Cabinet Office
Digital

Using AI to Review Records in the Cabinet Office

27 June 2024

David Canning MBE, Head of Digital Knowledge and Information Management
Dr Kelcey Swain, Development Operations Lead

Enabling and Transforming the Cabinet Office

Inclusive ◦ Quality ◦ Transparent

Cabinet Office Digital

The Digital Paradigm

Issues affecting access to digital records in archives include:

- Volume
- Format
- Distribution
- File Plans!

It’s a digital problem, so we need digital solution

Cabinet Office Digital

Live systems corrupt data!

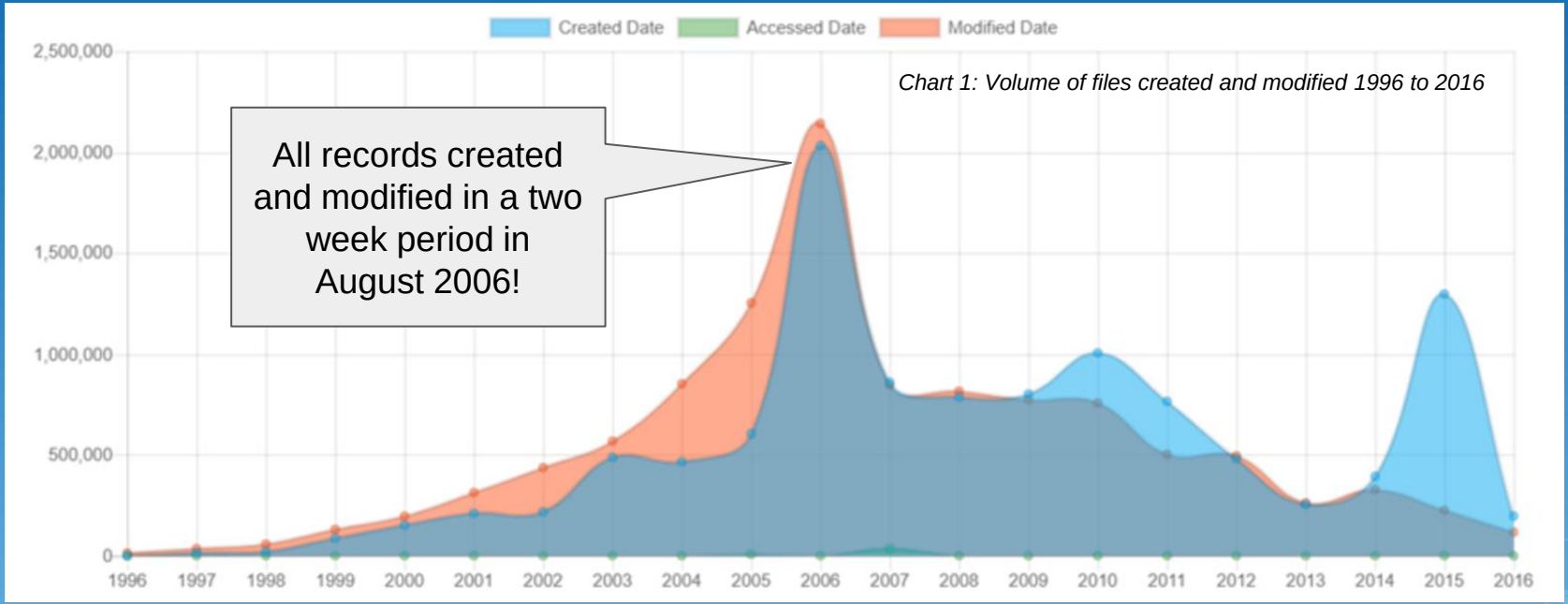
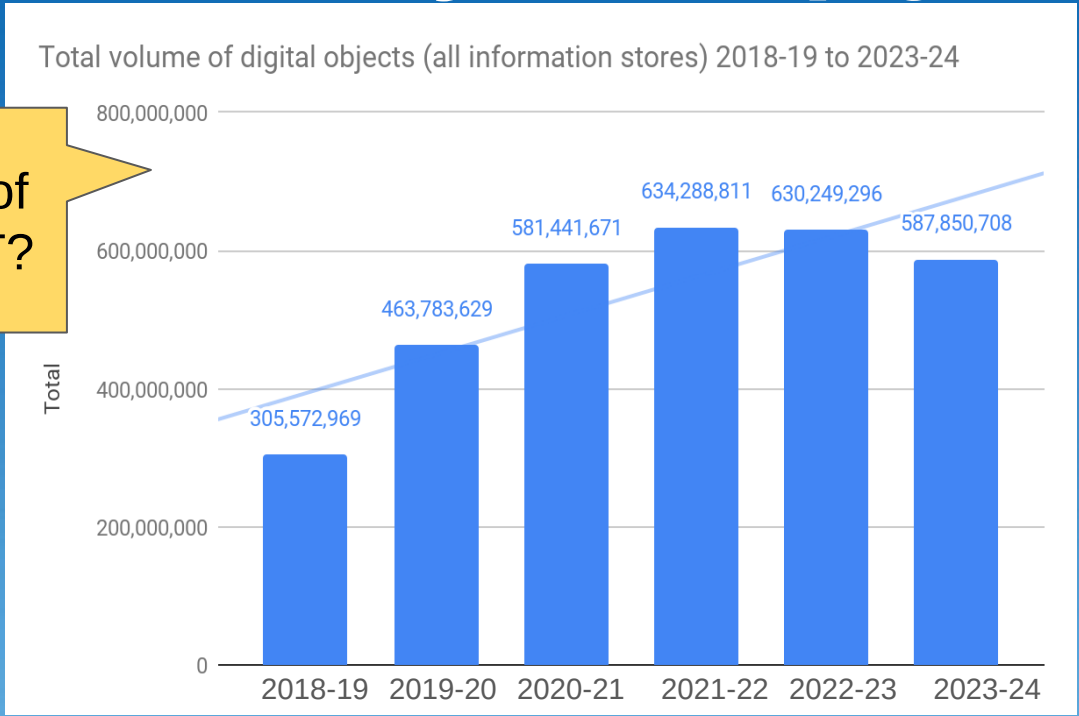


Chart 1: Volume of files created and modified 1996 to 2016

Cabinet Office Digital

Data will just keep growing

How much of this is R.O.T?

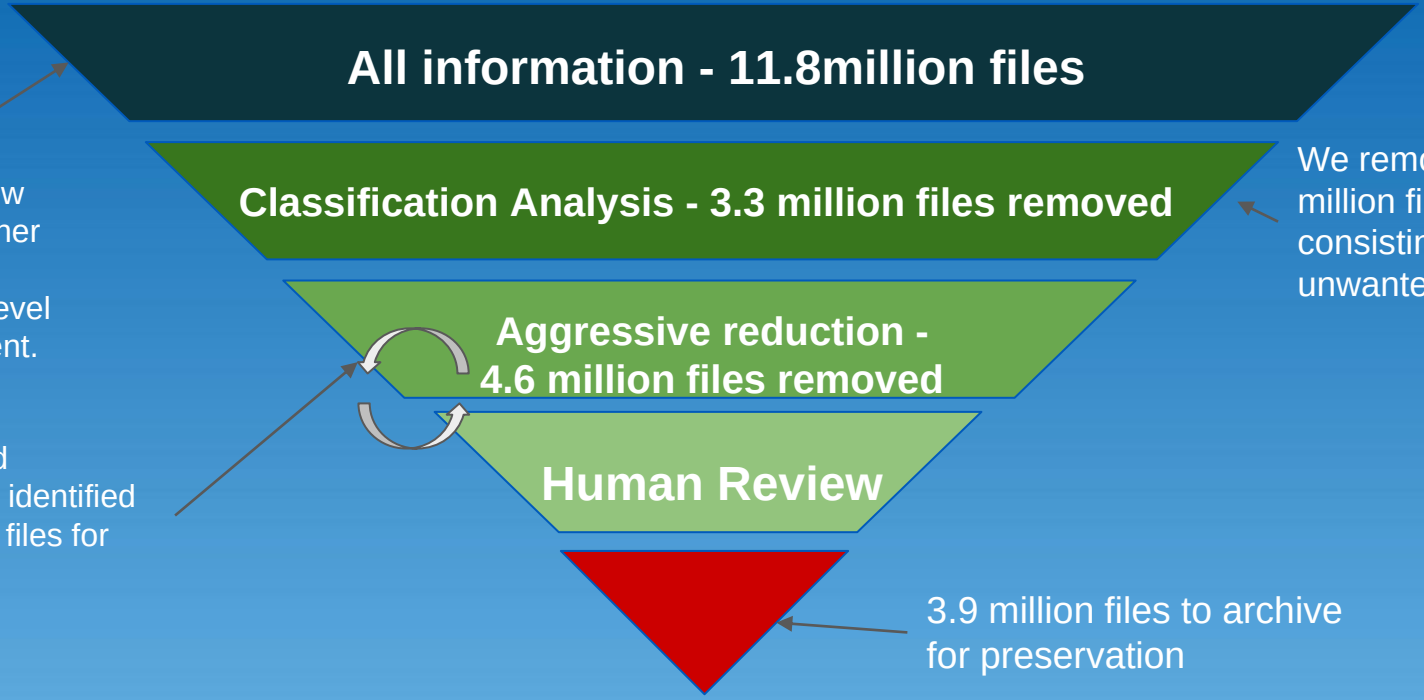


"The neuropsychological capacity of the human brain to process and record information may constitute the dominant limiting factor for the overall growth of globally stored information, with real-world economic constraints having only a negligible influence."

Gros, Kaczor, Markovic
Goethe University, Frankfurt, 2011

Cabinet Office Digital

Digital Disposal Methodology



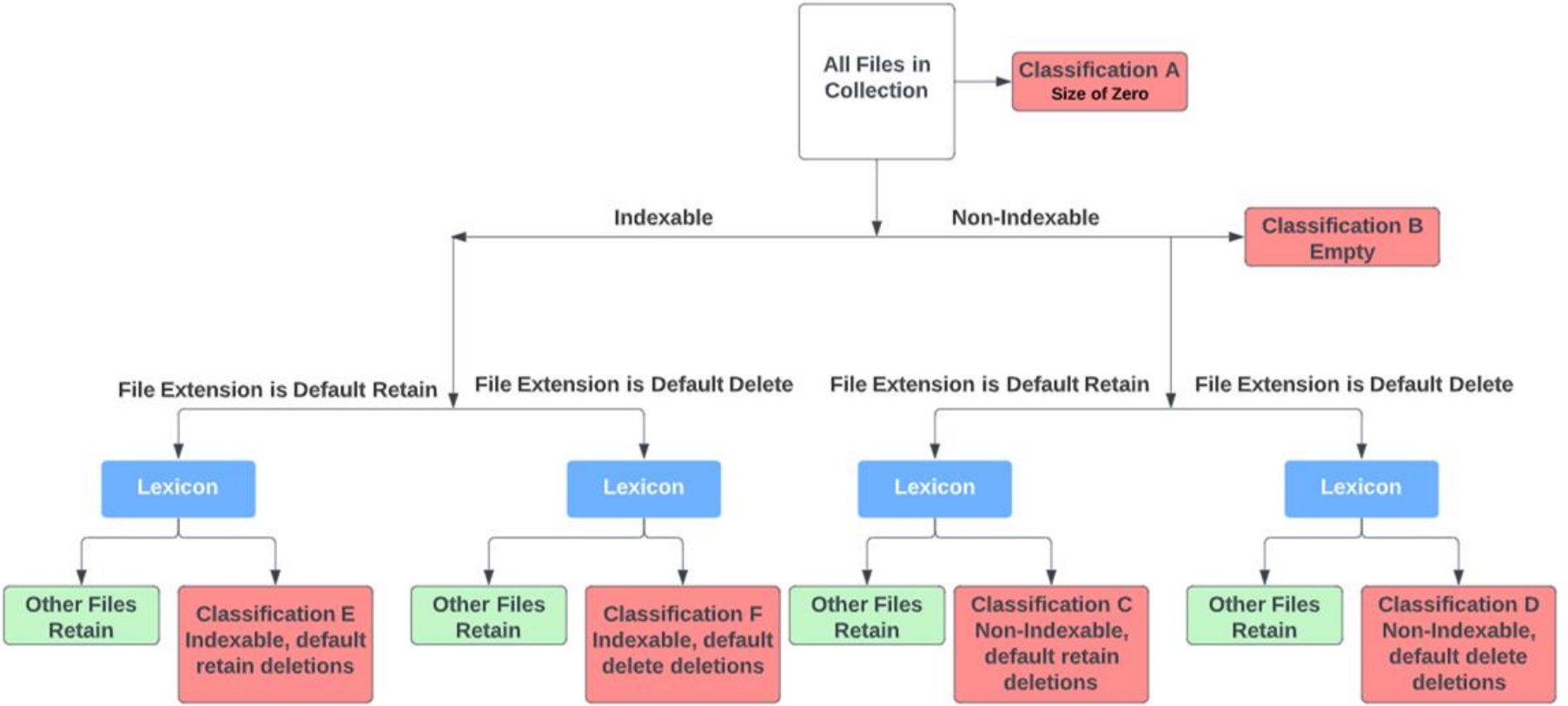
A manual review removed a further 2.8 million files through a top level review of content.

We removed 3.3 million files consisting of unwanted formats

Automated processes identified 1.8 million files for removal

3.9 million files to archive for preservation

Our Algorithmic Model for Document Review



Cabinet Office Digital

Lexicon terms were categorised according to their long-term value

Value
Submission
Minister
Bilateral
Richard Wilson

ROT
Annual leave
Blah blah
Christmas card
Stuff

ROT, even when in the presence of Value terms
Auto-response
Template
Timesheet
OOO

Weighting the algorithm - bias

Bias towards **retention** because there is an assumption that the document will contain valuable information based on its format

High - (Frequency of terms) - Low

Low - (Weighting of terms) - High

Default Value Extension Types	Value	ROT	ROT
Value	Retain	Retain	Delete
ROT	Retain	Delete	Delete
ROT	Delete	Delete	Delete

Files falling into this category may need to be human reviewed

Bias towards **deletion** because there is an assumption that the document will not contain valuable information based on its format

High - (Frequency of terms) - Low

Default ROT Extension Types	Value	ROT	ROT
Value	Retain	Delete	Delete
ROT	Delete	Delete	Delete
ROT	Delete	Delete	Delete

Enabling and Transforming the Cabinet Office

Cabinet Office Digital

Weighting the algorithm - Relevance

Term frequency (TF)

- The more times that a search term appears in the field we are searching in a document, the more relevant that document is.

Inverse document frequency (IDF)

- The more documents that contain a search term in the field that we are searching, the less important that term is.

Field length

- If a document contains a search term in a field that is very short (i.e. has few words), it is more likely relevant than a document that contains a search term in a field that is very long (i.e. has many words).

Cabinet Office Digital

Outcomes and Benefits

Cost avoidance/ efficiency	Consistency & accuracy	Speed
• Legacy: £2.2m • Backlog: £1.7m • Future: £190k pa	• Predictable • Auditable • Consistent and Reliable	• 59 years' work completed in one year

1xFTE pa required to review 200,000 documents

Cabinet Office Digital

Next Steps Development

Machine Learning Techniques
- Naive Bayes
- Markovian Discrimination

Mitigation
- Human-in-the-loop

Future-proof
- Retaining a trace of ROT