Identifying Benefits from the Curation and Open Sharing of Research Data

Jenny Fry, Suzanne Lockyer and **Charles Oppenheim**
Information Science, Loughborough University
John Houghton and **Bruce Rasmussen**
Centre for Strategic Economic Studies, Victoria University, Melbourne

Loughborough University

JISC

CENTRE FOR STRATEGIC ECONOMIC STUDIES

Background

- Lyon (2007) need for a cost-benefit analysis of data curation and preservation infrastructure
- Beagrie et al (2008) costs of data repositories are an order of magnitude greater than those suggested for e-print repositories
- Lags between expenditure and impact make it difficult to present clear cost-benefit comparisons
- Research data are heterogeneous and disciplinary needs and practices vary greatly; comprehensive generalisable assessment is not possible
- Embedded case study approach; whereby particular data activities within specific disciplines can be examined as examples of the issues and possible ranges of costs and benefits

Aims and Objectives

- Identify the benefits of curating and sharing research data in an open access kind of way

- The project's objectives were to:
  - Identify a methodology to estimate benefits
  - Identify benefits to UK HE and the scientific community more broadly
  - Use the methodology to derive an estimate, expressed in financial terms, for each identified benefit
  - Document case studies and examples of data re-use, where that reuse led to *tangible* benefits

Costs/Benefits

	Costs	Benefits
Direct/ Indirect	Staff IT Infrastructure Tool Development (Search & Analytical) Training User Support	
Diffuse	Development of standards Ownership Time & Effort to Deposit Skills for Re-use	

Costs/Benefits

	Costs	Benefits
Direct/ Indirect	Staff IT Infrastructure Tool Development (Search & Analytical) Training User Support	Wider Access New Collaborations New Discoveries Increased Validation Reduced Duplication Knowledge industries Time Savings
Diffuse	Development of standards Ownership Time & Effort to Deposit Skills for Re-use	

Costs/Benefits

	Costs	Benefits
Direct/ Indirect	Staff IT Infrastructure Tool Development (Search & Analytical) Training User Support	Wider Access New Collaborations New Discoveries Increased Validation Reduced Duplication Knowledge industries Time Savings
Diffuse	Development of standards Ownership Time & Effort to Deposit Skills for Re-use	Transparency Education Enhanced Skills Data Quality Good Practice Enhanced Visibility

Benefits Accrue in Different Ways

- Cost savings
- Efficiency gains
- Opportunities to create value by doing things in new ways and by doing new things

- Successively more difficult to quantify given that;
  - They emerge over time
  - Can only be realised in the future

European Bioinformatics Institute

Funders	Core: EMBL External: Wellcome Trust, MRC, BBSRC, EU, NIH
Incentives	*Data Policies* - (actively support data sharing; e.g. availability of funding, explicit data policies) *Journals* - Submission of MIAME-compliant data to ArrayExpress has been adopted by most scientific journals as a condition for publishing a paper
Ownership	Priorities for sequence data are high quality and early release; general philosophy, but hold until publication mechanisms in place; population-based data complicated by confidentiality, privacy and IPR
Visibility	By-products, e.g. methods, are publishable – though this doesn't happen as often as could; scientists will cite EBI services as authority on method (using URL to website)
Benefits	Interesting discoveries e.g. human copy-number variation in genotypes, 'push-a-button' access.

Qualidata

Funders	ESRC and JISC
Incentives	Ethos of data sharing; ESRC projects must offer to UKDA (@30% rejected 07/08) - ESRC does not require data mgt plan; funding available for data preparation
Ownership	Consent issues contribute significantly to the rejection rate, researchers' personal relationship with data, initiatives to explore potential for qualitative data sharing
Visibility	Enhancing data sets (value-added services) - data that is just 'there', data which is enhanced (combined with other data), and data which may promote, e.g. themes, news items, workshops.
Benefits	Fixed capacity for processing, continued attachment to individual scholarship and publishable outputs, access is dialogue focused

	EBI	Qualidata
Holdings	Exponential growth: *'somewhere between enormous and terrifying'*	Limited; slower growth of acquisitions; selection based on 'show cases'; 48 datasets accepted in 2007-08
Curatorial tasks	Data integration (computer assisted)	Metadata creation at point of submission; enhanced data on specific collections (manually)
Usage	Various measures: Web hits (2,260,965); Unique investigators 300,000 to 1M.	Active users (47,635 whole UKDA), 179 user queries; 46 relating to usage (26%) and 133 to grant applications/depositing (74%)
Infrastructure costs	Mainly staffing; falling costs of IT; high costs of sequencing technologies	Mainly staffing
Cost savings	Reagents, equipment	Costs to set-up & data collection
Time savings (via value-added services)	Bioinformatics would not exist without data sharing; bench biologists 2 years lab work	Difficult to estimate for users. Enhancing data sets @20% additional work in addition to curation and preservation

Example I: Cost Saving

Direct cost savings from data use/re-use		Value
Data requirements:		
Cost of data collection/creation (faced by the depositor)	C1	£200,000
Cost of any additional preparation for sharing (faced by the depositor/repository)	C2	£10,000
Cost of searching for and accessing the data (faced by the user/withdrawer)	C3	£2,500
Annualised cost of storage of the data concerned (faced by the repository)	C4	£10,000
The life of the data in years	L	10
Number of times the data are used/re-used over the life-cycle	N	6
Direct cost savings (Step 1):		
Direct Costs = C1 + C2 + (C3 * N) + (C4 * L)	DC	£325,000
Direct Benefits = C1 * N	DB	£1,200,000
Direct benefit/cost ratio = (C1 * N) / (C1 + C2 + (C3 * N) + (C4 * L))	DBCR	3.7
Indirect cost savings (Step 2):		
Additional R&D spending = DB - DC	ARD	£875,000
Additional returns to R&D @ 20% = ADR * 0.20	AR	£175,000
Total benefits = DB + AR	TB	£1,375,000
Total benefit/cost ratio = TB / DC	TBCR	4.2

Notes to users:
Input data are the cost parameters C1 to C4, the estimated useful life-span of the data and estimates of likely use/re-use.
Simply enter your data into the box, over-writing the values (red text).

http://johnhoughtons.homeip.net/Data-Repository-Template.exe

Recommendations

1. Baseline Reporting
   - Development of guidelines for reporting
   - Classificatory work e.g. disciplinary factors
2. Model Questionnaire
   - Building on questions outlined in project report
   - Reduce duplicative effort with web based data gathering instrument
3. Developing a Community Resource
   - Model questionnaire as shared resource
   - Might include consistent collection of following types of data: annual acquisitions, annual usage, citations, external funds and annual spend

Acknowledgements

- This presentation is based on research funded by the JISC under the project title:

- "Identifying benefits arising from the curation and open sharing of research data produced within UK Higher Education and research institutes".