Data Audit Framework Development (DAFD) Project

Sarah Jones, HATII

Delos Summer School, 8th – 13th June 2008

JISC

Talk will consider:

1. Background to the Data Audit Framework (outline problem, proposed solution, and projects JISC have funded in this area)
2. DAF Methodology (discussion of the workflow and five stages)
3. Initial audits (progress, lessons learned, what this may mean for the future)
4. Conclusions

The problem

Lack of knowledge

- what types of data are present within UK institutions?
- how they are managed?
- where they are deposited for long-term preservation?

JISC

•Little is known about research data is held by institutions, either publicly funded or private researcher's collections. General conclusions can be drawn but little concrete evidence.

•Research data is often forgotten about at the end of a project – there may not be a data management policy, or it might not be implemented

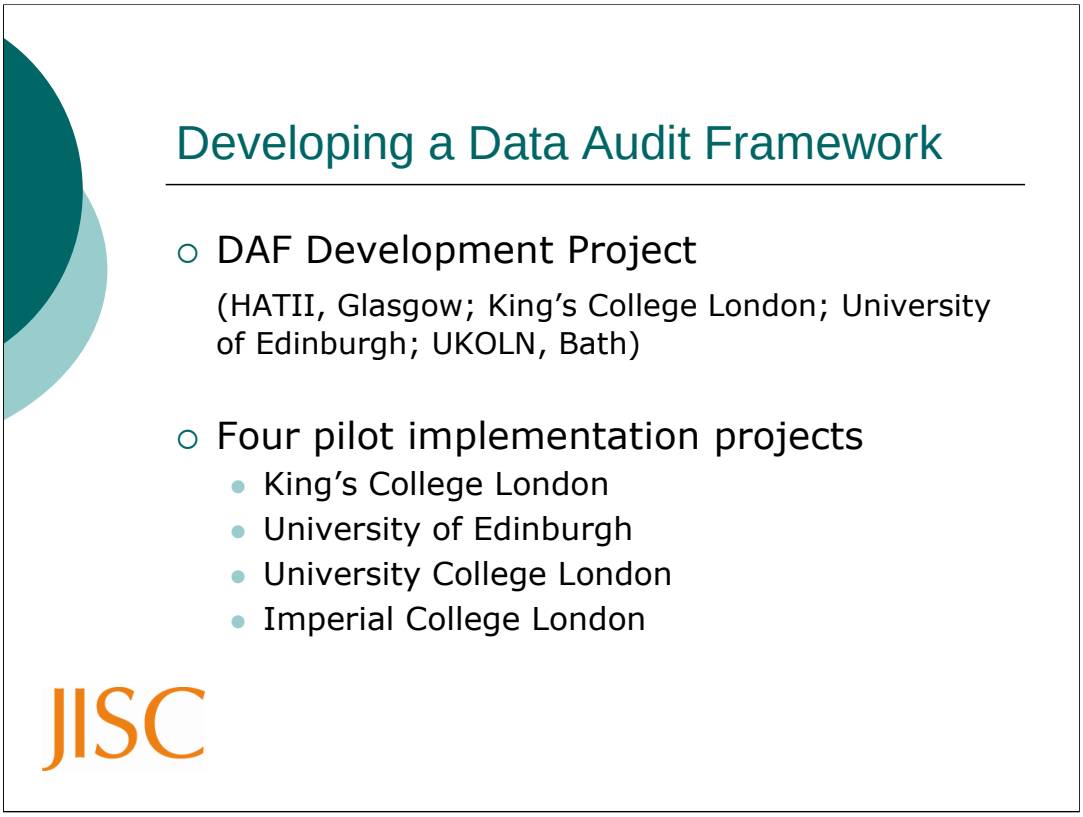
•There isn't always a mandate to deposit data with an archive so the products of research grants are not always properly maintained.

The solution

"JISC should develop a Data Audit Framework to enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation"

Liz Lyon, Dealing with Data: Roles, Rights, Responsibilities and Relationships, (2007)

JISC



JISC funded five projects: one overall development project to create an audit framework and online tool and four implementation projects to test the framework and encourage uptake.

DAFD schedule

April	Develop methodology for collecting data
May-June	Test preliminary methodology through pilot audits Glasgow: archaeology Edinburgh: geosciences King's College: medical UKOLN: engineering
June	Define system requirements & develop prototype
July-August	Implementation and iterative development
September	Release and dissemination

JISC

•DAFD is a short-term project (6 months) and will be running concurrently with some of the implementation projects.

•We're currently at the stage where the methodology has been developed and is being tested through preliminary audits. The audits are still ongoing but the initial feedback we have has been passed to the developers so they can write a requirements document and start development.

•Two of the implementation projects will be starting in June so they'll also be able to contribute to the ongoing development. The others will begin once the online tool is released.

DAF Methodology

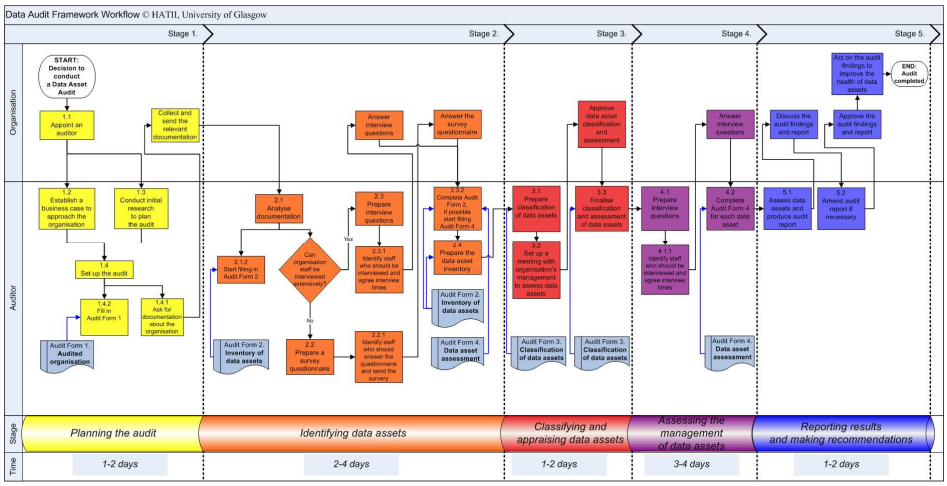
Five stages:

- Planning the audit;
- Identifying data assets;
- Classifying and appraising data assets;
- Assessing the management of data assets;
- Reporting findings and recommending change.

JISC

Will talk briefly through the key points of each stage

DAF workflow



JISC

Stage 1: Planning the audit

- Selecting an auditor
- Establishing a business case
- Research the organisation
- Set up the audit

JISC

•The auditor needs sufficient time to dedicate to the audit – suggestion in trial audits is that post grads may be ideal candidates as they understand subject area, are familiar with staff and research of department, have access to internal documents and can be paid to focus effort on this

•Selling the audit needs consideration – benefits should be tailored to specific circumstances of institution

•Key principle in this stage is getting as much as possible done in advance so time on-site can be optimised. Research should be conducted into departmental staff so the auditor knows who the best people to contact are and as many interviews as possible should be set up in advance.

Stage 2: Identifying data assets

- Collecting basic information to get an overview of departmental holdings

Audit Form 2: Inventory of data assets				
Name of the data asset	**Description of the asset**	**Owner**	**Reference**	**Comments**
Bach bibliography database	A database listing books, articles, thesis, papers and facsimile editions on the works of Johann Sebastian Bach	Charles Fairall	RAE return for 2007, http://www....ac.uk/...	An MS Access database in H:\Research\Bach\Bach_Bibliography.mdb.

JISC

The information to populate the inventory can come from desk research, surveys and/or interviews.

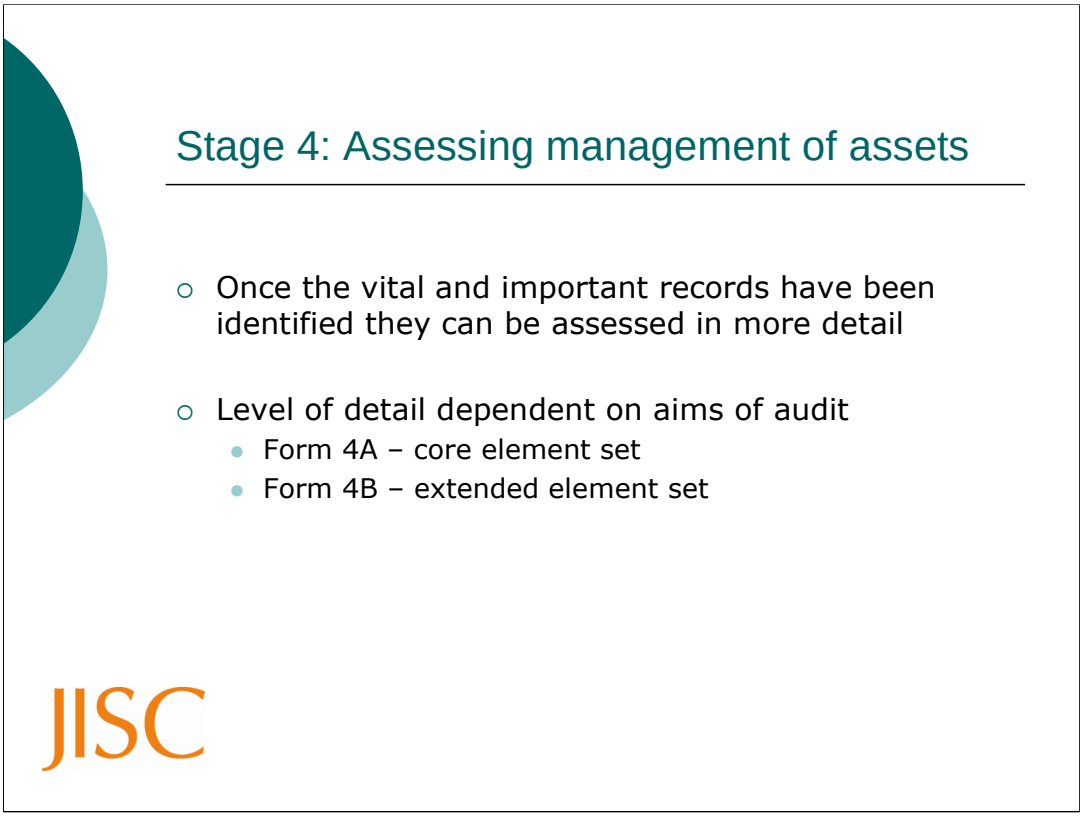
By the end of this stage there should be a complete inventory of data assets

Stage 3: Classifying and appraising assets

- Classifying records to determine which warrant further investigation

Vital	Vital data are crucial for the organisation to function such as those: ○ still being created or added to; ○ used on frequent basis; ○ that underpin scientific replication e.g. revalidation; ○ that play a pivotal role in ongoing research.
Important	Important data assets include the ones that: ○ the organisation is responsible for, but that are completed; ○ the organisation is using in its work, but less frequently; ○ may be used in the future to provide services to external clients.
Minor	Minor data assets include those that the organisation: ○ has no explicit need for or no longer wants responsibility for; ○ does not have archival responsibility e.g. purchased data.

JISC



This stage of the audit will provide the basis of the final recommendations. The current management and curation practices will be assessed to identify weaknesses and risks.

Form 4a collects a basic set of 15 data elements based on Dublin Core. The extended set collects 50 elements (28 mandatory, 22 optional). These are split into six categories:

-Description

-Provenance

-Ownership

-Location

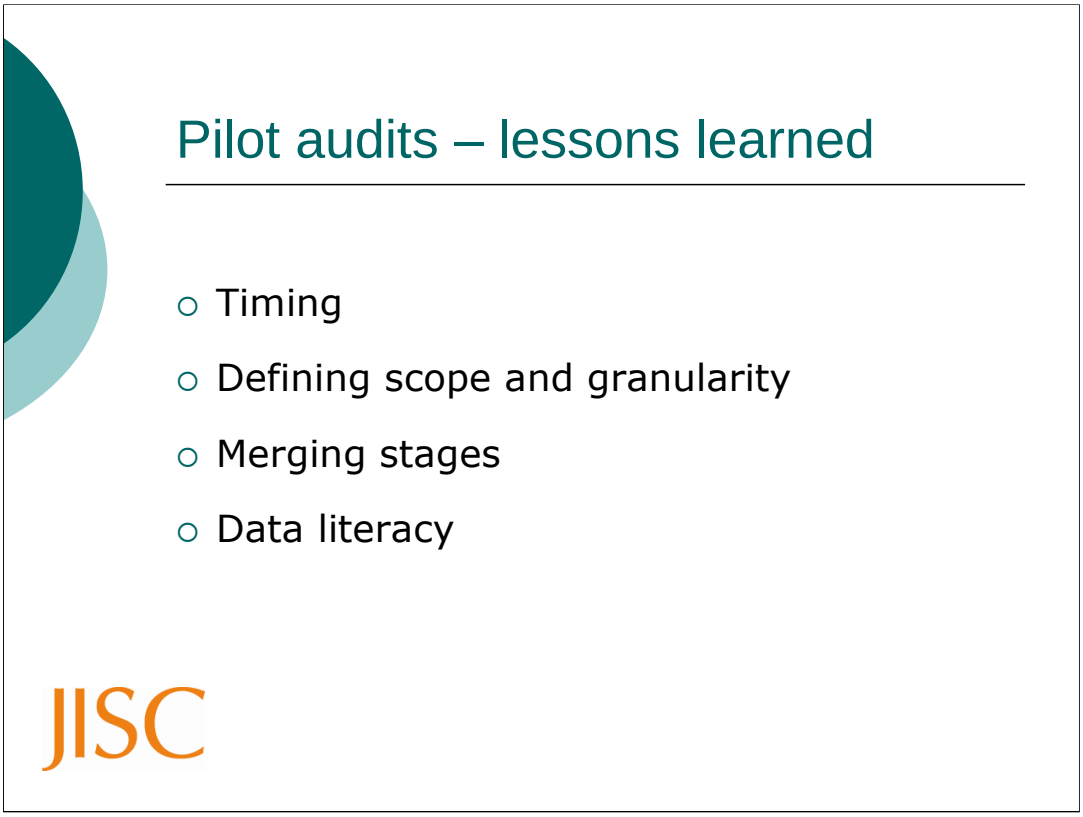
-Retention

-Management

Audit Form 4A: Data asset management (Core element set)

No	Parameter	Comment
1	**ID**	*A unique identification assigned by the audior or organisation to each data asset*
2	**Title**	*Official name of the data asset, with additional or alternative titles or acronyms if they exist*
3	**Description**	*A description of the information contained the data asset*
4	**Subject**	*Information and keywords describing the subject matter of the data asset*
5	**Purpose**	*Reason why the asset was created, intended user communities or source of funding / original project title*
6	**Coverage**	*Intellectual domain or subject area covered by the information in the data asset. Spatial and temporal coverage*
7	**Source**	*The source(s) of the information found in the data asset*
8	**Author**	*Person, group or organisation responsible for the intellectual content of the data asset*
9	**Date**	*The date on which the data asset was created or published*
10	**Updating frequency**	*The frequency of updates to this dataset to indicate currency*
11	**Language**	*The language(s) of the data asset content*
12	**Type**	*Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)*
13	**Format**	*Physical formats of data asset, including file format information*
14	**Rights**	*Basic indication of the user's rights to view, copy, redistribute or republish all or part of the information held in the data asset*
15	**Relation**	*Description of relations the data asset has with other data assets and any any DOI ISSN or ISBN references for publications based on this data*

The report feeds back on the results of each stage of the audit and identifies changes that would lessen risks and improve data management.



The audits are all underway, 2 are nearly finished and 2 are in the planning stages.

Timing

•Important to time the audit conveniently – be aware of holidays, exam times, periods out of the office

•Lead in time for audit required. We estimate man hours to be 2-3 weeks but elapsed time to be up to 2 months given the time needed to set up interviews

Defining scope & granularity

•Audits can take place across whole institutions and schools / faculties or within more discrete departments and units. Level of granularity will depend on the size of the organisation being audited and the kind of data it has. There may be numerous small collections or a handful of large complex ones. Scope and granularity will depend on circumstances.

Merging stages

•Methodology flows logically from one stage to the next, however initial audits have found it easier to identify and classify at once – this will be amended into one stage

Data literacy

•General experience has shown basic policies are not followed even in data literate institutions – no filing structures, naming conventions, registers of assets, standardised file formats or working practices. Approaches to digital data creation and curation seem very ad hoc and defined by individual researchers.

Conclusion

- Outcomes very preliminary but positive
  - Experience confirms data audit is needed
  - Time needed is longer than initially anticipated but still manageable
  - Results will support various other data projects

JISC

•Initial audits demonstrate vast amount of data is being created and confirm there is little documentation / knowledge of what exists. The un-standardised approach to creating and managing assets we encountered suggest the audit will provide institutions with the basics to address current data management issues.

•Time required is quite extensive, however if scoped well the audit will remain manageable. Inventory need not always be comprehensive but could be a representative sample from which the most important recommendations could be drawn.

•The DAF team are collaborating with several other JISC data projects:

- •UKRDS which is defining costs of preservation – identifying assets would be a useful baseline for them;
- •DataShare which is developing new models, workflows and tools for academic data sharing;
- •Skills, role and career structure of data scientists & curators – by Key Perspectives, Alma Swan;
- •DCC Summer School which will provide a venue to promote the online audit tool and offer a practical workshop to train potential auditors.