Cost and Value analysis of digital data archiving

ANNA PALAIOLOGK

Charles Beagrie

ECORYS

Motivation

Detailed and meaningful **cost information** allows:

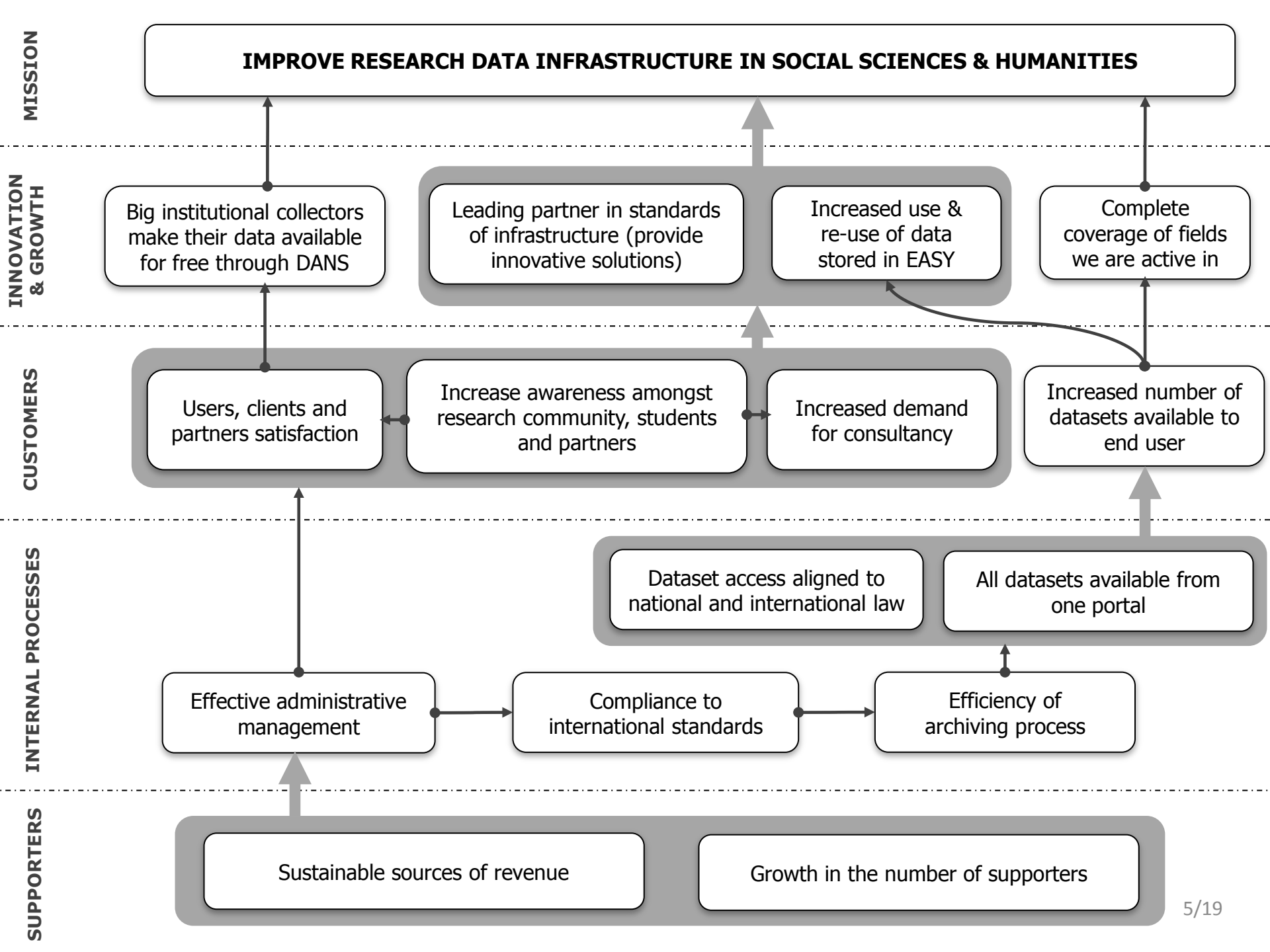
- more accurate planning
- better forecasting and control
- more accountability and transparency
- prioritise/control the level of ambition - realistic strategy (e.g. collection levels and preservation aims, quality-quantity balance, etc)

Challenges + Terminology

- funding does not grow in line with information growth
- curation vs. storing of the data
- acquisition and ingest
- guidelines vs. regulation on preferred formats
- legal requirements and grant terms
- access - most variable area of costs

DANS case study

- Data Archiving & Networked Services (DANS) is an institute of the Dutch Royal Academy of Arts and Sciences (KNAW)
- an independent digital archive
- collection: 14.000 datasets (1,5 TB) available to public and 10 datasets (20 TB) not available to the public
- 51 employees
- work processes based on Open Archival Information System (OAIS) - ISO 14721:2003
- mixed budget of approximately 3,8 million euro/ year
- costs measured in Euros per dataset
- next slide depicts **processes** and **vision** of DANS

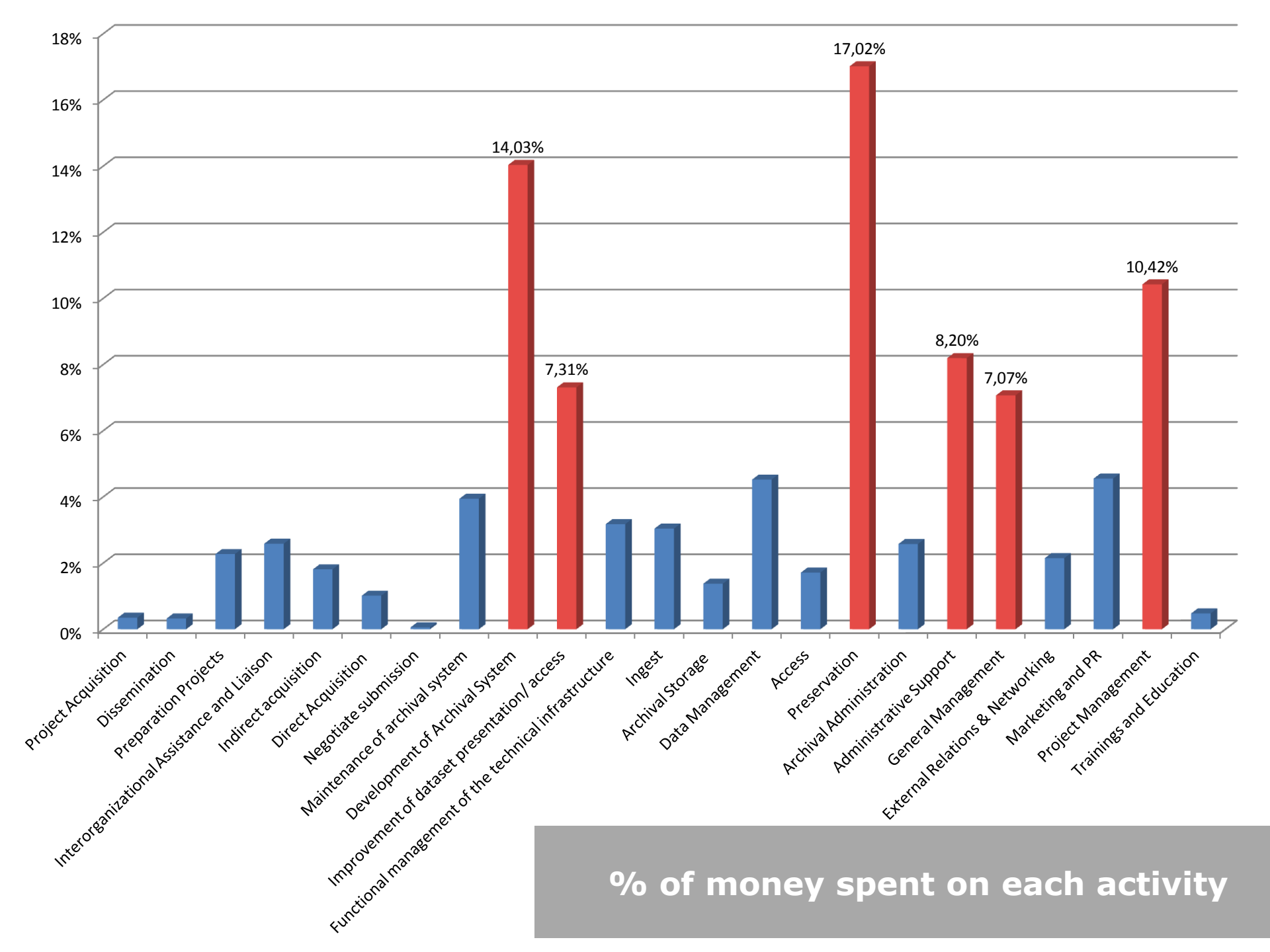

MISSION
IMPROVE RESEARCH DATA INFRASTRUCTURE IN SOCIAL SCIENCES & HUMANITIES
INNOVATION & GROWTH
Big institutional collectors make their data available for free through DANS
Leading partner in standards of infrastructure (provide innovative solutions)
Increased use & re-use of data stored in EASY
Complete coverage of fields we are active in
CUSTOMERS
Users, clients and partners satisfaction
Increase awareness amongst research community, students and partners
Increased demand for consultancy
Increased number of datasets available to end user
INTERNAL PROCESSES
Dataset access aligned to national and international law
All datasets available from one portal
Effective administrative management
Compliance to international standards
Efficiency of archiving process
SUPPORTERS
Sustainable sources of revenue
Growth in the number of supporters

Budget distribution

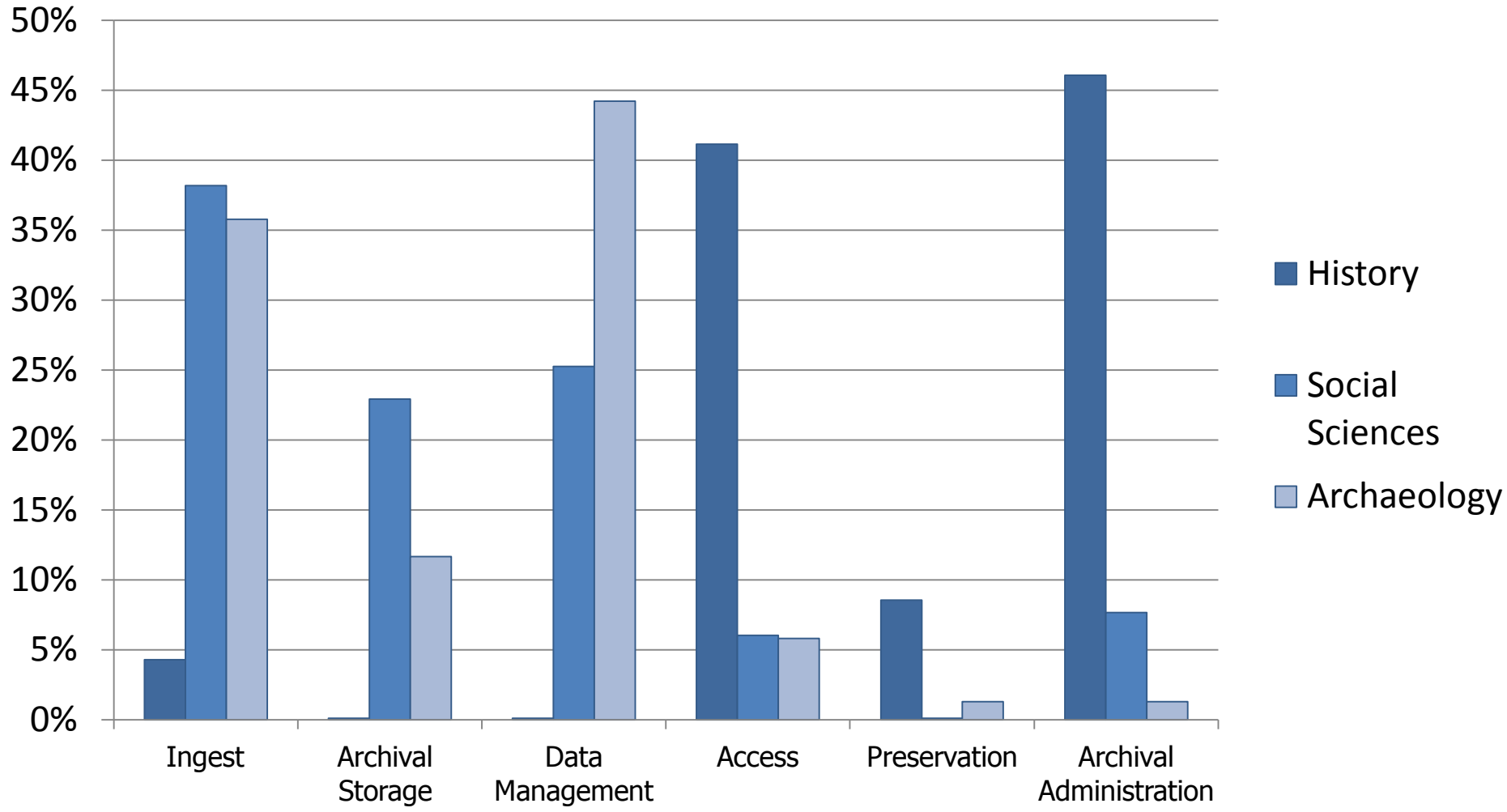
Staff is the major resource pool in digital archiving, up to 65–70% of total expenses

Data Acquisition	Office	IT services and equipment	Staff	Total
14,2%	14,3%	7%	**64,5%**	100%

Staff needs to do tasks bringing the most value. Rest needs to be automated.


18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
14,03%
7,31%
17,02%
8,20%
7,07%
10,42%
Project Acquisition
Dissemination
Preparation Projects
Interorganizational Assistance and Liaison
Indirect acquisition
Direct Acquisition
Negotiate submission
Maintenance of archival system
Development of Archival System
Improvement of dataset presentation/ access
Functional management of the technical infrastructure
Ingest
Archival Storage
Data Management
Access
Preservation
Archival Administration
Administrative Support
General Management
External Relations & Networking
Marketing and PR
Project Management
Trainings and Education
% of money spent on each activity

Workload allocation per discipline



Key findings

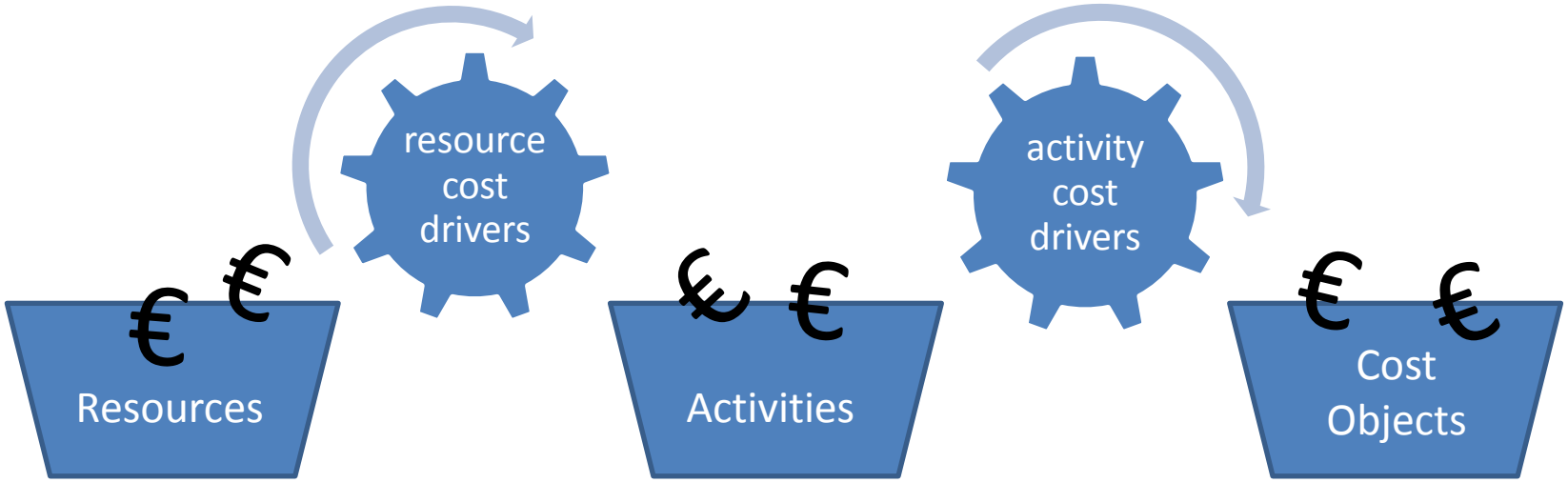
In long term data archiving:

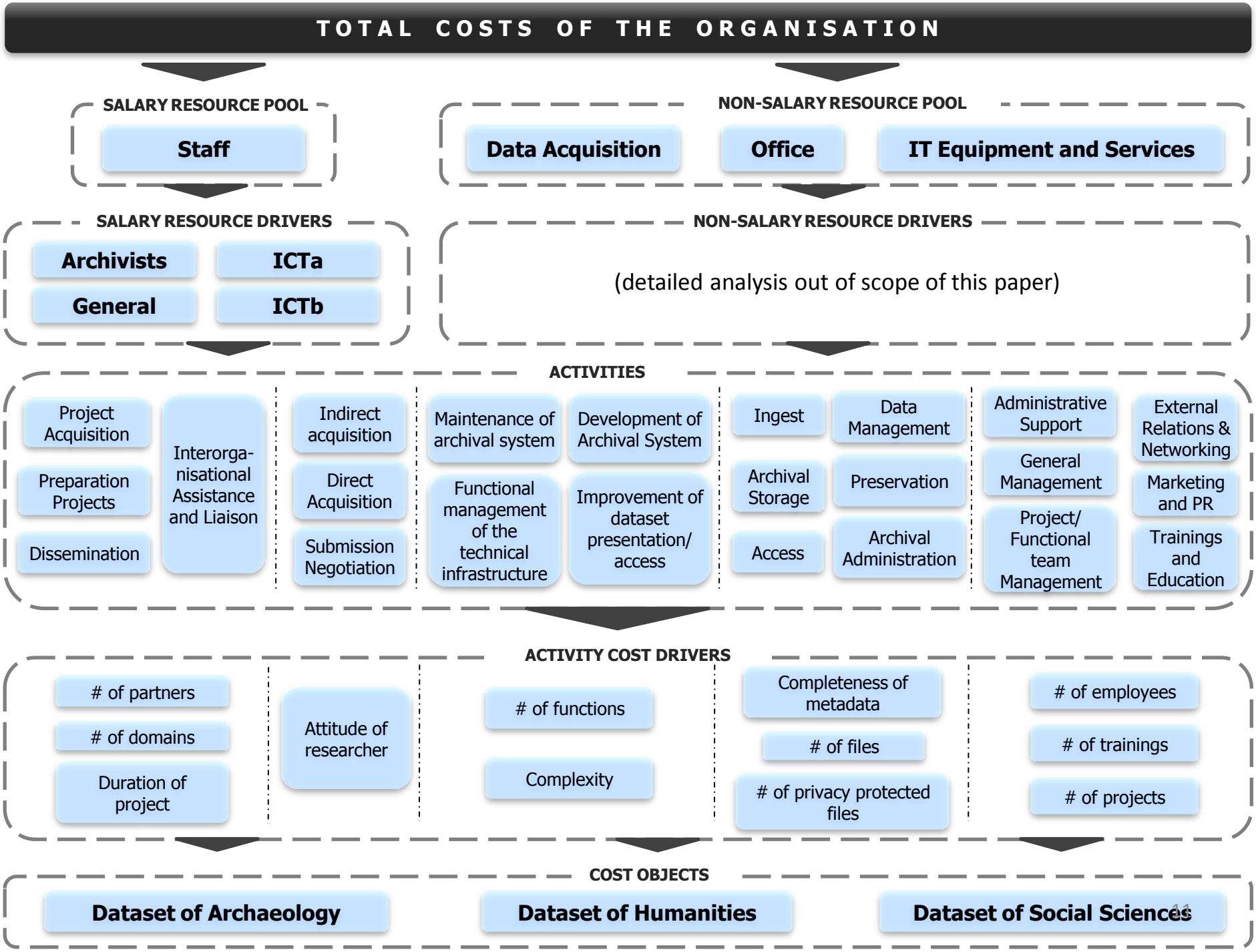
1. Up-front costs of acquisition and ingest of data (70-90% of total) dominate the long-term costs of storage and preservation.

2. Up-front costs dominated by staff time rather than hardware or other technology costs.

3. Long-term costs scale weakly, if at all, with the size of an archive. Preserving 10 TB is not that much more expensive than preserving 10 GB.

ABC methodology




TOTAL COSTS OF THE ORGANISATION
SALARY RESOURCE POOL
Staff
NON-SALARY RESOURCE POOL
Data Acquisition
Office
IT Equipment and Services
SALARY RESOURCE DRIVERS
Archivists
ICTa
General
ICTb
NON-SALARY RESOURCE DRIVERS
(detailed analysis out of scope of this paper)
ACTIVITIES
Project Acquisition
Preparation Projects
Dissemination
Interorga-nisational Assistance and Liaison
Indirect acquisition
Direct Acquisition
Submission Negotiation
Maintenance of archival system
Development of Archival System
Functional management of the technical infrastructure
Improvement of dataset presentation/ access
Ingest
Data Management
Archival Storage
Preservation
Access
Archival Administration
Administrative Support
General Management
Project/ Functional team Management
External Relations & Networking
Marketing and PR
Trainings and Education
ACTIVITY COST DRIVERS
# of partners
# of domains
Duration of project
Attitude of researcher
# of functions
Complexity
Completeness of metadata
# of files
# of privacy protected files
# of employees
# of trainings
# of projects
COST OBJECTS
Dataset of Archaeology
Dataset of Humanities
Dataset of Social Sciences

Practicalities of ABC methodology

ABC data collection "How-To"

- Dedicating a person to be responsible for collecting the cost information
- Do not overwhelm staff with information
- Do not expect all staff to be "on the same page" from the beginning
- Run a trial for a day or a week
- Ask staff to report separately on activities outside the Model
- Allow for a general comments field
- Leave, sickness or absence should be specified separately

Value + Economic Impact Analysis

Methods being applied to:

- *report published*

- *in progress*

- *in progress*

Benefits data collection

Desk-research sources:

- Organisation and infrastructure evaluation reports
- Documentation on data usage and users
- Internal (management) reports
- Annual and mid-term reports

Interviews with:

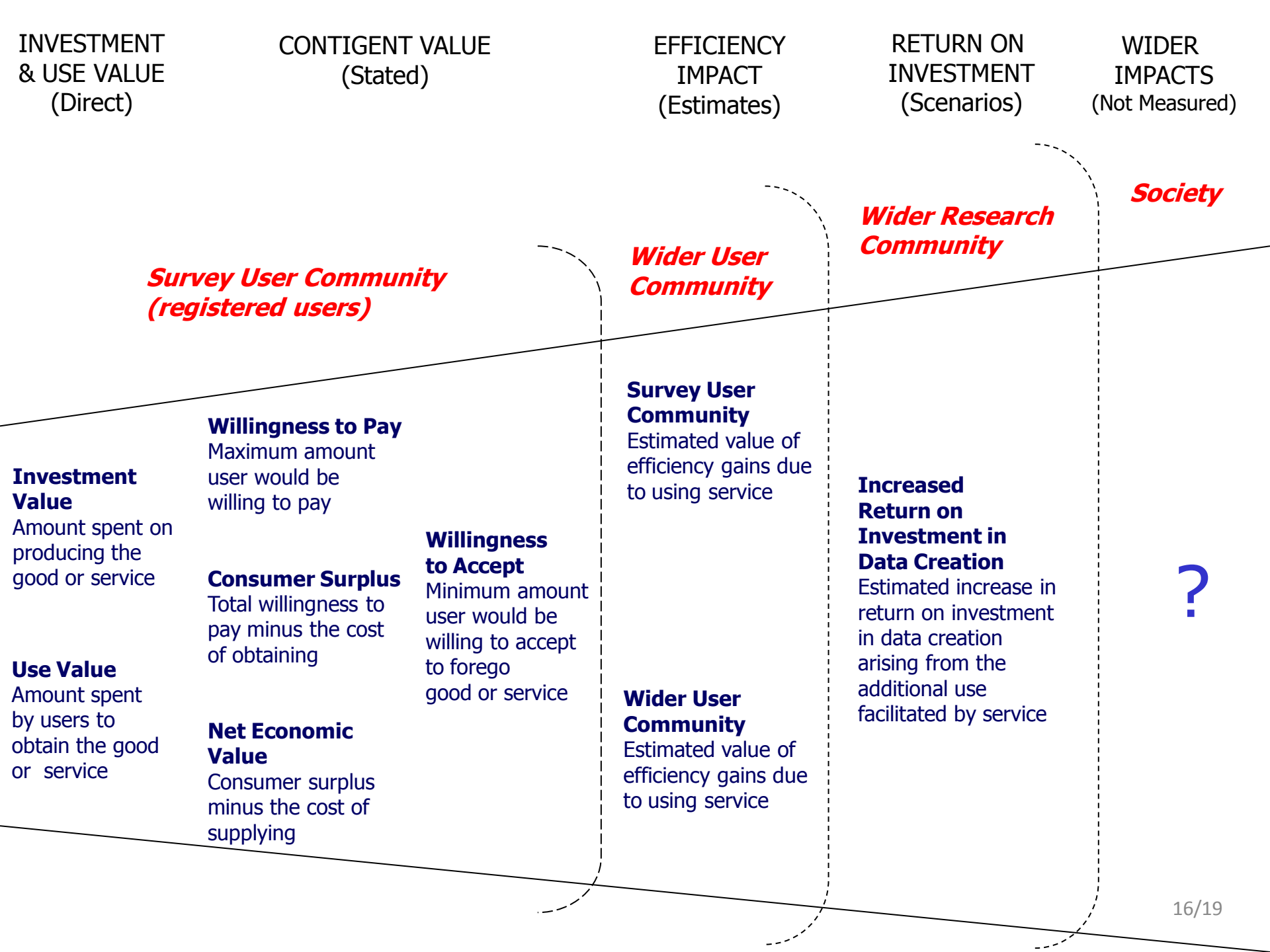
- Organisation management and staff
- Policy makers and practitioners
- Government institutions
- Non-academic and private sector representatives

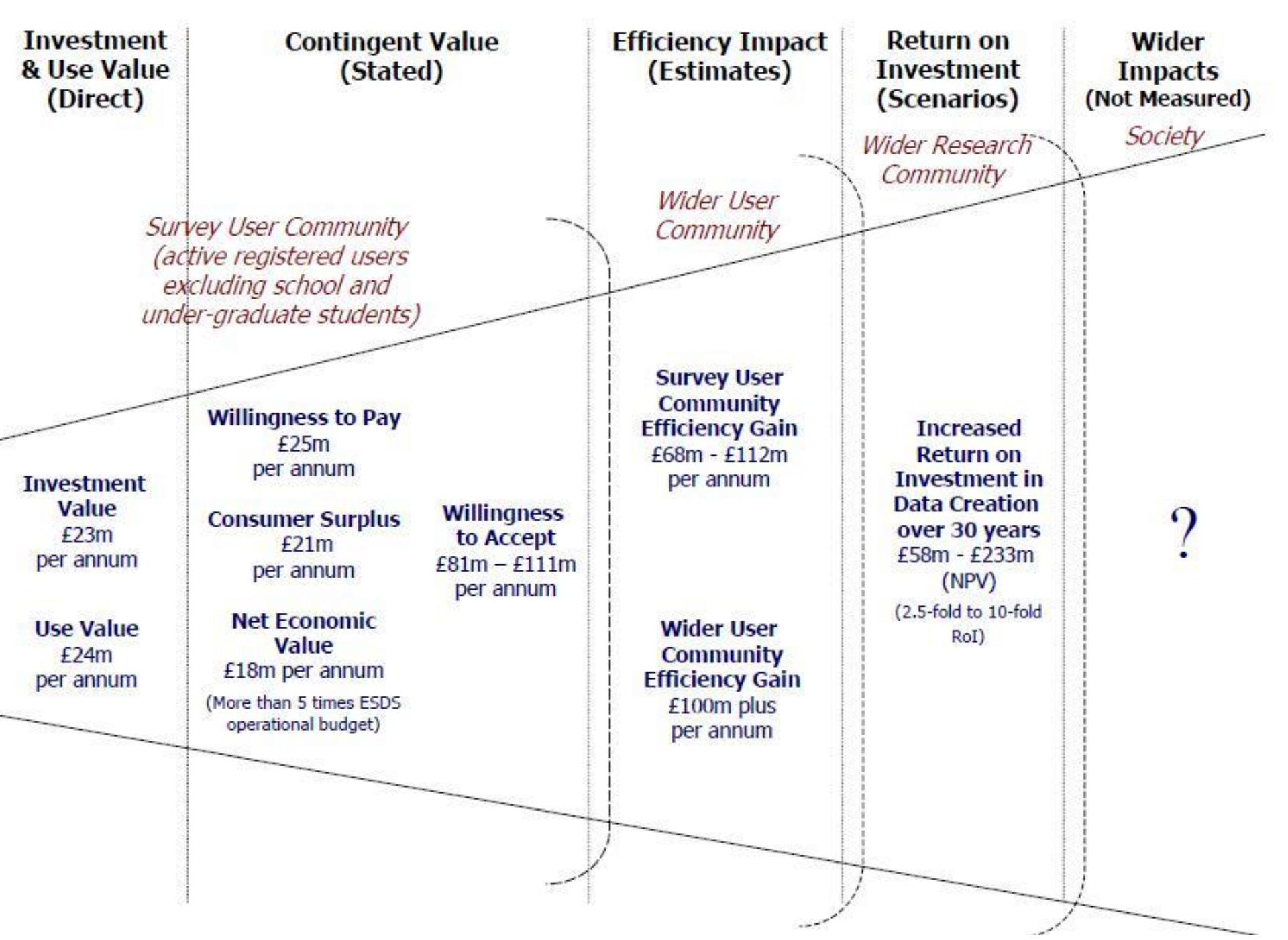
Online-survey addressed to:

- Depositors and users

Economic measures of value

- **Investment value**: annual operational funding & the costs that depositors face in preparing data for deposit and in making that deposit
- **Use value**: average user access costs **x** number of users
- **Contingent value**: the amount users are "willing to pay“ or “willing to accept” in return for giving up access
- **Efficiency gain**: user estimates of time saved by using the Data Service resources
- **Return on investment**: estimated return with time (30yrs)


INVESTMENT & USE VALUE (Direct)
CONTIGENT VALUE (Stated)
EFFICIENCY IMPACT (Estimates)
RETURN ON INVESTMENT (Scenarios)
WIDER IMPACTS (Not Measured)
Survey User Community (registered users)
Wider User Community
Wider Research Community
Society
Investment Value
Amount spent on producing the good or service
Use Value
Amount spent by users to obtain the good or service
Willingness to Pay
Maximum amount user would be willing to pay
Consumer Surplus
Total willingness to pay minus the cost of obtaining
Net Economic Value
Consumer surplus minus the cost of supplying
Willingness to Accept
Minimum amount user would be willing to accept to forego good or service
Survey User Community
Estimated value of efficiency gains due to using service
Wider User Community
Estimated value of efficiency gains due to using service
Increased Return on Investment in Data Creation
Estimated increase in return on investment in data creation arising from the additional use facilitated by service
?


Investment & Use Value (Direct)
Contingent Value (Stated)
Efficiency Impact (Estimates)
Return on Investment (Scenarios)
Wider Impacts (Not Measured)
Wider Research Community
Society
Wider User Community
Survey User Community (active registered users excluding school and under-graduate students)
Survey User Community Efficiency Gain
£68m - £112m per annum
Willingness to Pay
£25m per annum
Investment Value
£23m per annum
Consumer Surplus
£21m per annum
Willingness to Accept
£81m – £111m per annum
Increased Return on Investment in Data Creation over 30 years
£58m - £233m (NPV)
(2.5-fold to 10-fold RoI)
?
Use Value
£24m per annum
Net Economic Value
£18m per annum
(More than 5 times ESDS operational budget)
Wider User Community Efficiency Gain
£100m plus per annum

Next steps

Costs

- Refine cost drivers
- Allocate the other-than-staff costs to activities
- Experiment with other cost objects
- Develop the "matrix of dataset complexity"
- Apply economic adjustments
- Test reliability and accuracy
- Develop/Customise software to make ABC easy to use

Value/Benefits

- Develop the benefits framework further
- Collect more diverse/detailed data
- Verify results

References

1. Rob Baxter, *EUDAT Sustainability Plan*, WP2-D2.1.1-v.1.0, October 2012. Available from: http://www.eudat.eu/deliverables/d211-sustainability-plan

2. Neil Beagrie, Julia Chruszcz, Brian Lavoie and Matthew Woollard*, Keeping Research Data Safe 1 and 2, 2008 and 2010.* Available from: *http://www.beagrie.com/krds.php*

3. Neil Beagrie, John Houghton, Anna Palaiologk and Peter Williams, *Economic Impact Evaluation of the Economic and Social Data Service*, 2012. Available from: *http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf*

4. Anna Palaiologk, Anastasios Economides, Heiko Tjalsma and Laurents Sesink, *An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS*, International Journal on Digital Libraries, 2012. Available from: *http://link.springer.com/content/pdf/10.1007%2Fs00799-012-0092-1*