a data management toolkit for research led institutions

Sarah Jones, Seamus Ross & Raivo Ruusalepp
HATII, University of Glasgow

CNI Task Force Meeting, Washington DC, USA
8th December 2008
Outline

• Background context
• Methodology and tool
• Pilot audit findings
• Conclusions
• Discussion
The problem

How can organisations realise the value of their research data when it is unclear:

• what data is held;
• where it is located;
• and how it is being managed?
A recommendation

“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”

Scope of work

• DAF Development Project
  (HATII* University of Glasgow; King’s College London; University of Edinburgh; UKOLN, University of Bath)

• Four pilot implementation projects
  – University of Edinburgh
  – Imperial College London
  – King’s College London
  – University College London
Benefits of audit

Awareness of data holdings
- Capacity planning
- Facilitate data sharing and reuse
- Monitor data holdings and avoid data leaks

Recognition of data management practices
- More efficient of resources and improved workflows
- Ability to manage risks – data loss, inaccessibility, compliance
- Enables the development of a data strategy
The methodology

Stage 1: planning

Objective
• Prepare in advance to optimise workflow and time with staff

Process
• Define scope / expected outcomes
• Research organisational context
• Set up interviews
Online tool: stage 1
Stage 2: identifying

Objective
- Create an inventory

Process
- Identify data
- Classify assets
- Hold interviews, questionnaires, desk-research
Stage 3: assessing

Objective

• Identify weaknesses in data management and potential risks

Process

• Assess most crucial assets to organisation
• Consider access, reuse, and data management
<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>A unique identification assigned by the auditor or organisation to each data asset</td>
</tr>
<tr>
<td>2</td>
<td>Data creator(s)</td>
<td>Person, group or organisation responsible for the intellectual content of the data asset</td>
</tr>
<tr>
<td>3</td>
<td>Title</td>
<td>Official name of the data asset, with additional or alternative titles or acronyms if they exist</td>
</tr>
<tr>
<td>4</td>
<td>Description</td>
<td>A description of the information contained the data asset and its spatial, temporal or subject coverage</td>
</tr>
<tr>
<td>5</td>
<td>Subject</td>
<td>Information and keywords describing the subject matter of the data</td>
</tr>
<tr>
<td>6</td>
<td>Creation date</td>
<td>The date(s) on which the data was collected or created</td>
</tr>
<tr>
<td>7</td>
<td>Purpose</td>
<td>Reason why the asset was created, intended user communities or source of funding / original project title</td>
</tr>
<tr>
<td>8</td>
<td>Source</td>
<td>The source(s) of the information found in the data asset</td>
</tr>
<tr>
<td>9</td>
<td>Updating frequency</td>
<td>The frequency of updates to this dataset to indicate currency</td>
</tr>
<tr>
<td>10</td>
<td>Type</td>
<td>Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)</td>
</tr>
<tr>
<td>11</td>
<td>Format</td>
<td>Physical formats of data asset, including file format information</td>
</tr>
<tr>
<td>12</td>
<td>Rights and restrictions</td>
<td>Basic indication of user’s rights to view, copy, redistribute or republish all or part of the information held in the data asset. Access restrictions on the data itself or any metadata recording its existence should also be noted</td>
</tr>
<tr>
<td>13</td>
<td>Usage frequency</td>
<td>Estimated frequency of use and if known required speed of retrieval to determine IT infrastructure and storage needs</td>
</tr>
<tr>
<td>14</td>
<td>Relation</td>
<td>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</td>
</tr>
<tr>
<td>15</td>
<td>Archiving policy</td>
<td>Number of copies of the data asset that are currently stored, frequency of back-up and archiving procedures</td>
</tr>
<tr>
<td>16</td>
<td>Management to date</td>
<td>History of maintenance and integrity of the data asset e.g. edit rights / security, and any curation activities performed</td>
</tr>
</tbody>
</table>
Audit form 3B: extended

- **Description**  title, type, subject, abstract, keywords, language...
- **Provenance**  purpose, source, context, dates, modifications....
- **Ownership**  creators, curators, rights, usage constraints...
- **Location**  references, versions, relations, potential archive...
- **Retention**  value, disaster recovery, archive / preservation policy...
- **Management**  documentation, audit trail, costs, funding...
### Online tool: Stage 3

#### Data Audit Framework

**Assets**

<table>
<thead>
<tr>
<th>Asset Name</th>
<th>Classification</th>
<th>Date Created/Updated</th>
<th>AF3 Form Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>View M74 assessment</td>
<td>vital</td>
<td>2008-10-19</td>
<td>core</td>
</tr>
<tr>
<td>View Rannoch Moor - HLF</td>
<td>important</td>
<td>2008-10-19</td>
<td>core</td>
</tr>
<tr>
<td>View Water Row Govan evaluation</td>
<td>minor</td>
<td>2008-10-19</td>
<td>core</td>
</tr>
<tr>
<td>View Kirk of St Nicholas Skeletal analysis</td>
<td>vital</td>
<td>2008-10-19</td>
<td>core</td>
</tr>
<tr>
<td>View Dun Eistean Archeology Project (DEAP)</td>
<td>important</td>
<td>2008-10-19</td>
<td>core</td>
</tr>
</tbody>
</table>

**Element set**

<table>
<thead>
<tr>
<th>Element set</th>
<th>Include</th>
<th>Exclude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Audit Tool Audit Stage 3**

- **Data Creator 1**: John Leslie (project management)
- **Data Creator 2**: Albert Hunter (archaeology)
- **Data Creator 3**: John Telfer (archaeology)
- **Data Title**: Dun Eistean Archeology Project (DEAP)
- **Description**: Photographs with panic, low and several consensus / trench tables proportional to an archaeological evaluation ed x 10 ha site proposed for... (continued)
- **Subject**: Archaeological report
- **Creation Date 1**: 2008-09-14
- **Creation Date 2**: 2008-05-01
- **Purpose**: Demonstration by the methodology group is to conduct a methodology within the planning permission granted by South Lannacester Council.
- **Source**: Fieldwork / Historic, drawn with e.g., desk-based assessment
- **Updating Frequency**: Daily during excavation
- **Data Type 1**: photographs
- **Data Format 1**: PDF
- **Data Format 2**: IIQ
- **Rights & Restrictions**: Accessed internally during construction and write-up. Owner, management and GICSC time likely to be made publicly available
- **Usage Frequency**: Daily during construction and write-up
- **Relation**: Desk-based assessment conducted by John Leslie
Stage 4: recommendations

Objective

- Recommend changes to improve data management

Process

- Collate audit results
- Analyse data
Online tool: stage 4

Audit Summary

- Number of assets: 7
  - University-funded: 2
  - Multiple funders: 3
  - Other: 2
  - Arts Council: 1
  - ESRC: 1
  - JISC: 1

- Most common formats: RAW, (2), PDF, (1), XML, (1)

Department of Theatre Film and TV Studies (TFTS): Audit Form 1

Organisation details

- Address: Theatre Film and Television Studies Department, University of Glasgow, The Gilmorehill Centre, University Avenue, Glasgow G12 8QQ

Description: Theatre Film and Television Studies (TFTS) at the University of Glasgow is one of the most well-established departments for these subjects in the UK. We have a strong team of over twenty teaching and research staff who specialise in diverse areas of performance and media arts. Several taught courses are offered. Through these students build up critical knowledge, creative approaches and confidence in their own skills. An excellent subject library situated on-site provides access to web resources, course literature and theatre production materials. The Department places a strong emphasis on research and benefits greatly from financial support offered by previously wholly, most recently, Dame Margaret McArthur. Several widely renowned research projects stem from McArthur Fellowships as well as from other research council funding secured by TFTS. These include Changing Approaches to Theatrical Talent, Glasgow Cinemas and Shakespearean performance database.

Strategies: Methods of working vary between staff members. No policy on creating and preserving research data is in place at present. Best practice guides from the Performance Data Archive (PDA) are used on research council projects and some staff have drawn these processes into their everyday work.

Responsibilities: Defined largely by individual researchers. Librarian maintains some research outputs and others disposed with data archives.

Budget: Research proposals are allocated funds to preservation. Internally, it’s part of library...
Data Audit Framework

Example recommendations

Devise and implement a data policy
  – Use a consistent approach to file naming and directory structures
  – Store data appropriately – not on external HDDs, USB sticks, CDs...
  – Maintain data integrity (checksums, access restrictions, edit rights)

Develop practical guidance and training (from PhDs / ECR)

Budget for curation and reuse from outset

Work with local services such as IR or archive
Pilot audits

- School of GeoSciences at University of Edinburgh
- Innovative Design and Manufacturing Research Centre (IdMRC) at University of Bath
- Glasgow University Archaeological Research Division (GUARD) at University of Glasgow
Main findings

- Lots of data being created
- Few policies for data creation, storage and management
- Researchers unsure where to begin
- Unaware of available support
- Often no place of deposit or funds for preservation

International Journal of Digital Curation
http://www.ijdc.net/ijdc/article/view/91/109
Implementation tips

- Scope audit carefully
- Timing needs to be appropriate
- Find an advocate in the department
- Collect information at once where possible
- General discussion to build rapport, communicate purpose of audit and understand organisation

Conclusions

Premise of DAF project is founded
- Lack of knowledge of what data is held
- Poor understanding of how to curate it

Key areas for future work
- Support on policies, standards and best practice guidance
- Skills lacking – basic training & professional workforce needed
- Develop robust, sustainable UK infrastructure for curation
- Recognition of data and active push for sharing and reuse critical
Next steps for DAF

• Support and training

• Implementation projects → redevelopment

• Results feed into other data research
  – UKRDS, Data Skills study
  – DCC policy study
  – Glasgow and Edinburgh University curation strategy
Further information


http://www.data-audit.eu

Sarah Jones: s.jones@hatii.arts.gla.ac.uk
Why curate?

What role should researchers assume in the curation lifecycle?

How can we overcome ownership and control issues to make data open?

What are the most pressing priorities?

What is required of co-ordinating bodies such as the DCC?

data sharing and reuse

a skilled workforce

appropriate repositories