DataShare & Data Audit Framework projects at Edinburgh

Robin Rice
Data Librarian
EDINA and Data Library
Information Services
University of Edinburgh

JISC

Edinburgh eScience Collaborative Workshop
12 June 2008

Overview

- Background, Data Library
- About DISC-UK DataShare project
- National related initiatives
- Context – Edinburgh's participation in the Data Audit Framework
- About DAF project

What is a data library?

A **data library** refers to both the content and the services that foster use of collections of numeric and/or geospatial data sets for secondary use in research. A data library is normally part of a larger institution (academic, corporate, scientific, medical, governmental, etc.) established to serve the data users of that organisation.

The data library tends to house local data collections and provides access to them through various means (CD-/DVD-ROMs or central server for download). A data library may also maintain subscriptions to licensed data resources for its users to access.

Data Library Resources

- Large-scale social science survey data
- Country and regional level time series data
- Population and Agricultural Census data
- Data for mapping
- Resources for teaching

Data Library Services

- Finding…

"I need to analyse some data for a project, but all I can find are published papers with tables and graphs, not the original data source."

- Accessing …

"I've found the data I need, but I'm not sure how to gain access to it."

- Using …

"I've got the data I need, but I'm not sure how to analyse it in my chosen software."

- Managing …

"I have collected my own data and I'd like to document and preserve it and make it available to others."

DISC-UK DataShare project

Data Information Specialists Committee - UK

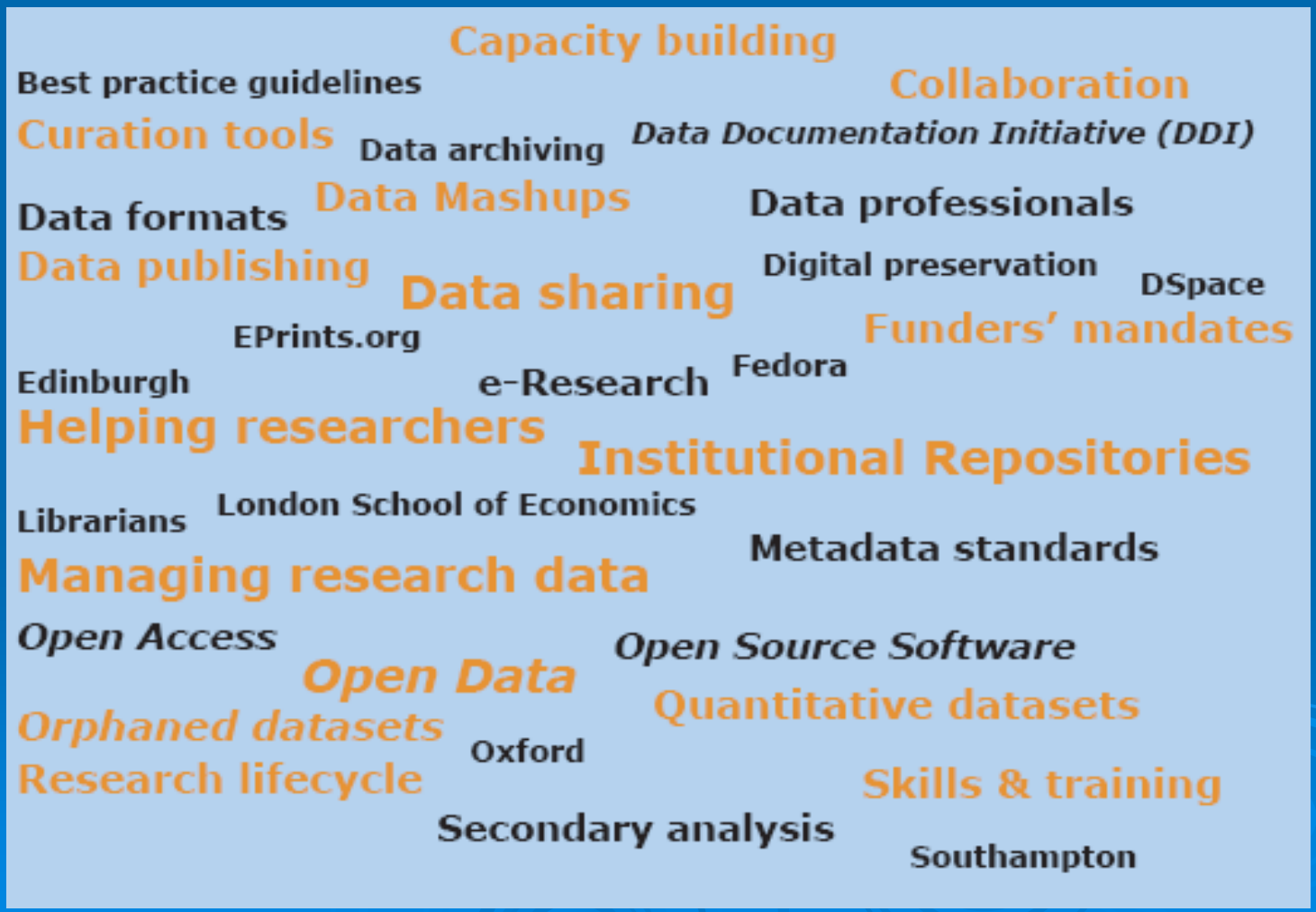
DISC-UK DataShare Project: March 2007 - March 2009

Funded by JISC Digital Repositories and Preservation Programme

The **project's overall aim** is to contribute to new models, workflows and tools for academic data sharing within a complex and dynamic information environment which includes increased emphasis on stewardship of institutional knowledge assets of all types; new technologies to enhance e-Research; new research council policies and mandates; and the growth of the Open Access / Open Data movement.

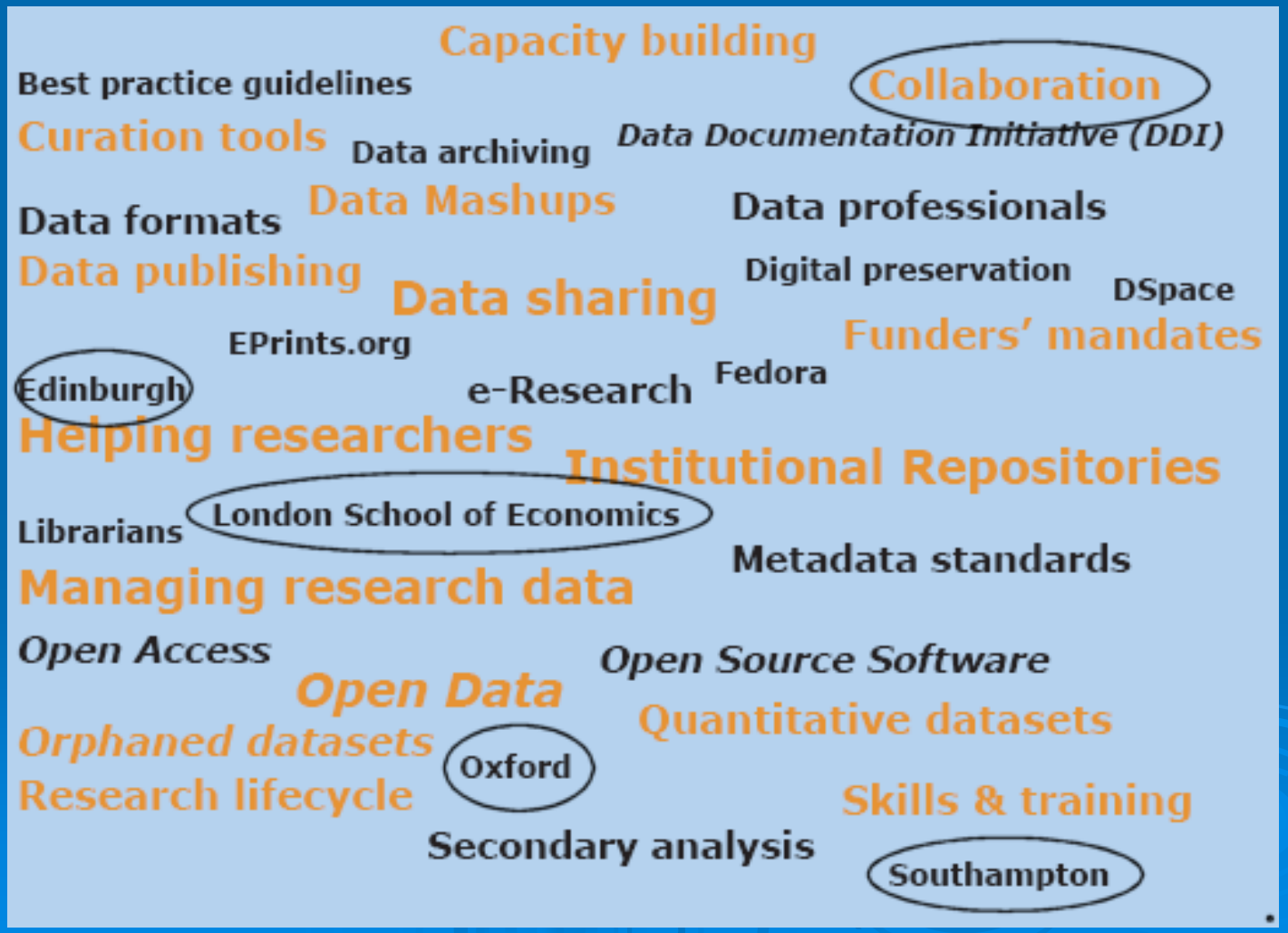
DISC-UK DataShare project

Data Information Specialists Committee - UK



DISC-UK DataShare project

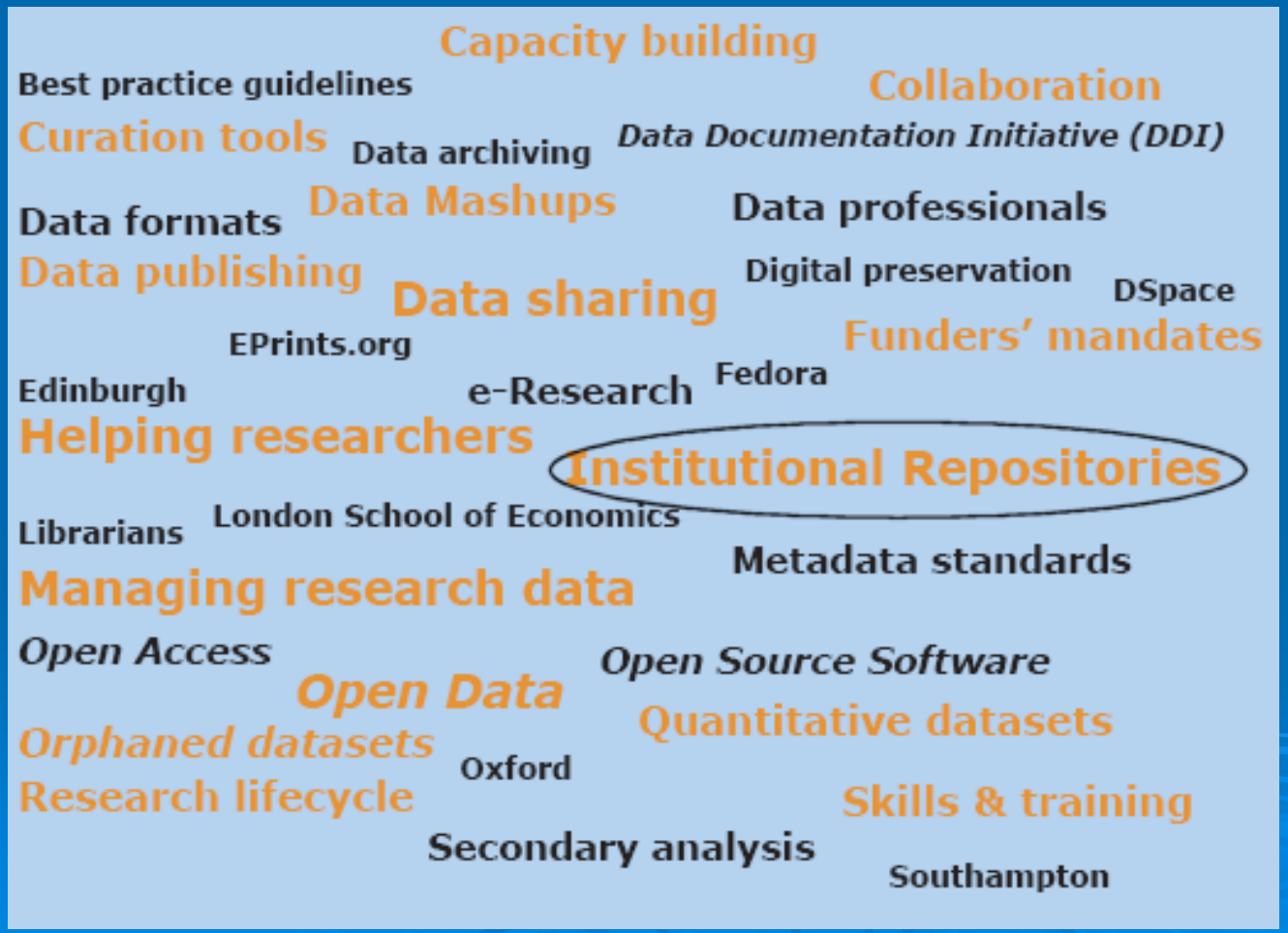
Data Information Specialists Committee - UK



Project Partners

DISC-UK DataShare project

Data Information Specialists Committee - UK



Digital Repositories

According to Heery and Anderson (*Digital Repositories Review*, 2005) a **repository** is differentiated from other digital collections by the following characteristics:

- content is deposited in a repository, whether by the content creator, owner or third party
- the repository architecture manages content as well as metadata
- the repository offers a minimum set of basic services e.g. put, get, search, access control
- the repository must be sustainable and trusted, well-supported and well-managed.

OA & Institutional Repositories

- *Open access* repositories allow their content to be accessed openly (e.g. downloaded by any user on the WWW) as well as their metadata to be harvested openly (by other servers, e.g. Google or scholarly search engines).

- WIKIPEDIA *definition*: Open access (OA) is free, immediate, permanent, full-text, online access, for any user, web-wide, to digital scientific and scholarly material, primarily research articles published in peer-reviewed journals.

- *Open Knowledge* Foundation: “A piece of knowledge is open if you are free to use, reuse, and redistribute it.”

- *Institutional repositories* are those that are run by institutions, such as Universities, for various purposes including showcasing their intellectual assets, widening access to their published outputs, and managing their information assets over time. These differ from subject-specific or domain-specific repositories, such as Arxiv (for Physics papers) and Jorum (for learning objects).

DISC-UK DataShare project

Data Information Specialists Committee - UK

Incentives for researchers to manage and share data: meeting funders' requirements

- Followed by the OECD 2004 ministerial declaration was the 2007 "*OECD Principles and guidelines for access to research data from public funding*."
- In 2005 Research Councils UK published a draft position statement on 'access to research outputs', followed by a public consultation, covering both scholarly literature and research data.
- ESRC added a mandate for deposit of research outputs (publications) into a central repository along with existing data deposit mandate into its central data archive (UKDA).
- In 2007 the Medical Research Council added a new requirement for all grant recipients to include a data sharing plan in their proposals.

Barriers to data sharing in IRs

- Reluctance to forfeit valuable research time to prepare datasets for deposit, e.g. anonymisation, code book creation.
- Reluctance to forfeit valuable research time to deposit data.
- Datasets perceived to be too large for deposit in IR.
- Concerns over making data available to others before it has been fully exploited.
- Concerns that data might be misused or misinterpreted, e.g. by non-academic users such as journalists.
- Concerns over loss of ownership.
- Concerns over loss of commercial or competitive advantage.
- Concerns that data are not complete enough to share.
- Concerns that repositories will not continue to exist over time.
- Concerns that an established research niche will be compromised.
- Unwillingness to change working practices.
- Uncertainty about ownership of IPR.
- IPR held by the funder or other third party.
- Sharing prohibited by data provider as a condition of use.
- The desire of individuals to hold on to their own ‘intellectual property’.
- Concerns over confidentiality and data protection.
- Difficulties in gaining consent to release data.

Benefits to Data Deposit in IRs

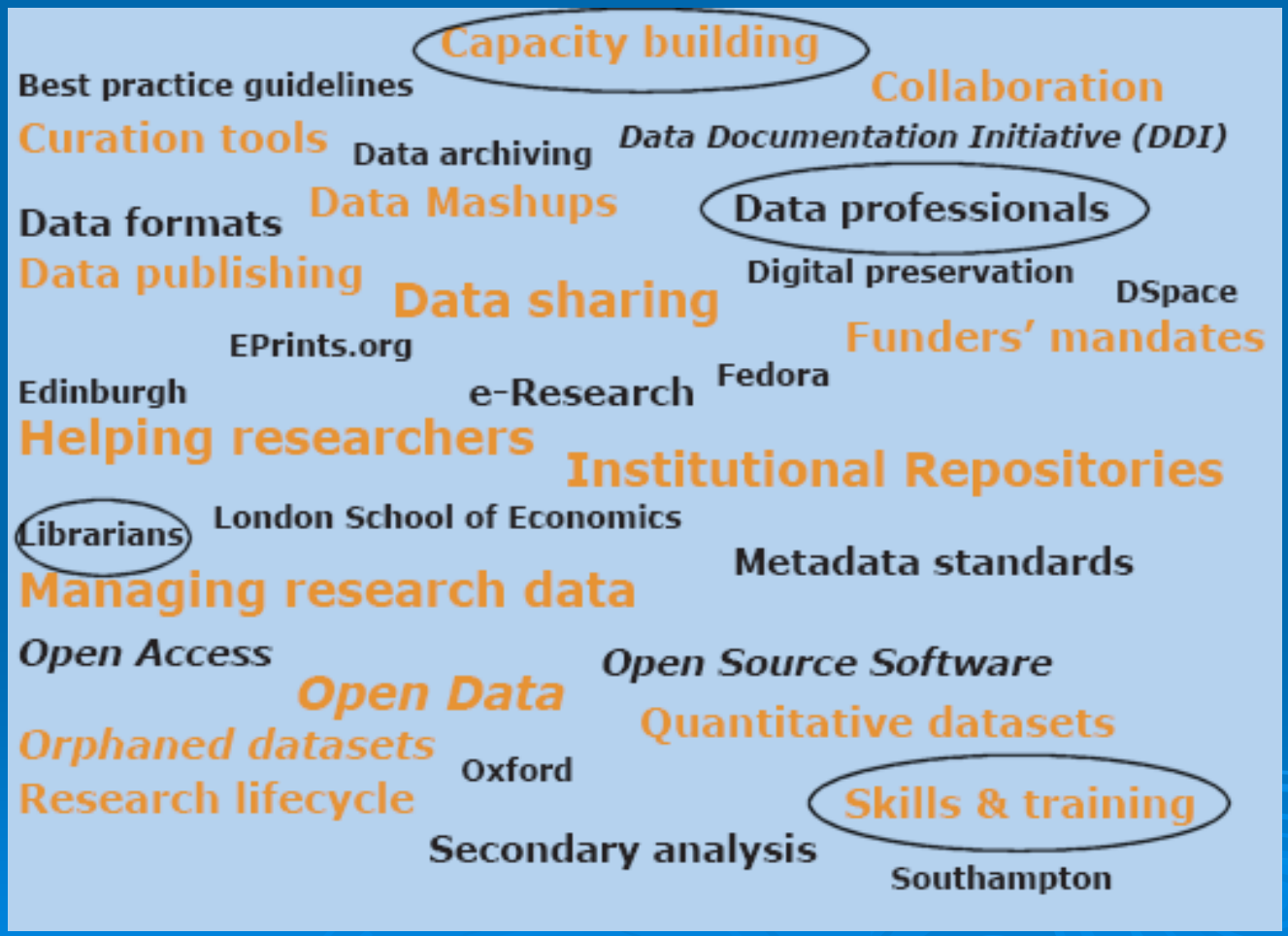
- IRs provide a suitable deposit environment where funders mandate that data must be made publicly available.
- Deposit in an IR provides researchers with reliable access to their own data.
- Deposit of data in an IR, in addition to publications, provides a fuller record of an individual's research.
- Metadata for discovery and harvesting increases the exposure of an individual's research within the research community.
- Where an embargo facility is available, research can be deposited and stored until the researcher is ready for the data to be shared.
- Where links are made between source data and output publications, the research process will be further eased.
- Where the Institution aims to preserve access in the longer term, preservation issues become the responsibility of the institution rather than the individual.
- A respected system for time-stamping submissions such as the 'Priority Assertion' service developed by R4L (Southampton, 2007) would provide researchers with proof of the timing of their work, should this be disputed.

Barriers and Benefits taken from

DISC-UK DataShare: State-of-the-Art Review (H Gibbs, 2007)

DISC-UK DataShare project

Data Information Specialists Committee - UK



Capacity building, skills & training, and professional issues

- Much current attention in UK on the capacity of HEIs to provide services for data management

- Debate about roles and responsibilities: researchers? funders? 'data scientists'? librarians? IT services?

- *Dealing with data: roles, responsibilities and relationships.* (L Lyon 2007)
- *Stewardship of digital research data: a framework of principles and guidelines.* RIN (2008)
- "Data Librarianship - a gap in the market." *CILIP Update* (Elspeth Hyams interview with L Martinez Uribe and S Macdonald, June 2008).
- *Data Seal of Approval*, DANS, the Netherlands (2008)
- Digital Curation Centre *Autumn school on data curation* (1 week)
- *Data Skills/Career Study:* JISC-commissioned study to report on 'the skills, role and career structure of data scientists and curators: an assessment of current practice and future needs.'
- *UK Research Data Service*: a RUGIT/RLUK feasibility study on developing and maintaining a national shared digital research data service for the UK HE sector

Enter Data Audit Framework

- Edinburgh University Data Library wishing to facilitate data curation, management, sharing
- Edinburgh DataShare repository looking for data depositors
- Information Services examining its support for research data (Research Computing Survey of staff across all Colleges)
- Edinburgh Compute and Data Facility and SAN
- CC&ITC looking at data storage requirements for College of Science and Engineering

Recommendation to JISC: Data Audit Framework

"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)

DAFD & DAFIs

(Thanks to Sarah Jones, HATII,
DAFD Project Manager for slides content)

- 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. All started in April and to finish by October.

- DAF Development Project, led by Seamus Ross

  (HATII, Glasgow, for DCC; King's College London; University of Edinburgh; UKOLN at Bath)

- Four pilot implementation projects
  - King's College London
  - University of Edinburgh
  - University College London
  - Imperial College London

Methodology

Based on Records Management Audit 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.

Stage 1: Planning the audit

- Selecting an auditor

  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

- Establishing a business case

  Selling the audit needs consideration of context

- Research the organisation

  Much information may be on their website

- Set up the audit

  Key principle in this stage is getting as much as possible done in advance so time on-site can be optimised. 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	Senior lecturer	RAE return for 2007, http://www....ac.uk/...	An MS Access database in H:\Research\Bach\Bach_Bibliography.mdb.

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.

Stage 4: Assessing management of assets

- Once the vital and important records have been identified they can be assessed in more detail

- Level of detail dependent on aims of audit
  - Form 4A – core element set
  - Form 4B – extended element set

- 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**

Stage 5: Report and recommendations

- Summarise departmental holdings
- Profile assets by category
- Report risks
- Recommend change

Pilot audits – lessons learned

- **Timing:** 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 and 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

- DISC-UK trying to track these and other tools and guidelines through its social bookmarks and tag cloud, blog, and a selected bibliography.
- Collective Intelligence page –
- http://www.disc-uk.org/collective.html
- For further information about DataShare –
- http://www.disc-uk.org/datashare.html
- Feel free to contact me, R.Rice@ed.ac.uk