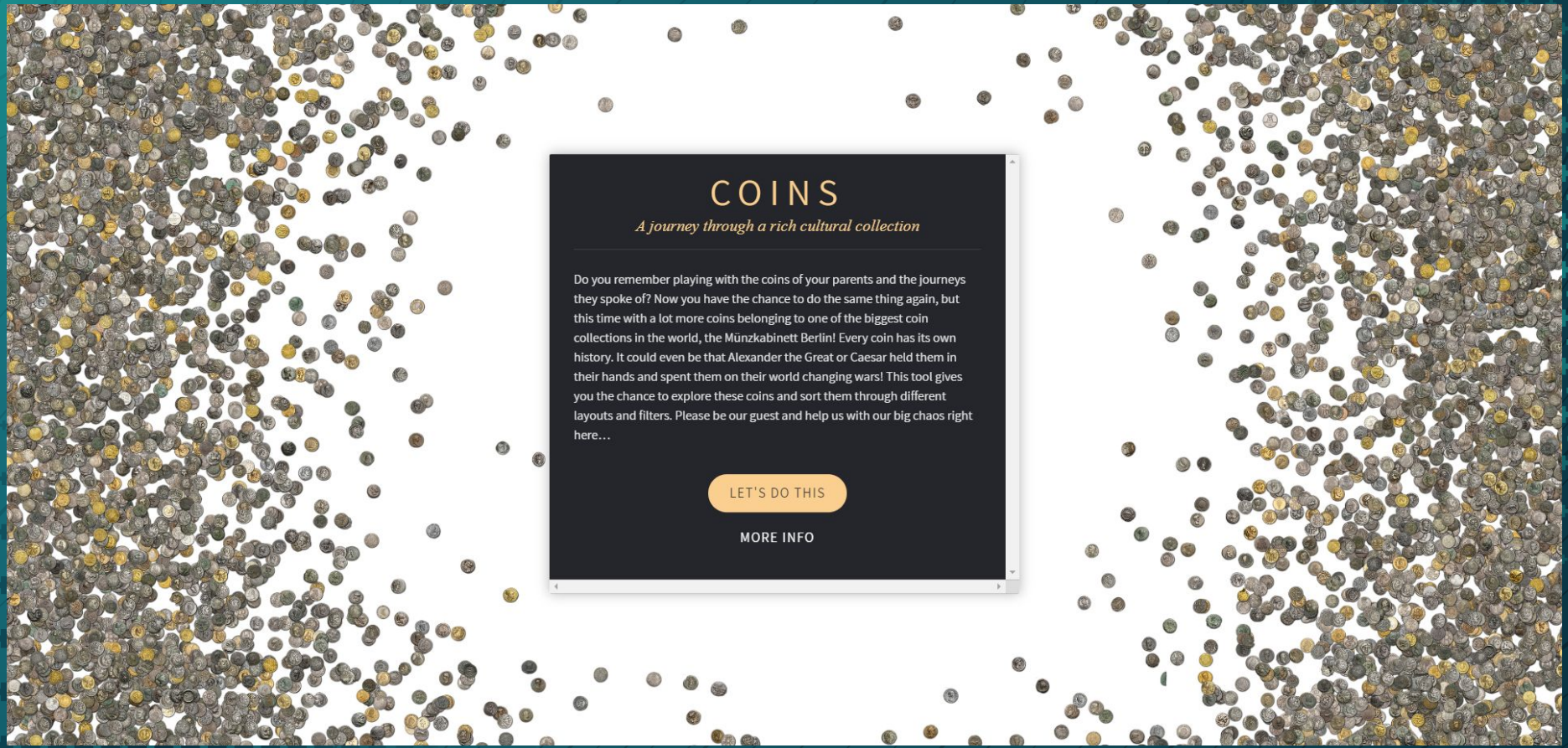
Wellcome / Good, Form & Spectacle

http://alpha.wellcomelibrary.org



National Archives and Records Administration

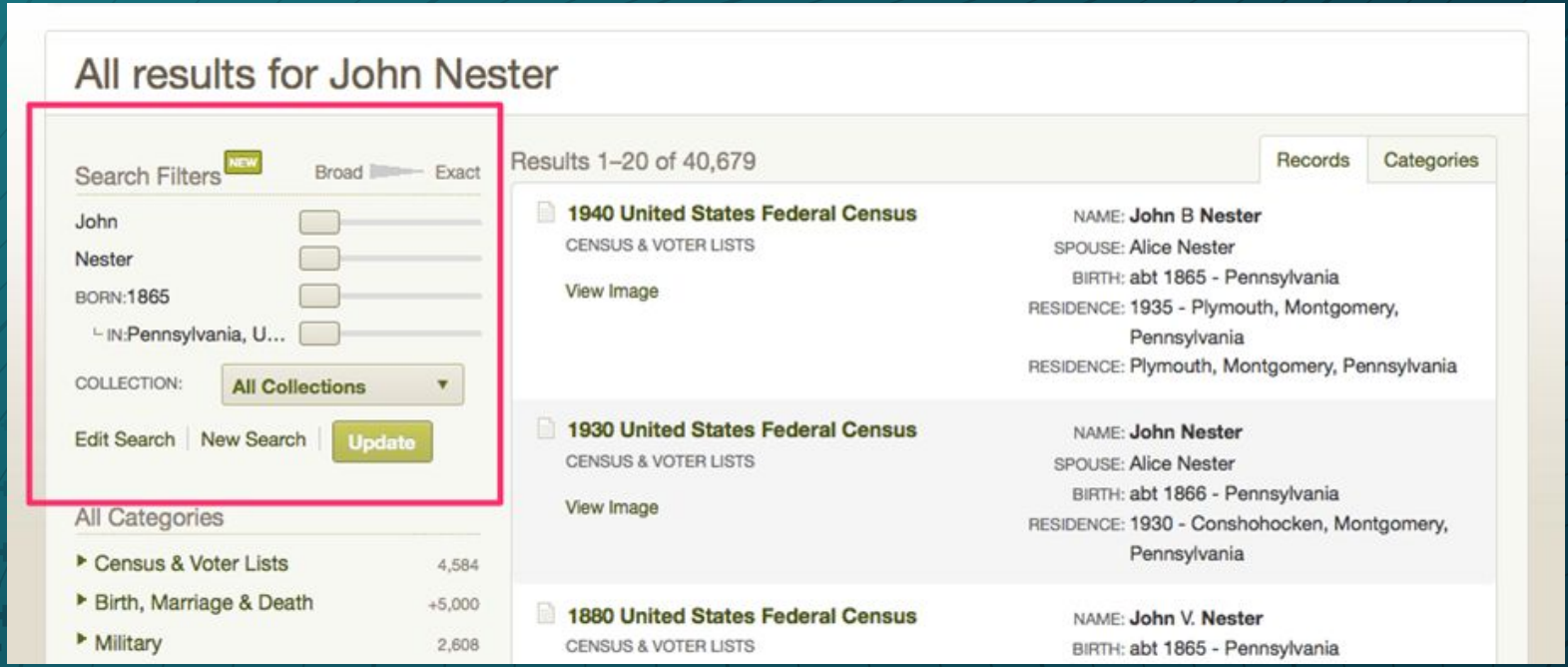
https://www.archives.gov/findingaid/explorer



UoAS Potsdam

https://uclab.fh-potsdam.de/coins/

Giving Control (direct manipulation)



Visual Cues

National League of the Blind and Disabled » Operations » Relations with other organisations

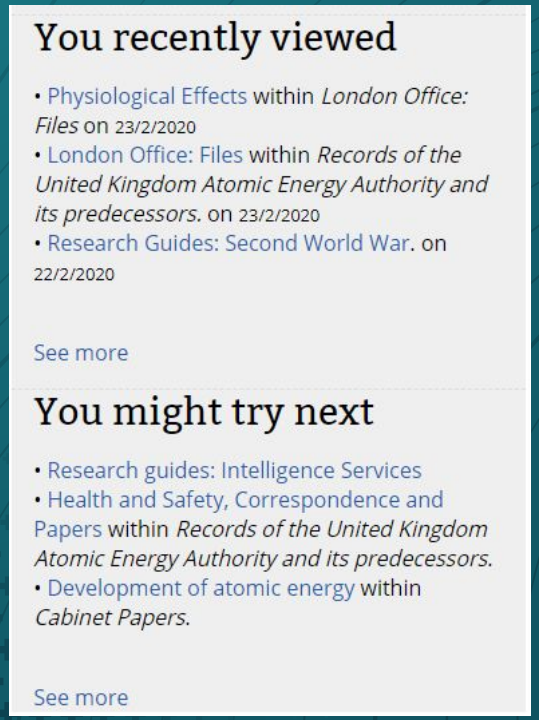
National Union of Mineworkers, 1975

Karta/ritning

Kart over Tromsé amt ... [Skala 1:4000.000. Utgiven av Norges geografiske opmaaling 1886. (bl. 1) - Kartor i skala 1:200.000 över nordost-och-västliga resp. sydost-och-västliga delar. (bl. 2-5).] [5 bl. papper, varav 2 klistrade på väv 61x84; 65x89; 63x83; 60x82,5. Blad nr 3 utgivet 1874 av Den geografiske Opmaaling.]

Skiljedomstolen i renbetesfrågan 1909

Textual cues



You recently viewed

- Physiological Effects within *London Office: Files* on 23/2/2020
- London Office: Files within *Records of the United Kingdom Atomic Energy Authority and its predecessors.* on 23/2/2020
- Research Guides: Second World War. on 22/2/2020

See more

You might try next

- Research guides: Intelligence Services
- Health and Safety, Correspondence and Papers within *Records of the United Kingdom Atomic Energy Authority and its predecessors.*
- Development of atomic energy within *Cabinet Papers.*

See more

- Onward paths
- Inline help
- Defining difficult terms
- Richer descriptions

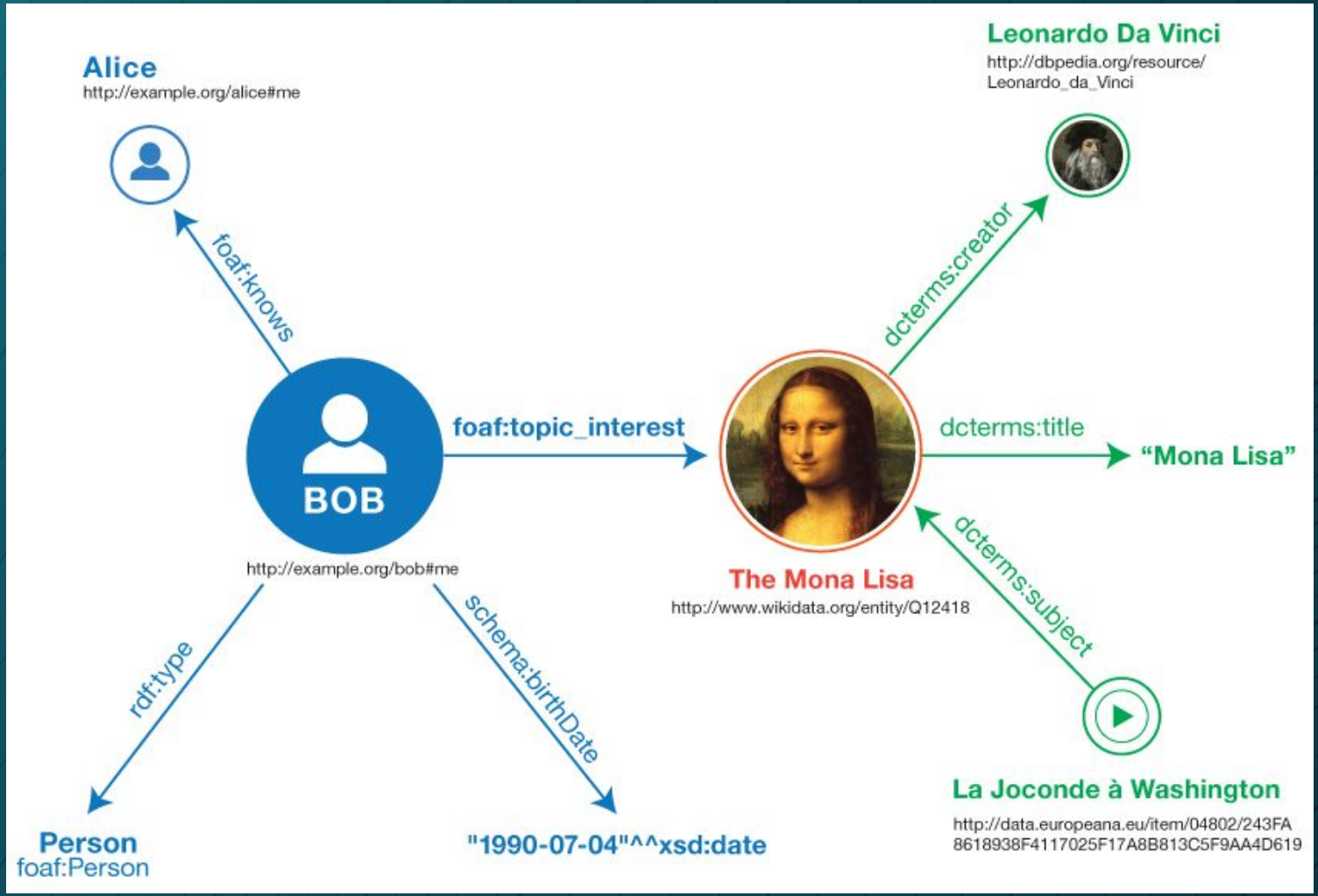
7 Pillars of Metadata (Hillyard, 2018)

1. Legacy
2. Primary
3. Secondary
4. Supplementary
5. Derived
6. Control
7. Meta(!)

Graph catalogues

- Nodes, edges and properties
- Express conceptual relationships
- Different ways of expressing relationships (RDF vs LPG)
- Different rules for the types of relations (RiC, Matterhorn RDF)

All catalogues will be structured like this soon


Alice
http://example.org/alice#me
Leonardo Da Vinci
http://dbpedia.org/resource/
Leonardo_da_Vinci
foaf:knows
dcterms:creator
BOB
http://example.org/bob#me
foaf:topic_interest
dcterms:title
"Mona Lisa"
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:subject
rdf:type
schema:birthDate
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA
8618938F4117025F17A8B813C5F9AA4D619
Person
foaf:Person
"1990-07-04"^^xsd:date

Access Nirvana: how do we get there?

- Ask for more from vendors
- Make use of specialised tools (e.g. EPADD) and identify strong features in available systems
- Think about who we are providing access for and for what purpose

4. User research

How do we know what we need to offer?

How do we learn about user behavior?

- Directly
  - Observation
  - Testing (guerilla to highly controlled)
- Indirectly
  - Surveys
  - Data analysis (e.g. log files)

“Digital information is really just people in disguise.”

Jaron Lanier, ‘Who owns the future?’, 2013

Things we don't know about access

- What characteristics of a born digital file do users most care about?
- What is the ideal length for a catalogue description?
- What features in a description maximise findability?
- How do we communicate access restrictions and next steps clearly?
- In an ideal world, how would researchers prefer to work with born digital content?

Sloyan et al (2018)

"Many researchers have limited experience of using born-digital archives in their research and so can only speculate about how they would like to access and use such records."

"All found it difficult to consult the two [catalogue and files] in tandem and easily switch from viewing the digital files to finding the corresponding online catalogue record."

Sloyan et al (2018)

"Participants…divided between those who enjoyed being able to browse and explore the directories and those who were frustrated…This was caused partly by being asked to navigate a personal filing system that did not necessary chime with their expectations, but also by the **use of opaque filenames** that did not reveal the content. Most participants saw the value in preserving the original directory structure and filenames…but they also **wished to have alternative arrangements and filenames** that enabled quicker searching and easier identification of topics and data. A desire was also expressed for the **catalogue to be more helpful in identifying the location of specific themes, subjects and file formats.**"

Study design and techniques

- Within and between subjects
- Think aloud protocol
- Testing scripts
- Record sessions
- Heuristic evaluation

Nielsen and Molich

https://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html

“The system should speak the user’s language”

“Consistency and standards”

“Recognition rather than recall”

“Help users…recover from errors”

Task: Heuristic Evaluation

Choose one of two prototypes:

https://alpha.nationalarchives.gov.uk/collectionexplorer/

https://alpha.nationalarchives.gov.uk/scopehistogram/

1. What do you think the purpose of the interface is? (What goal is it helping users accomplish?)
2. How is it trying to help?
3. How does it perform against one of Nielsen's heuristics?

5. Delivering a golden minimum

Can we do enough to make Trevor Owens happy?

"If we took this approach seriously for working with digital content…[we] would acquire content, create a short collection level description, then produce a container list (which as digital files and folders provide the instantaneous ability to provide a list is rather trivial) and then…simply upload the files to a directory on the web…**This golden minimum should be the default**…"

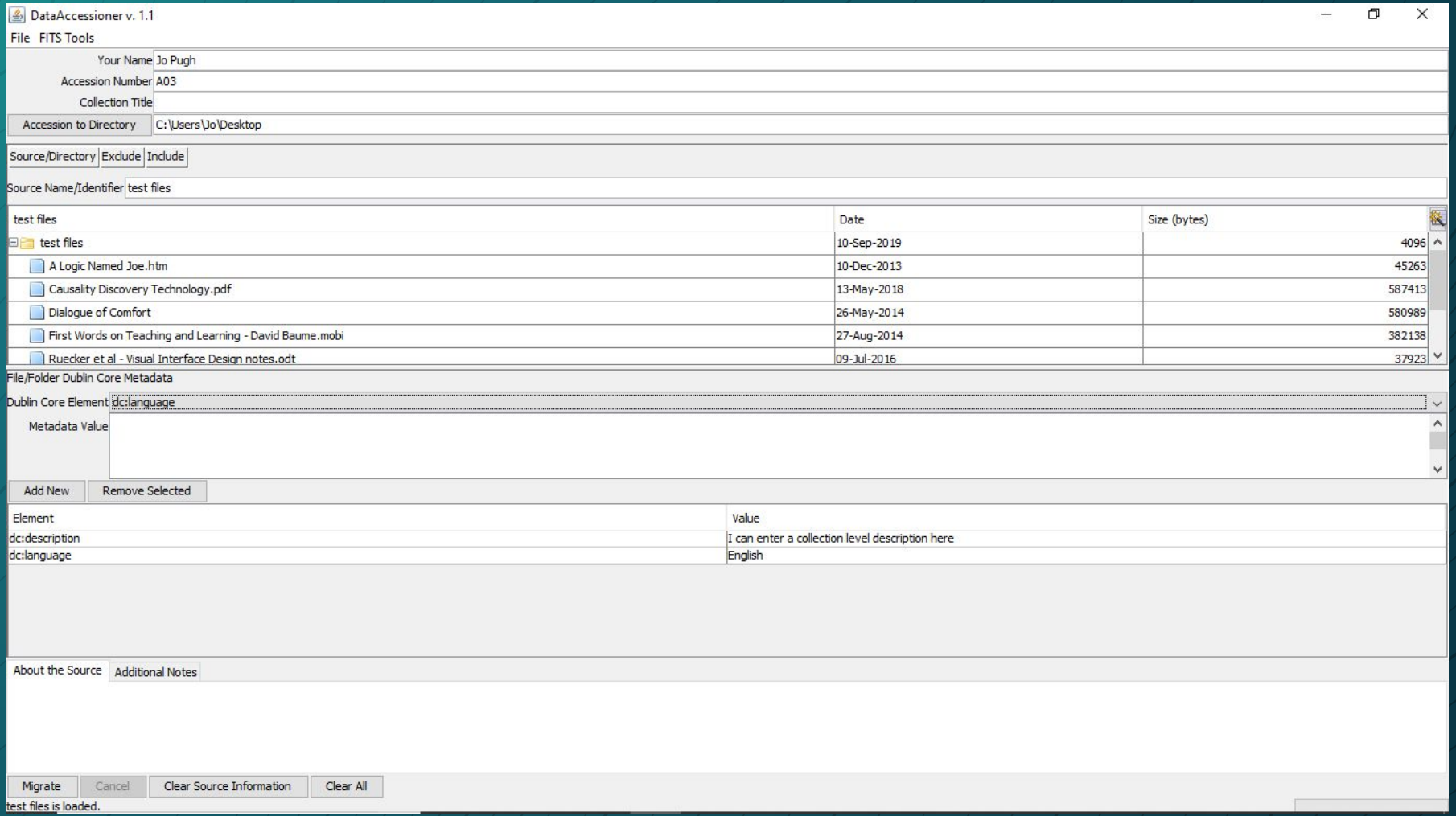
Trevor Owens, 'Theory and Practice of Digital Preservation', 2018

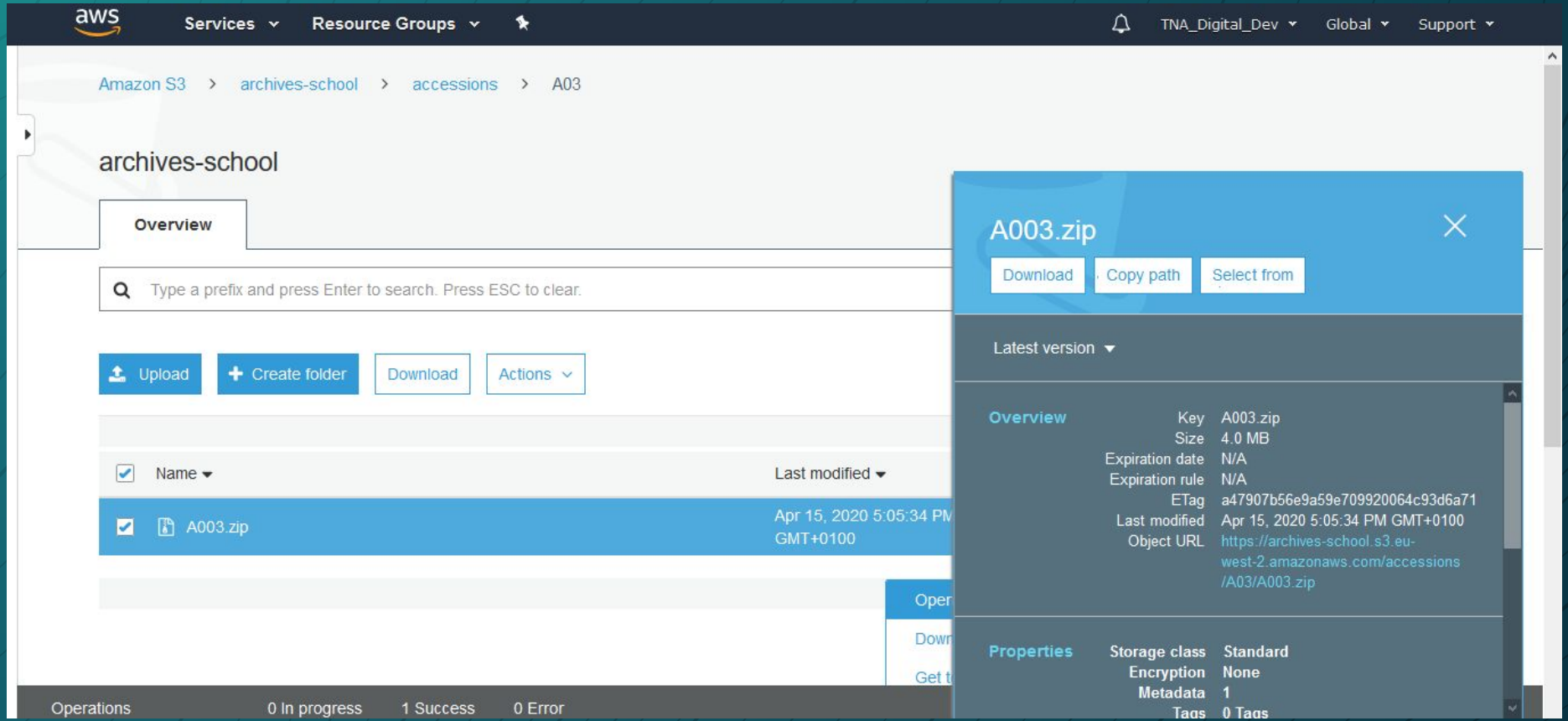
"From my observations over the years, my sense is that **almost no one is doing this** for digital materials."

Trevor Owens, 'Theory and Practice of Digital Preservation', 2018

Workflow

1. Run Data Accessioner over collection to generate ‘bag’ and metadata
2. Put bag in (cloud) storage
3. Run transformation on metadata
4. Publish results to catalogue


DataAccessioner v. 1.1
File FITS Tools
Your Name Jo Pugh
Accession Number A03
Collection Title
Accession to Directory C:\Users\Jo\Desktop
Source/Directory Exclude Include
Source Name/Identifier test files
test files Date Size (bytes)
test files 10-Sep-2019 4096
A Logic Named Joe.htm 10-Dec-2013 45263
Causality Discovery Technology.pdf 13-May-2018 587413
Dialogue of Comfort 26-May-2014 580989
First Words on Teaching and Learning - David Baume.mobi 27-Aug-2014 382138
Ruecker et al - Visual Interface Design notes.odt 09-Jul-2016 37923
File/Folder Dublin Core Metadata
Dublin Core Element dc:language
Metadata Value
Add New Remove Selected
Element Value
dc:description I can enter a collection level description here
dc:language English
About the Source Additional Notes
Migrate Cancel Clear Source Information Clear All
test files is loaded.


aws
Services
Resource Groups
TNA_Digital_Dev
Global
Support
Amazon S3 > archives-school > accessions > A03
archives-school
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
Download
Actions
Name
Last modified
A003.zip
Apr 15, 2020 5:05:34 PM GMT+0100
Operations
0 In progress
1 Success
0 Error
A003.zip
Download
Copy path
Select from
Latest version
Overview
Key A003.zip
Size 4.0 MB
Expiration date N/A
Expiration rule N/A
ETag a47907b56e9a59e709920064c93d6a71
Last modified Apr 15, 2020 5:05:34 PM GMT+0100
Object URL https://archives-school.s3.eu-west-2.amazonaws.com/accessions/A03/A003.zip
Properties
Storage class Standard
Encryption None
Metadata 1
Tags 0 Tags

Task: PREMIS to ISAD(G)

- Where in ISAD(G) would you allocate the PREMIS fields in the sample output?
- Is there any information about the files that is not captured that should/could be?

```
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://dataaccessioner.org/schema/dda-1-1" name="Files associated with thesis research">
  <accession number="A003">
    <ingest_note>Files associated with thesis research transferred by Jo Pugh on Mon Jan 14 13:18:00 GMT 2019</inge
    <ingest_time>00:02:03.123464</ingest_time>
    <folder name="test files" last_modified="2019-01-14T13:17:08.767">
      <file name="A Logic Named Joe.htm" last_modified="2013-12-10T14:08:08.934" size="45263" MD5="10c4729131459d93
      <file name="Causality Discovery Technology.pdf" last_modified="2018-05-13T23:09:58.158" size="587413" MD5="23
        <premis:object xmlns:premis="info:lc/xmlns/premis-v2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          <premis:objectIdentifier>
          <premis:objectCharacteristics>
            <premis:compositionLevel>0</premis:compositionLevel>
            <premis:fixity>
              <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
              <premis:messageDigest>23989868710415bf2ba7af457c7f0804</premis:messageDigest>
              <premis:messageDigestOriginator>OIS File Information</premis:messageDigestOriginator>
            </premis:fixity>
            <premis:size>587413</premis:size>
            <premis:format>
              <premis:formatDesignation>
                <premis:formatName>Portable Document Format</premis:formatName>
                <premis:formatVersion>1.3</premis:formatVersion>
              </premis:formatDesignation>
              <premis:formatRegistry>
                <premis:formatRegistryName>http://www.nationalarchives.gov.uk/pronom</premis:formatRegistryName>
                <premis:formatRegistryKey>fmt/17</premis:formatRegistryKey>
              </premis:formatRegistry>
              <premis:formatNote>application/pdf</premis:formatNote>
```

```
# Transforming output from data accessioner into a form importable into Discovery via MYC

import re
import codecs
import untangle
import os

count = 0

output_file = open("new_collection.txt", "a") # a for append, w for write
path = '/Users/Jo/Desktop/Digital Preservation/Archives School/Accessioner to MYC/metadat

collection_creator = input("Who was the original creator of this collection?: ")
collection_url = input("At what URL is this collection available to download?: ")

print("Thinking...")

output_file.write("Level\t" + "Title\t" + "Creator\t" + "Description\t" + "Reference\t" +
                  "Related units of description (URL)\t" + "Note (fixity)\n")

# level uses folder followed by file
# title is file.name
# creator is user defined for fonds level only
# description we could leave blank but let's use the ingest note
# reference is the accession number with augmentation for each file/folder
# start date and end date are the last modified date twice. Take the first 10 characters
# Extent is 'Digital file' followed by PREMIS:Size (gives a number in bytes / by 1024 to
# physical description is the file format in premis:formatName
# note is the checksum - the file's MD5 value


for filename in os.listdir(path):
    filename_p = path + filename

    accession_file = codecs.open(filename_p, 'r', encoding='utf-8', errors='replace')
    xml = untangle.parse(accession_file)

    # Produce fonds level information

    level = 'fonds'
```

Level	Title	Creator	Description	Reference	Start Date	End Date	Extent	Physical description	Related units of desc	Note (fixity)
fonds	Files associated	Jo Pugh	Files associat	A003	20190114	20190114	8 digital files		www.myarchive.com	
file	test files	Jo Pugh		A003/1	20190114	20190114	8 digital files			
item	A Logic Named Joe.htm			A003/1/1	20190114	20190114	Digital file (44 Kb)	Hypertext Markup Language		MD5 checksum: 10c472
item	Causality Discovery Technology.pdf			A003/1/2	20190114	20190114	Digital file (573 Kb)	Portable Document Format		MD5 checksum: 239898
item	Dialogue of Comfort			A003/1/3	20190114	20190114	Digital file (567 Kb)	Plain text		MD5 checksum: 5f424b
item	First Words on Teaching and Learning -			A003/1/4	20190114	20190114	Digital file (373 Kb)	Hypertext Markup Language		MD5 checksum: 74efac
item	Ruecker et al - Visual Interface Design r			A003/1/5	20190114	20190114	Digital file (37 Kb)	OpenDocument Text		MD5 checksum: 4e77f1
item	Storm Harding - Jumping the Paywall.oc			A003/1/6	20190114	20190114	Digital file (1356 Kb)	OpenDocument Presentation		MD5 checksum: a892ae
item	Temporal Modelling.doc			A003/1/7	20190114	20190114	Digital file (1899 Kb)	Microsoft Word Binary File Format		MD5 checksum: 505d77
item	Yakel - Thinking inside and outside the			A003/1/8	20190114	20190114	Digital file (151 Kb)	Portable Document Format		MD5 checksum: 42422a

6. Basic sensitivity detection

How do we know what large digital collections contain?

Knowing what you have (again)

- DROID does an excellent job of telling us about the characteristics of files
- But what about their content?

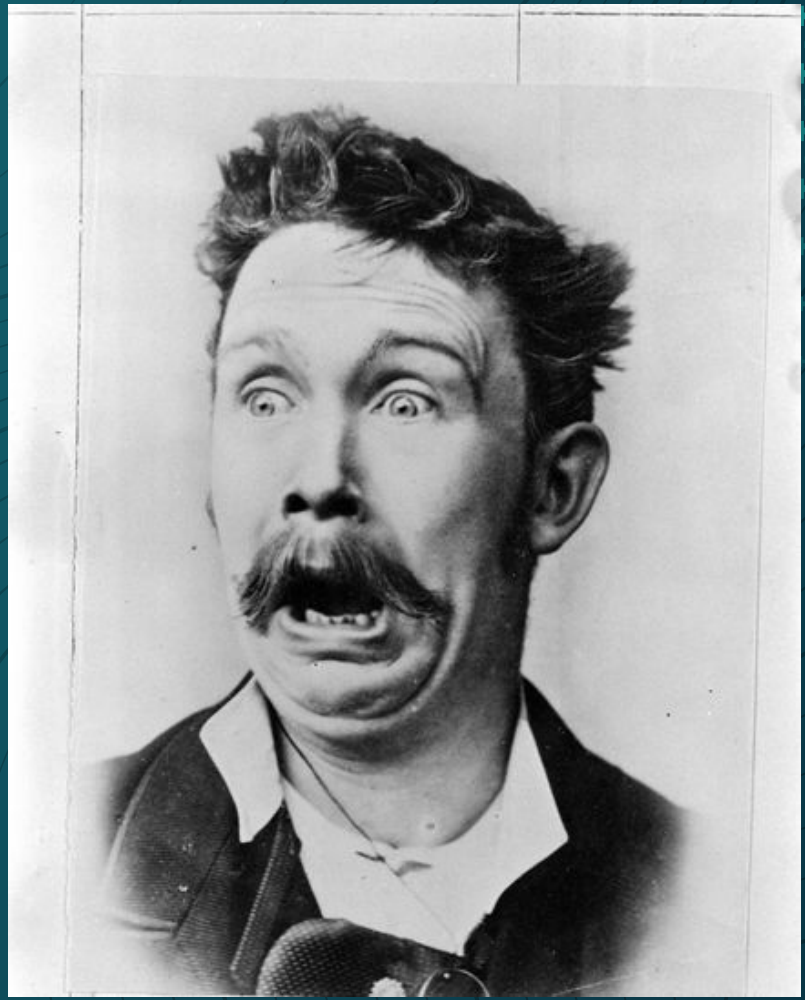
Data crunching for access

- Topic modelling
- Similarity matching
- ‘Significance’ matching (using another corpus)
- Catalogue metadata modelling

What makes access scary?

We need to have some way of assessing and managing risk.

How do we do this for very large numbers of files?




Information purgatory -
where records are neither
open nor closed.

Pattern matching with Regex

- We write rules to look for patterns that might indicate personal data (in a DPA context)
- We investigate high concentrations of sensitivity

Does this idea of writing rules sound like an algorithm? You're right!

Regular Expression examples

[spoon] matches anything containing those letters (e.g. spoonless, sporran)

^spoon$ only matches spoon

^soon matches anything beginning soon

Hide|seek matches hide or seek

..te matches any four characters ending in te

M[\d]{1,2} matches motorways

Task: Regex 101

Go to https://regex101.com/ (select Python 'flavor')

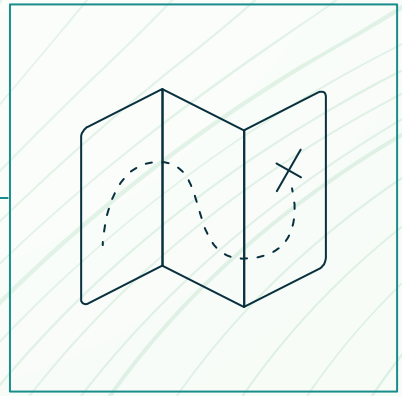
Try:

- (07[\d]{8,12})
- .+\@.+\..+
- ^(?:[0-9]{1,3}\.){3}[0-9]{1,3}
- (^([Gg][Ii][Rr]0[Aa]{2})|((([A-Za-z][0-9]{1,2})|(([A-Za-z][A-Ha-hJ-Yj-y][0-9]{1,2})|(([A-Za-z][0-9][A-Za-z])|([A-Za-z][A-Ha-hJ-Yj-y][0-9]?[A-Za-z])))) [0-9][A-Za-z]{2}))

Can you work out what personal information they might test? How robust do they seem?

Conclusions

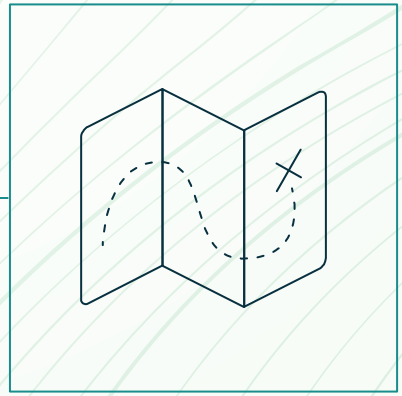
- If access is an afterthought, you're doing it wrong
- Access standards are great, demonstrably meeting user needs is better
- It's easy to generate and publish metadata about collections and provide downloadable files
- Doing this really well is not so easy
- The time to begin is today



Thanks!

Any questions?

You can find me at @mentionthewar &
jpugh@nationalarchives.gov.uk



Thanks!

Feedback forms are online at:

https://tinyurl.com/y3vo5h8t