

Preserving databases as records: experience at NDAD

Kevin Ashley

Head of Digital Archives Dept

ULCC

Overview

- How NDAD works
- Selection
- Accession process
- Documentation
- Description
- User access
- Contrast with other database archives

What is NDAD?

- A service run by ULCC under contract to National Archives
- Preserves UK government records which exist as 'structured information'
- Established in 1997 - service in March 1998
- First service by a national archive to provide online public access to preserved material
- Selection undertaken by National Archives and government departments
- Everything else is down to us

NDAD contains...

- Data not original systems
- Contextual material
- Datasets which:
 - ⇒ show development & implementation of major UK policy & legislation
 - ⇒ relate to major issues
 - ⇒ contain information of potential interest to all future researchers

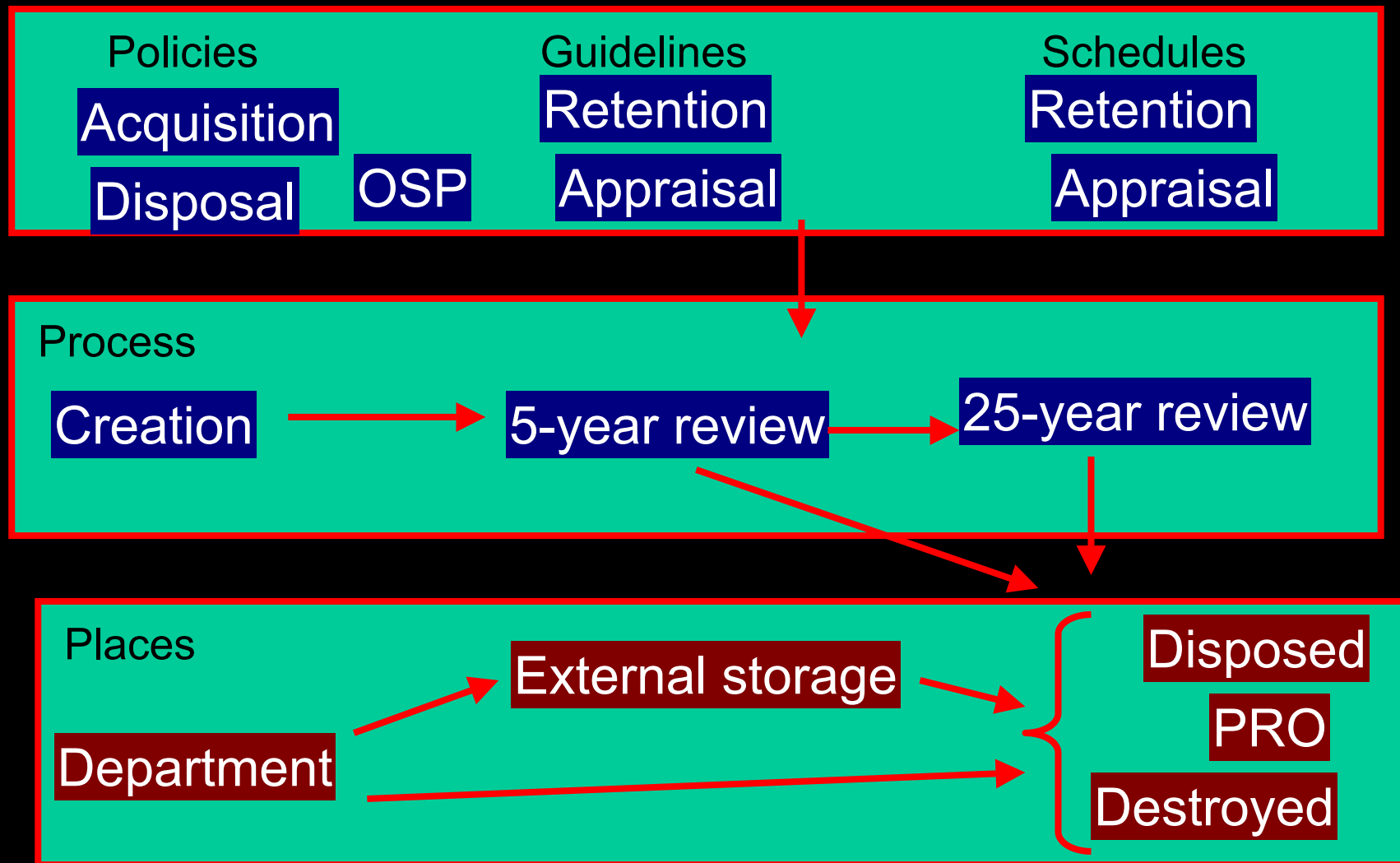
Our role

- Acquire
- Preserve
- Describe
- Provide access + copies
- Provide support to users
- Promote

Selection

- The only selection decisions I make are for my own department's records!
- Selection of organisational records is a multi-stage process
- Records are created by, used by, and document the activities of an organisation or individual

gov.uk records lifecycle



The way it has been

- Based on paper process
- Record content still primary driver, but...
 - ⇒ Some selected because of IT significance
 - ⇒ Action taken earlier because of obsolescence
 - ⇒ Ease of preservation probably has more significance than with paper
- Lack of inventory significant barrier
- Digital form allows new types of content to be preserved for effective re-use

Resultant selection decisions

- Digital records continuing from preserved paper records
- Large, easily identifiable systems
- Systems on brink of retirement or replacement
- Many digital records still invisible to manual selection process
- Many actors involved are new to process of records management

Accession process

- Information collected from records manager and data owners: format, titles, documentation, etc
- Digital and paper material acquired and listed
- Check for:
 - ⇒ completeness
 - ⇒ accuracy
 - ⇒ readability
- May lead to further acquisition or negotiation

Preservation

- Data transformed to canonical form - originals kept
- Paper documentation digitised
- Technical metadata produced or transformed
- Consistency checks applied:
 - ⇒ For transformation process
 - ⇒ Against original system
 - ⇒ Against published information
 - ⇒ Internal cross-checks

About 'errors'

- Checks may tell us our process is wrong
- More often it is people or the system or the documentation that is wrong
- We describe inconsistencies - we do not correct
- Incorrect data is still a true record
- But re-use must be informed by knowledge of data quality
 - ⇒ Many sources for this

Examples of inconsistencies

- Metropolitan Police Crime Records:
 - ⇒ Coding system for police stations incomplete and possibly inaccurate
 - ⇒ Some codes not explained
 - ⇒ Some codes reused
- Data type errors frequent: dates which are not dates; integers which are text
- Published statistics do not always agree with source data

Inconsistencies (2)

- Schools census - 4 datasets per year for different school types
- But 1976 only has 3 - no nursery schools
- Further examination shows files have been merged
- Confirmation came from completed census forms held by schools - not by government department

Documentation

- 3 sources:
 - ⇒ Comes with accession: paper or digital form
 - ⇒ Produced by NDAD
 - ⇒ Gathered from elsewhere

Sources

- The system itself
- Documentation prepared by supplier
- Documentation produced by user
- Internal records of the organisation
- Publications
- Oral history
- Specialist knowledge

Description

- Professional archivists work with data specialists
- Goal is to document the system, its use and its context
- Information is historical and technical
- For some users, the catalogues are all that is needed
- For others, catalogues can assist in data reuse

Metadata

- Micro meta-data:
 - ⇒ Data types, field descriptions, data ranges...
- Meta-data on a broader level about:
 - ⇒ The data itself
 - ⇒ The systems
 - ⇒ Uses
 - ⇒ Results of that use
 - ⇒ Who provided, used, was influenced by, the data

Metadata/catalogues

- About series
 - ⇒ dataset & table relationships
 - tables
 - fields
- Model is broadly relational
- Some attributes multi-level and can be inherited:
 - ⇒ Access restrictions

Field-level metadata

- Includes:
 - ⇒Name
 - ⇒Datatype
 - ⇒Description/original description
 - ⇒Access restrictions
 - ⇒Statistical info (max/min etc)
 - ⇒Coding information

Documenting functionality

- Understanding how a system worked is important
- Current computing techniques may make data access far easier now than in 20,30,40 years ago
- Just because information is *in* a dataset does not mean it was usable
- But preserving software and access methods does not help current users

Access

- Access is via website and linked to hierarchical catalogues
- Not all data open to access
- Simple viewing of rows from one or more tables:
 - ⇒ Ability to choose fields
 - ⇒ Ability to use query language or forms to select rows
 - ⇒ Refine query or fields selected at any point
- Links back to catalogues + documentation
- Ability to order copies of data and/or documentation or download data

Access - goal

- Intention is NOT to duplicate original system
- Nor do we provide advanced analysis tools
- Simple viewing is the goal via a generic tool
- Viewing caters for multimedia datatypes and is extensible via object-like design
- Traditional database systems not up to task without significant additional effort
- Hence much software home-grown

What's different with digital?

- Size of record collection no longer a preservation barrier
 - ⇒but complexity is more important than it was
- No need to sample case files, etc
- Quantitative research now trivial, hence...
- ... Qualitatively different research now possible
- Cost of acquisition completely different
- Acquisition process significantly more complex
 - ⇒but this is a short-term phenomenon

Archive vs data bank

- The role of an archive is to preserve
- We do not correct or 'improve' material
- The holdings of an archive purport to be a record of what was used
- Data validation ensures that errors or discrepancies are identified and described
- Knowing that incorrect or incomplete information was in use can be important

And finally...

- NDAD was a quick solution to an urgent problem
- Keeps intellectual control with PRO:
leaves details to service provider
- Provided a base for training archivists and information technologists
- Positive experience for all partners