Costs of Digital Archiving

the case of DANS

Anna Palaiologk | palaiologk.anna@gmail.com
Heiko Tjalsma | heiko.tjalsma@dans.knaw.nl
Laurents Sesink | laurents.sesink@dans.knaw.nl

IAASSIST 2010

Social data and social networking: Connecting social science communities across the globe.

June 2, 2010 | Cornell University | Ithaca | US
Index

• Challenge
• Objectives
• What’s new?
• Approach
• Components
• How does it work?
• Results Analysis
• Next Steps
• Qs & As
Challange

- increase in research data deposits
  - E-science
  - policies (Open Access mandates etc)
- new (complex) data formats
  - video, GIS, 3D,
- new ways to disseminate and (re)use of research data
  - Visualization/analyzing tools, enhanced publications, Research Infrastructures
Challange

- support the:
  - increase in research data
  - demand for new functionality
  - Research Infrastructures

- with a (nearly) fixed budget
Objectives

- more accurate planning
- better forecasting and control
- more accountability and transparency
- control the level of ambition - realistic strategy

Specific questions:
- What are the costs of archived datasets?
- What are the most labour intensive activities?
Building on existing research

- NASA Cost Estimation Toolkit (CET)
  - (Booth et al. 2006; Fontaine et al. 2007; Hunolt, 2006a; Hunolt, 2006b; Hunolt et al. 2006)

- LIFE Costing Model
  - (McLeod, Wheatley & Ayris, 2006; Ayris et al., 2008)

- Keeping Research Data Safe (KRDS)
  - (Beagrie, Chruszcz & Lavoie, 2008)

- Common features:
  - OAIS reference model
  - Activity Based Costing (ABC) approach
Approach: ABC

- staff is the major resource pool

<table>
<thead>
<tr>
<th>Data Acquisition</th>
<th>Office</th>
<th>IT services and equipment</th>
<th>Staff</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,19%</td>
<td>14,33%</td>
<td>6,94%</td>
<td>64,54%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Based on an average of the years 2006-2008

- the processes of DANS are based on the Open Archival Information System (OAIS) Reference Model
Why another model?

- focus on research datasets of arts, humanities and social sciences
- a Data Archiving institute as an independent entity (not a library, not an educational institution)
- costs measured in euro’s per dataset
- activity specific cost drivers to avoid indirect costs
- concept of “matrix of dataset complexity” is introduced
- results analyzed with the help of a balanced-scorecard
Components

- ABC Model (Cooper and Kaplan (1988))
  - improving tactical and strategic decision-making
  - understand the use of scarce organisational resources in various business activities

- Balanced Scorecard (BSC) (Kaplan and Norton, 1997)
  - translates an organisation’s mission and existing business strategy into a limited number of specific strategic objectives that can be linked and measured operationally.
Components of an ABC model

- resources are first traced to activities and activity costs are then traced to products/services
organisational resources are logically grouped into resource pools
  - Office, Data Acquisition, IT Services & Equipment and Staff

Resource cost drivers have been created only for the resource pool “Staff” Based on activity and salary.
  - Archive, General, ICTa and ICTb
Activities of DANS are listed in the Activity-based Reference Model of DANS
- It is a reflection of the operational functions taking place in DANS

Activity cost drivers are the elements that enable the calculation of future costs
- They are used in the model as an input which activates the estimation process
ABC model for DANS

Total Costs

- Data Acquisition
- Office
- Staff
- IT Equipment & Services

Archivists
- e.g. Project Acquisition
  - e.g. #of partners

General
- e.g. Negotiate Submission
  - e.g. -attitude of researcher

ICTa
- e.g. Maintenance of Archival System
  - e.g. #of functions

ICTb
- e.g. Ingest
  - e.g. -metadata completeness

- e.g. Project Management
  - e.g. #of employees

Dataset of Archaeology
Dataset of Humanities
Dataset of Social Sciences
## Activity based reference model

<table>
<thead>
<tr>
<th>Category</th>
<th>Activities</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NETWORKED SERVICES</strong></td>
<td>Project Acquisition</td>
<td>DANS: Market research for projects that could potentially contribute to the aims of the organisation (e.g. generate research data).</td>
</tr>
<tr>
<td></td>
<td>Dissemination</td>
<td>DANS: Increasing awareness amongst the potential data consumers and data producers.</td>
</tr>
<tr>
<td></td>
<td>Preparation Projects</td>
<td>DANS: Preparing funding proposals and applications for projects that are internal initiatives.</td>
</tr>
<tr>
<td></td>
<td>Interorganisational Assistance and Liaison</td>
<td>DANS: Providing technical assistance and information in liaison with other organisations, private individuals, educational institutions, etc.</td>
</tr>
<tr>
<td></td>
<td>Indirect acquisition</td>
<td>DANS: Promotion of data-preservation-and-access-friendly ideology that will potentially result into more research data deposits (initialization of policy making, depositor support, etc.).</td>
</tr>
<tr>
<td></td>
<td>Direct Acquisition</td>
<td>KRDS: The processes involved in acquiring research data for an archive.</td>
</tr>
<tr>
<td></td>
<td>Negotiate submission</td>
<td>LIFE: The specification of submission requirements for producers/depositors together with communication and negotiation with producers/depositors.</td>
</tr>
<tr>
<td><strong>DATA ACQUISITION</strong></td>
<td>Maintenance of archival system</td>
<td>DANS: Maintaining the archival system supporting the core services needed to operate and administer the datasets.</td>
</tr>
<tr>
<td></td>
<td>Development of Archival System</td>
<td>DANS &amp; KRDS: Further developing the archival system; providing an interface between new application software and the platform.</td>
</tr>
<tr>
<td></td>
<td>Improvement of dataset presentation/ access</td>
<td>DANS: Software and system design or standards development facilitating data discovery, presentation and access.</td>
</tr>
<tr>
<td></td>
<td>Functional management of the technical infrastructure</td>
<td>DANS: Selection of and communication with ICT and storage services providers. Hardware maintenance and setting up of service processes.</td>
</tr>
<tr>
<td></td>
<td>Ingest</td>
<td>NASA CET &amp; OAI: The &quot;ingest&quot; functional area includes receiving, reading, quality checking, cataloguing, of incoming data (including metadata, documentation, etc.) to the point of insertion into the archive. Ingest can be manual or electronic with manual steps involved in quality checking, etc.</td>
</tr>
<tr>
<td></td>
<td>Archival Storage</td>
<td>DANS: Error checking and correction and update of Archival Information Package (AIP) entries.</td>
</tr>
<tr>
<td><strong>ARCHIVING</strong></td>
<td>Data Management</td>
<td>OAI: The services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive.</td>
</tr>
<tr>
<td></td>
<td>Access</td>
<td>OAI: Services and functions for populating and accessing a wide variety of data by the project. Services and functions which make the archival information holdings and related services visible to Consumers.</td>
</tr>
<tr>
<td></td>
<td>Preservation</td>
<td>KRDS: The services and functions for monitoring, providing recommendations, and taking action, to ensure that the information stored in the archive remains accessible over the long term, even if the original computing environment becomes obsolete. Preservation covers the process of performing actions on digital objects in order to ensure their continued accessibility. It includes evaluation and quality assurance of actions, and the acquisition or implementation of software to facilitate the preservation actions.</td>
</tr>
<tr>
<td></td>
<td>Archival Administration</td>
<td>KRDS: User support to users of the archival system and establishing and maintaining the archive's standards and policies (e.g. initial format standards, documentation standards, model deposit agreements, the archive’s selection policy and procedures to be followed during the Ingest process). The latter normally involve a large initial effort to develop and then regular review and small updates over time and rarer major re-drafting.</td>
</tr>
<tr>
<td><strong>ADMINISTRATIVE</strong></td>
<td>Administrative Support</td>
<td>KRDS: Administrative support and control provided by office managers, personal assistants and secretaries. Management of supply chain, movement of goods, and</td>
</tr>
</tbody>
</table>
## Detailed activity

<table>
<thead>
<tr>
<th>Cat.</th>
<th>Activity</th>
<th>Description</th>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHIVING</td>
<td>Access</td>
<td>OAIS: Services and functions which make the archival information holdings and related services visible to Consumers.</td>
<td>Publish data</td>
<td>DANS: Publish the processed data. Specify type of access (e.g. “Restricted” for privacy-sensitive/protected files).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Check AIP</td>
<td>OAIS: Check the AIP through search functions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Create jump off page</td>
<td>DANS: Illustrate the contents of the dataset.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inform depositor</td>
<td>DANS: Notify depositor of AIP publication and have depositor check the AIP.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Storage hierarchy management</td>
<td>DANS: Link dataset to related datasets and/or external pages/archives.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transfer to another archive</td>
<td>KRDS: Transfer material to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.</td>
</tr>
</tbody>
</table>
Connecting activities to cost objects

- For each activity there is one or more cost driver connecting the activity to the cost object

<table>
<thead>
<tr>
<th>Category</th>
<th>Activity</th>
<th>Tasks</th>
<th>Activity Cost Drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHIVING</td>
<td>Ingest</td>
<td>- Receive submission (SIP)</td>
<td>- Number of files</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Quality check</td>
<td>- Completeness of metadata (X entries/ 4 obligatory entries)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Verification SIP creation</td>
<td>- Complexity of metadata. DDI3.0, Dublin Core</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Number of file formats (spps, mpeg4, GIS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Size of file formats</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Complexity of file formats</td>
</tr>
</tbody>
</table>


How to estimate costs

(1) WHO WILL THE PEOPLE TO WORK ON THIS DATASET BE?
(Please choose a category)

- ARCHIVIST
- GENERAL
- ICTa

(2) WHAT KIND OF ACTIVITIES WILL EACH ONE OF THE RESOURCE DRIVERS PERFORM?
(Please choose activities)

- Direct Acquisition
- Ingest
- Maintenance of archival system
- Preservation
- General Mgmt

(3) BASED ON PREVIOUS SELECTIONS THE INVOLVED ACTIVITY DRIVERS APPEAR
(Please fill in)

- Attitude of researcher
- #of functions
- Metadata completeness
- #of files
- #of employees

(4) WHAT ARE THE ATTACHED COSTS OF THESE ACTIVITIES?
(business sustaining costs are calculated automatically)

- Rent (services included)
- Banking
- IT services

(5) COSTS OF SPECIFIED DATASET APPEAR

X euros / dataset of social sciences
The Balanced ScoreCard uses Success Factors to indicate the success of the mission of the organization.

The mission of DANS is to improve the research data infrastructure of social sciences and humanities.
Four perspectives

- By viewing the organisation from different perspectives, the balanced scorecard provides a more comprehensive understanding of its current performance.
Succes Factors

Each of the 15 illustrated Success Factors (SF) is described further by a set of Performance Indicators (PIs).

<table>
<thead>
<tr>
<th>Perspective</th>
<th>Success Factor</th>
<th>Performance Indicator</th>
<th>Target</th>
<th>Current State</th>
<th>Activity Connected</th>
<th>Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENABLERS/ USERS</td>
<td>Increase number of datasets available to end user</td>
<td># of contracts signed with research funders</td>
<td>10</td>
<td>3</td>
<td>-Direct Acquisition</td>
<td>-Check the procedure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-Indirect Acquisition</td>
<td>-Allocate more staff</td>
</tr>
</tbody>
</table>
Results

- Expanding the scope of DANS
  - Social Sciences and Humanities to all sciences.
- ABC model gives information about differences in different disciplines.
Archiving in different disciplines

History
- Access 41%
- Archival Storage 0%
- Data Management 0%
- Preservation 9%
- Ingest 4%

Social Sciences
- Archival Admin. 8%
- Archival Storage 23%
- Data Management 25%
- Access 6%
- Preservation 0%

Archaeology
- Archival Admin. 1%
- Archival Storage 12%
- Data Management 44%
- Access 6%
- Preservation 1%
- Ingest 36%
- Ambition level of the Academy of Science and (KNAW) and the Netherlands Organization for Science (NWO)
  - First mover
  - BSC model will give information what is needed.
  - Link between BSC, ABC model what it costs.
% of money spent on each activity
Next Steps

- Refine **cost drivers**
- Allocate the **other-than-staff costs** to activities
- Develop further the **“matrix of dataset complexity”**
- Consider adding **more cost objects**
- Test reliability and accuracy
- Develop **software** to make ABC easy to use
Thank you for your attention!