

The cost of not archiving

InNet, September 2014

Nick Thieberger
PARADISEC
School of Languages and Linguistics
University of Melbourne

Modeling costs of archiving

How to establish costs is hard enough

How to establish the *value* of archiving is harder still

Archives establish different kinds of collections

- from those with detailed metadata, and interoperable data structures
- to those with minimal metadata and heterogeneous data structures

Cost implications of enrichment



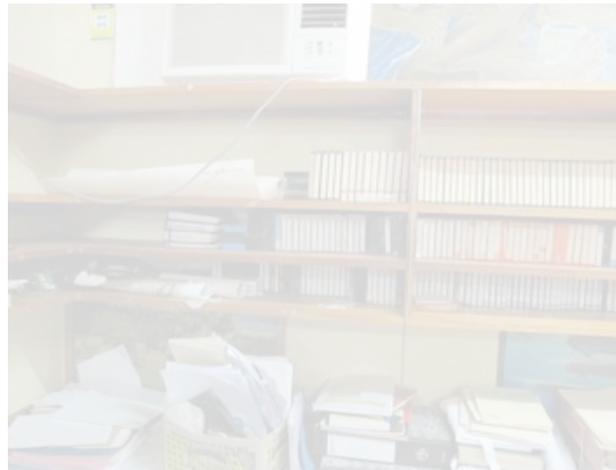
Solomon Islands Museum, 2012: ~800 tapes, most include oral tradition in local languages



Economics of creating collections

Grant to digitise 200 tapes, ~£10,000

Value of each tape is its inherent value (how to cost that?)
plus £50



The cost of an archival deposit

A collection donated to PARADISEEC can be valued, and its assessed monetary value can be claimed by the donor as a tax deduction.

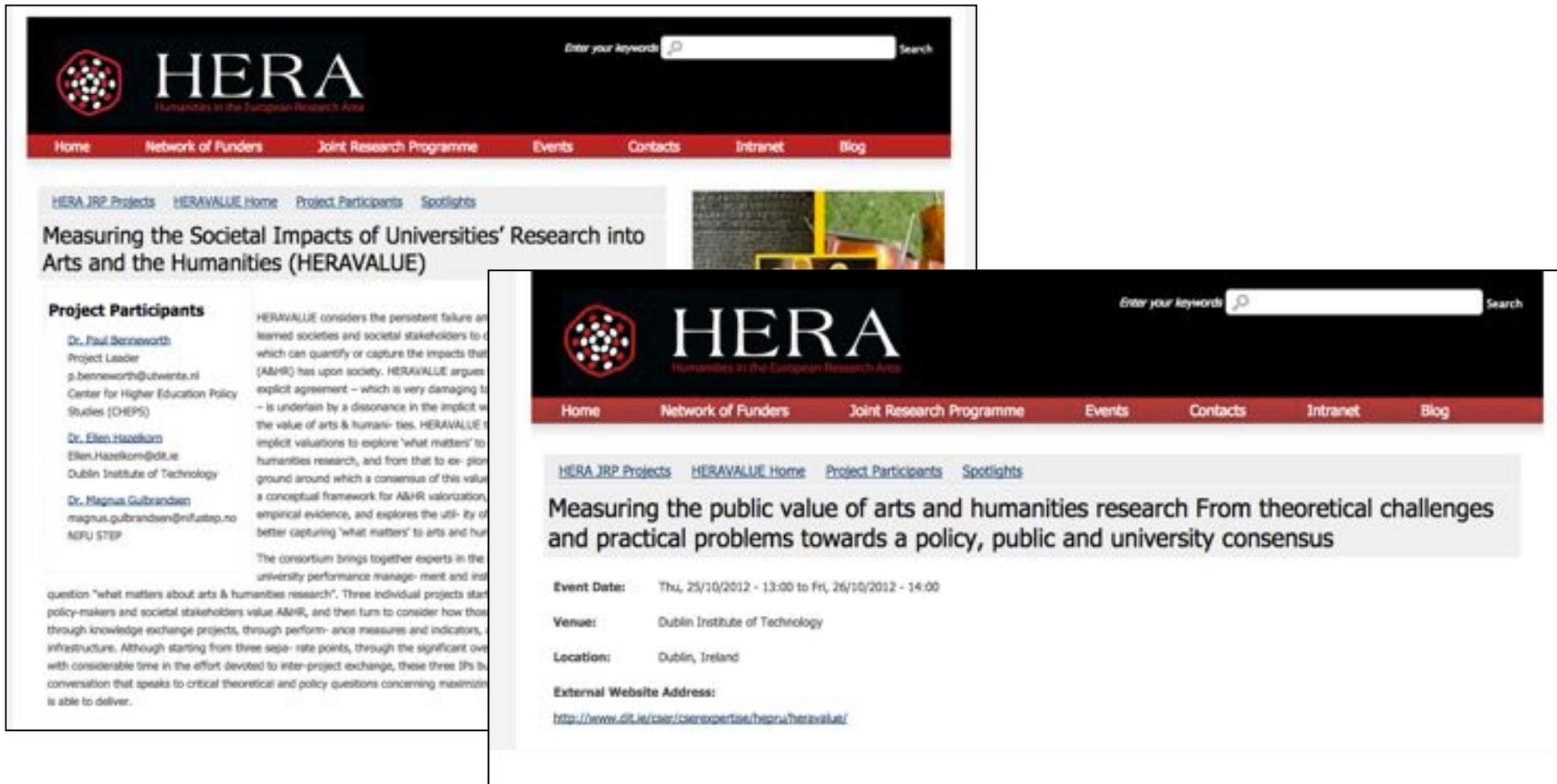
In general, an hour of field recording can be valued at \$200 or more, depending on the quality of the annotation, the amount of preparation required to make the recording, among other factors.

PARADISEEC asks that you consider donating a portion or the whole amount of the tax benefit you gain in this way. This donation is itself tax-deductible, and the donation will help to keep PARADISEEC functioning.

So, for example, if you pay 30% tax and you donate 30 tapes, each valued at \$200, they will reduce your taxable income by \$6,000. This saves you the tax you would have paid on \$6,000, or \$1,800. You then donate that to PARADISEEC and can claim a further \$1,800 reduction in your taxable income.

(proposed wording, based on the National Library of Australia's established system)

Bigger picture – valuing HASS research



The screenshot shows the HERA website interface. At the top, there is a search bar and navigation links: Home, Network of Funders, Joint Research Programme, Events, Contacts, Intranet, and Blog. The main content area features a header for 'HERAVALUE Home' and a title for the project: 'Measuring the Societal Impacts of Universities' Research into Arts and the Humanities (HERAVALUE)'. Below the title, there is a section for 'Project Participants' listing three individuals: Dr. Paul Bennenorth (Project Leader), Dr. Ellen Hazelkorn, and Dr. Magnus Gulbrandsen. The text describes the project's focus on valuing arts and humanities research and its societal impacts. A second screenshot on the right shows a detailed view of the project, including the event date (Thu, 25/10/2012 - 13:00 to Fri, 26/10/2012 - 14:00), venue (Dublin Institute of Technology), location (Dublin, Ireland), and external website address (http://www.dit.ie/csr/csr-expertise/hera/heravalue/).

<http://heranet.info/heravalue/index>

Importance of costing to establishing models for projects

“Estimations of costs that cover all aspects should be part of the planning process to limit the risk that a project ends up as yet another digital black hole, as so many others have done.”

Number of admin staff per terabyte of storage, between 1 per TB to 1 per 100 Tb (google. Internet Archive)

Jonas Palm (n.d.) The Digital Black Hole, http://www.tape-online.net/docs/Palm_Black_Hole.pdf

Costing: the big picture

“Acquisition and ingest costs most.

....

early preservation action during ingest or pre-ingest produces lower costs over the lifecycle as a whole.” (KRDS1, p.25; KRDS2, pp.31-52)

KRDS1: Beagrie, N., Chruszcz, J., and Lavoie, B. (2008), *Keeping Research Data Safe: a cost model and guidance for UK universities*, Final Report April 2008, available from <http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>

KRDS2: Beagrie, N., Lavoie, B., and Woollard, M. (2010), *Keeping Research Data Safe 2*, Final Report April 2010, available from <http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>

“Activity-based costing”

Estimate the costs of research datasets
preserved in a repository

How to express the true value of the repository
to its stakeholders

Don't just focus on file sizes and numbers

[possibly collaborate with economists to build up
a model for language archives]

Palaiologk, Anna S., Anastasios A. Economides, Heiko D. Tjalsma, and Laurents B. Sesink. 2012. An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. *Int. J. on Digital Libraries* 12(4):195-214

“Activity-based costing”

Outline 23 activities in 5 categories with activities, e.g. ; Archiving; Administration.

Networked services;

- Project acquisition
- Dissemination
- Preparation projects
- Interorganisational assistance and liaison

Data acquisition;

- Indirect acquisition
- Direct acquisition
- Submission negotiation

ICT (R&D and maintenance)

- Maintenance of archival system
- Development of archival system
- Improvement of dataset presentation/access
- Functional management of the technical infrastructure

Palaiologk, Anna S., Anastasios A. Economides, Heiko D. Tjalsma, and Laurents B. Sesink. 2012. An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. *Int. J. on Digital Libraries* 12(4):195-214

Value

We need to develop models of the value of
storing research materials

Time taken to create the records

Cost to the creator:

Fieldwork planning, cost of getting to the field

Equipment cost

Cost to the repository:

Not all candidate collections require equal input from the archive,
and effort can depend on technical characteristics of the material
(naming, organisation, metadata)

Monetary value=?

Analysis and enrichment of primary records

Cost of:

Metadata creation

Transcription

Interlinearisation

Monetary value=?

Inherent value of unique records of a minority language

Unique art and market value, is there any correlation with language performance?

In 1973 the UNESCO Musical Sources collection released an LP titled Solomon Islands: Fateleka and Baegu Music from Malaita, recorded in 1969 and 1970 by Hugo Zemp

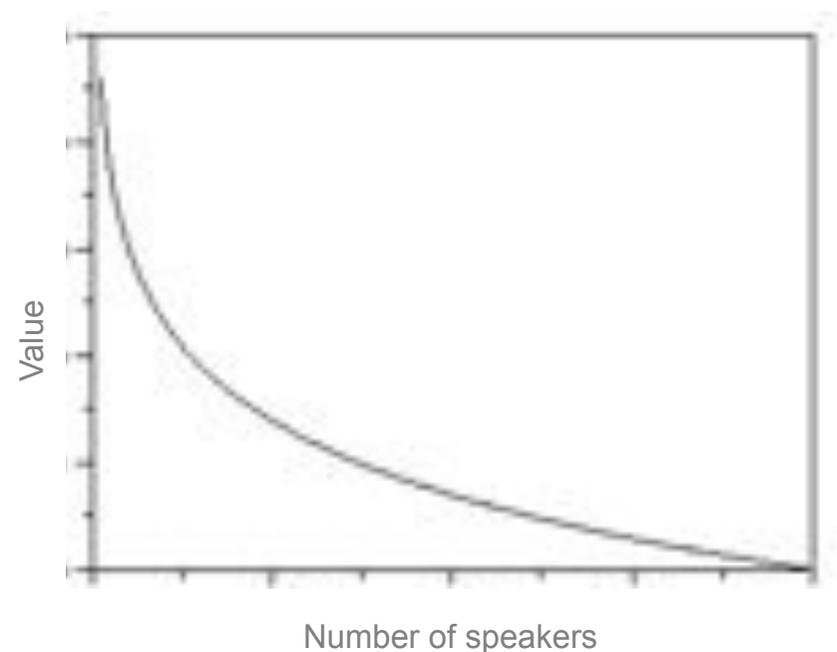
The song was used in a commercially successful CD (Deep Forest)

http://www.deepforest.co/dfpress_00-00-00sweetlullabyforworld.htm

Value to the source community of the records being accessible

As David Nash has noted (p.c.) the relationship of a person to their language becomes increasingly significant as the number of speakers declines, that is, the value of language records for languages with few or no speakers can rise inversely in proportion to the number of speakers or of people for whom it is an ancestral language

Monetary value=?



Value to the academic community of not being seen as expropriators of cultural information

- unlike the earlier model where the records are kept by a single researcher with no provision for access

Monetary value=?

Proper data management increases efficiency

“Data management reduces time and financial costs and greatly enhances the quality of the data you use too” (10)

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and Sharing Research Data: A Guide to Good Practice (Paperback)*. London: Sage Publications.

Data management and archiving

Creation of well-formed records prior to ingestion into the archive

Reduces costs

But: It relies on researchers knowing what to do

Requires a reward system for the creation of collections

cf ALS discussion <http://www.paradisec.org.au/blog/2012/11/counting-collections/>

Haspelmath, Martin and Susanne Michaelis. 2014. Annotated corpora of small languages as refereed publications: a vision [blog item:

<http://dlc.hypotheses.org/691>]

Use value

Key findings from the studies indicated that:

- “Very significant increases in research, teaching and studying efficiency were realised by the users as a result of their use of the data centres;
- The value to users exceeds the investment made in data sharing and curation via the centres in all three cases; and
- By facilitating additional use, the data centres significantly increase the measurable returns on investment in the creation/collection of the data hosted.”

Beagrie, Neil and John Houghton. 2014 *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/>

Use value

Archaeology Data Service (UK) suggests it has
a *use value* to its community of £7.4 million
per annum

Beagrie, Neil and John Houghton. 2014 *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/> [24/4/14]

Use value

Has an impact on research, teaching and studying efficiency

“We found that the total estimated efficiency impacts of ADS among its user community might be as much as £58 million per annum at 2011 activity levels”

Beagrie, Neil and John Houghton. 2014 *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/> [24/4/14]

The cost of not archiving

Loss of research investment in primary records

Eventual backlash from taxpayers who see loss of research data as wasteful of government funds

Loss of heritage as records are lost and become unusable

Time lost in finding research material if it still exists

Time lost in recreating research data if it has been lost



Australian Research Council – LIEF grants LE110100142,
LE0560711, LE0453247
ARC DP0450342, DP0984419

I acknowledge the support of the Alexander von Humboldt
Foundation.

<http://www.humboldt-foundation.de>

