Datanomics
Costs and Value of Research Data

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Neil Beagrie
(Charles Beagrie Ltd)
Counting on Reproducibility
DPC/Jisc Briefing Day, Birmingham, May 2019
Some (KRDS) Partners
Outline

• Costs

• Valuing Intangible Assets (Quantitative Approaches)

• Where Do We Go From Here?
  » “What to Keep” study
  » Infonomics – an industry perspective

• Concluding thoughts
Costs
Costs

Activity Cost Models

- Many examples – a few generic (intended for broadly based community), most organisation-specific (derived)
- Substantial effort to create

Activity Cost Data

- Can be created in a consistent form using a ACM
- Cost Data still takes significant effort to collect and may be incomplete. “Total Costs of Curation” can be distributed across many budget centres/departments

Cost Trends

- Cost data can give trends and “Laws” or “rules of thumb” that are very powerful tools
Effort and Use Knowledge Pyramid: for Costs
Costs Rules of Thumb (1)

- Rules for a prolonged but not eternal period of time ("Laws")
- "Kryder’s Law" – disk storage roughly halving in cost every year (comparable to Moores Law for processing power)
- A “re-set” in Kryder’s Law from 2010 onwards documented by Rosenthal and Gupta

Disk cost-per-byte

Kryder slowdown. David Rosenthal. Chart by Preeti Gupta at UCSC
Costs: KRDS Laws/Rules of Thumb

1. Getting data in takes about half of the lifetime costs, preservation about a sixth, access about a third.
2. Preservation costs decline over time.
3. Fixed costs are significant for most data archives.
4. Staff are the most significant proportion of archive costs.

Note recent Dutch Digital Heritage Network research provides further independent validation of “KRDS Laws”
Valuing Intangible Assets
Valuing Intangible Assets

- Valuable approach to digital preservation and intangibles by Laurie Hunter and since adapted for research data
- We measure value of data services not just data alone
- Measuring value of intangible assets is hard – much harder than for physical assets
- Economic methods are well established but difficult to get the cost and value data to use in them
- Counter-factuals – a baseline – are important
- Collaboration with John Houghton to move beyond qualitative value to financial measures of value
Tangible and Intangible Assets

Two Views of Data Archives

Physical Assets
- Buildings / Equipment
- Grid / ICT Networks
- Staff

Intangible Assets
- Data
- Intellectual Capital
- Technical/Organisational Environment
- Organisational Capital

Skills & Training
- Human Capital

Professional Networks
- Relationship Capital

Illustration by Charles Beagrie Ltd ©2016 incorporating images by Jorgen Stamp digitalbewaring.dk. CC-BY licensed
Value + Economic Impact Analysis

John Houghton (Victoria University) + Neil Beagrie (Charles Beagrie Ltd) 4 joint studies to date. Methods applied to:

- Economic & Social Data Service (ESDS)
- Archaeology Data Service
- British Atmospheric Data Centre
- European Bioinformatics Institute
Economic Metrics Used

- **Investment value**: annual operational funding plus the costs that depositors face in preparing data for deposit and in making those deposits
- **Use value**: weighted average user access costs multiplied by the number of accesses
- **Contingent value**: the amount users are "willing to pay“ for 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 in the data service**: standard ROI
- **Return on investment in the data creation**: the estimated increase in return on investment to the funder(s) in the data creation due to the additional use facilitated by the data service
Economic Methods Applied

### Investment & Use Value (Direct)
- **Investment Value**: Amount spent on producing the good or service
- **Use Value**: Amount spent by users to obtain the good or service

### Contingent Value (Stated)
- **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

### Efficiency Impact (Estimated)
- **Willingness to Accept**: Minimum amount user would be willing to accept to forego good or service

### Return on Investment in the Data (Estimated)
- **User Community**: Estimated value of efficiency gains due to using the good or service
- **Range of Time Savings**: (from time spent with data from the centre to overall work time)

### Wider Impacts (Not Directly Measured)
- **Society**: Estimated increase in return on investment in data creation arising from the additional use facilitated by the data centre
ESDS Value/Impact Analysis

Benefit/cost ratio of net economic value to ESDS operational costs

£5.40 to £1
ESDS Study: Researcher Efficiency Gains

Impact of using ESDS data and services on research efficiency
(after Beagrie et al 2012, p77, Figure 15)
Economic and Social Research Council © 2012 CC-BY licensed
“Ideally, economic impact assessments should estimate the counterfactual – i.e. what would occur in the absence of the facility...However, counterfactuals are rarely addressed in the [c.100] studies reviewed due to lack of data. We found two exceptions that address this issue partially. One is the evaluation of the economic impacts of ESDS (2012) which partially explores the counterfactual through a users’ survey...Another exception is a review of economic impacts of large-scale science facilities in the UK (SQW, 2008) ... however, this estimation is not done rigorously and relies mostly on the estimation of the local benefits.”

*Big Science and Innovation - Report to BIS - Technopolis 2013*
## Costs of Inaction

### Costs of Inaction: reported metrics for archiving via individual researchers

<table>
<thead>
<tr>
<th>Category</th>
<th>Metric Description</th>
<th>Rate of Loss</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute loss</td>
<td>Rate of loss of research data sets</td>
<td>17% per annum</td>
<td>(Vines et al 2014)</td>
</tr>
<tr>
<td></td>
<td>Rate of loss of working contact emails</td>
<td>7% per annum</td>
<td>(Vines et al 2014)</td>
</tr>
<tr>
<td></td>
<td>Rate of loss for web-links to data on personal websites</td>
<td>c. 5.5% per annum</td>
<td>(Pepe et al 2014)</td>
</tr>
<tr>
<td>Partial information loss</td>
<td>Data requests fulfilled</td>
<td>25.7%</td>
<td>(Wicherts et al 2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44%</td>
<td>(Krawczyk and Reuben 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>59%</td>
<td>(Vines et al 2013)</td>
</tr>
<tr>
<td>Access</td>
<td>Elapsed time to fulfill data requests</td>
<td>Up to 6 months</td>
<td>(Wicherts et al 2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within 1-3 weeks (mean 7.7 days)</td>
<td>(Vines et al 2013)</td>
</tr>
</tbody>
</table>

Illustration by Charles Beagrie Ltd ©2017. CC-BY licensed

Although these reported metrics are from studies of different disciplines and study dates, they contrast sharply with the excellent preservation record, very high fulfilment rates, and rapid online access rates of public data archives in the social sciences. The public data archives also are appreciating as opposed to deprecating assets with improving rather than decreasing trends in value over time.
Where Could We Go From Here?
What to Keep

Recent Jisc research data study
Recommendations

• **Recommendation 4**: Investigate the relative costs and benefits of differential curation levels, storage, or appraisal for what to keep for the two major use cases (Research Integrity, and Reuse) identified in the study.
Levels of Curation

US National Science Board 2005 Long-lived Data Collections

two-tier system with differential curation levels used by the UKDS or the DANS data archive’s systems - DataverseNL for short-term data management (up to 10 years) and EASY for long-term archiving, in the Netherlands. Both these examples in the UK and the Netherlands have different time horizons (how long the data is kept), costs in terms of metadata and preservation care (how it is kept) for their two systems, with the option to move from short-term to long-term systems and curation levels after future appraisal (or alternatively be maintained in their existing short-term system/ or deleted).
An industry-centric view of the value of information
Accounting for Information

Some Infonomics Quotes

• “Five or six decades since the beginning of the Information Age, the namesake of this age, and the major asset driving today’s economy, is still not considered an accounting asset”

• “Corporations typically exhibit greater discipline in tracking and accounting for their office furniture than their data”

• Bottom line - Data stewards are not alone in seeing this as an anomaly. There are others pressing for changes to insurance and accounting practices.
Gartner’s Information Valuation Models

**Foundational Measures**
- What to Keep study, NERC Data Value Checklist, etc

**Costs of Inaction**

**Financial Measures**
- Investment Value
- Contingent Valuation
- Return on Investment (ROI)

**Leading Indicator**
- Focused on improving information management discipline

**What is your objective for valuing information?**

**Lagging Indicator**
- Focused on improving information's economic benefits

From *Why and How to Measure the Value of Your Information Assets* by Douglas Laney

#GartnerSYM

© 2019 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and ITxpo are registered trademarks of Gartner, Inc. or its affiliates.
Conclusions

• We can use collections of cost data to look for trends – rules of thumb are probably the most widely useful cost information

• “Datanomics” and “Infonomics” have synergies - we may be able to leverage efforts within our community and industry

• Need to investigate the relative costs and benefits of differential curation levels, storage, or appraisal for the two major use cases (Research Integrity, and Reuse) identified in the What to Keep study.

• We have HSM in IT – in time can we look towards automating some decisions as Hierarchical Curation Management?
Further Information

• Costs, Benefits, and ROI for Research Data

• Economic Impact Studies of Research Data Services
  – The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres http://repository.jisc.ac.uk/5568/1/iDF308_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

• Douglas B. Laney 2017 Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage