UC3
UC Curation Center

Total Cost of Preservation (TCP)
Cost Modeling for Sustainable Services

Rev. 2.0 – 2012-10-12

1 Introduction

Information technology and resources are thoroughly integrated with, and indispensable to, today's Internet and web-based culture, commerce, science, education, and entertainment. The digital assets underpinning these activities, however, are inherently fragile with respect to ever increasing disruptive technological change. Without effective, affordable, and proactive curation management, today's digital assets will not remain viable and useful in the future. In order to achieve ubiquity of long-term preservation efforts and results, the full economic costs of preservation activities over time, the "*total cost of preservation*" (TCP), must be well understood.

TCP analysis can be applied usefully in the development of two specific price models for preservation service pricing:

- Pay-as-you-go
- Paid-up

The pay-as-you-go model is appropriate for situations where a reliable and predictable annual income stream is available to the client purchasing preservation services. When this is not the case, for example, with organizations facing irregular or boom-or-bust budgetary cycles or for fixed term, grant-funded research projects, the paid-up model may be more attractive; indeed, in many circumstances it may be the only realistic option.

2 Cost modeling

Long-term digital preservation is a complex activity, involving sophisticated technical infrastructure as well as significant human competencies, analysis, and decision making. Given the difficulties in accounting for the myriad aspects of preservation activity, and of projecting the cost of those activities into the indefinite future, any analysis of TCP must rely on a number of fundamental abstractions and assumptions about those activities and the environment in which they take place. (See Appendix A for a summary of prior work on preservation cost analysis.)

CC BY

2.1 Preservation landscape

The TCP analytical framework abstracts preservation activities and their environment into the following 10 high-level categories, each of which is a cost component in TCP modeling:

1. Content ***Owners*** or ***Curators***, with a primary stewardship responsibility for digital content ***Collections***, make use of …
2. Submission ***Streams*** to send content to a …
3. Preservation ***System***, composed of ***Ingest***, ***Data management***, and ***Access*** service functions, all running on …
4. ***Servers*** and making use of …
5. ***Storage***, to enable discovery and exploitation by …
6. Content ***Consumers***, with ongoing …
7. Preservation ***Analysis and planning*** and …
8. ***Interventions*** as necessary to ensure ongoing viability and accessibility, all subject to …
9. Direct operational ***Administration*** and …
10. High-level ***Management*** oversight.

This abstraction (see Figure 1) is based on the ISO 14721 Open Archival Information System (OAIS) reference mode, which is widely used within the digital library and preservation communities. However, the model and some of its terminology have been modified slightly in order broaden its applicability and facilitate understanding by a non-specialist audience. (More information about OAIS is found in Appendix F.)

A content *Owner* or *Curator* is an individual or corporate agent with primary *intellectual* stewardship responsibility for that body of content; the *technical* stewardship responsibility rests with the preservation *System*.

A submission *Stream* is the technical mechanism, and its accompanying workflow, used by *Owners* to transfer digital content to the preservation *System*

A preservation *Intervention* is a non-routine, non-periodic, or unpredictable activity, such as a format migration, necessary to ensure ongoing access to content managed within the preservation *System*.

CC BY

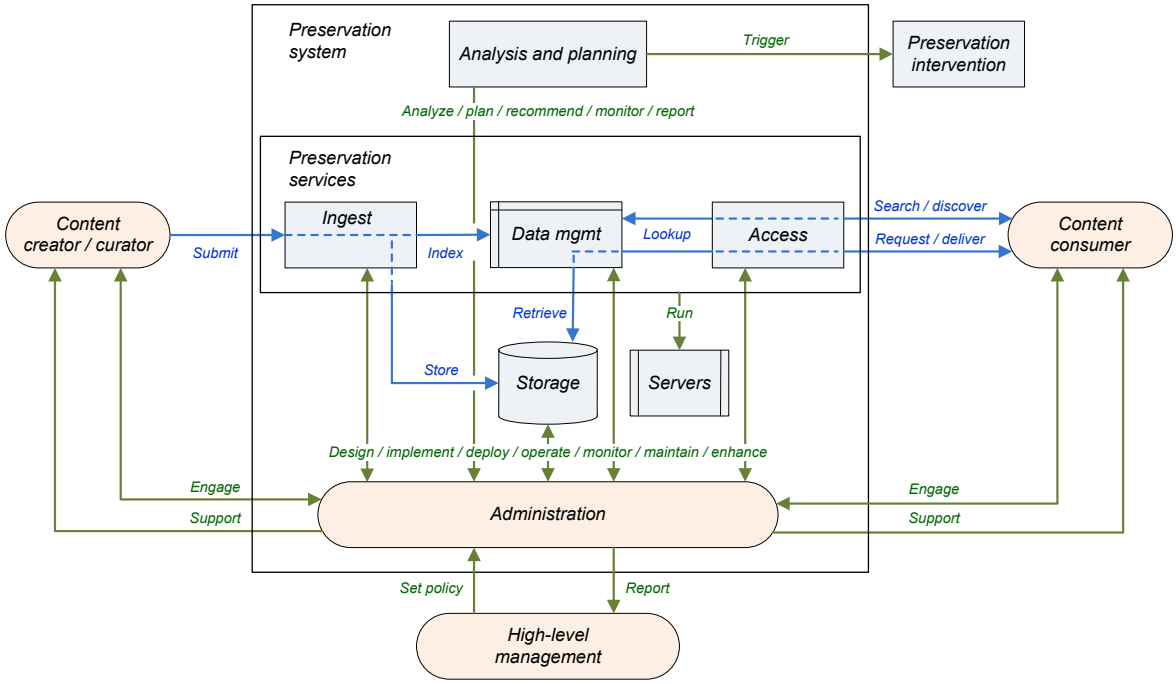


Figure 1 – TCP entities and interactions

2.2 Assumptions

Properly accounting for the full economic costs of long-term preservation is a complicated undertaking. In order to make the analysis tractable, the TCP analysis makes the following simplifying assumptions:

1. Only the costs pertaining to preservation service *providers* are considered. In particular, the costs associated with the local activities of content *Owner* or *Curators* (e.g., content creation or acquisition, reformatting, packaging, submission, etc.) are considered out of scope. On the other hand, the cost of supporting *Owners* in making use of preservation *Service* functions *is* in scope.
2. Individual cost components of the model can be categorized unambiguously as either *fixed*, which are incurred regardless of level of use, or *marginal*, which scale proportionally with use.
3. Individual cost components can be categorized unambiguously as *one-time*, *term*, or *annual*. Term costs are those incurred for an activity or capacity with an effective lifespan of more than one year. Term costs can be *annualized* on an inflation-adjusted basis over their effective lifespan. (See Appendix E.)
4. The values defined for individual cost components represent *nominal* costs, that is, the costs defined for generic instances of activities or system

capacities. This is reasonable under a further assumption of a policy of uniformity of preservation *effort* rather than *outcome*. For example, on the basis of its form, structure, accompanying metadata, etc., some digital content may be more inherently amenable to preservation care and will naturally receive a higher level of preservation service and outcome.

5. Most routine preservation actions performed on content, e.g., characterization, fixity, normalization, etc., are substantially automated or automatable. Thus, the main costs associated with an action are in the acquisition and deployment of the implementing software, which is *independent* of the number of objects against which the action is performed.
6. The preservation service provider can carry forward budgetary *surpluses* across fiscal year boundaries. Furthermore, these surplus funds can be *invested* at market rates, with the income further contributing to the surplus.
7. The model is ultimately *revenue neutral*, that is, there should be no surplus (or deficit) funds remaining at the end of the period under consideration.
8. Values can be determined for various *annual adjustment factors*, such as inflation, investment rate of return, cost of living adjustments (COLA), merit pay raises, and changes to the per-unit cost of preservation, which, although held *constant* over the full period of TCP consideration, are nevertheless reliably *predictive* of long-term economic, technological, and organizational trends.

The last assumption, although the mainstay of the standard economic forecasting technique of *discounted cash flow* (DCF) analysis, can be problematic over extended time periods.

2.2 Total cost of preservation

The total cost of preservation (TCP) to a preservation service provider across all of its content *Owners* is represented numerically as:

(1)

where:

	Total cost of preservation (incurred by the service provider across all *Owners*).
	Number of content *Owners* or *Curators*.

	Unit cost of a *Owner* or *Curator*.
	Number of unique submission *Streams*.
	Unit cost of a *Stream*.
	Fixed cost of the preservation *System*.
	Number of *Servers*.
	Unit cost of a *Server*.
	Number of units of *Storage*.
	Unit cost of *Storage*.
	Number of content *Consumers*.
	Unit cost of a *Consumer*.
	Fixed cost of *Analysis and Planning*.
	Number of preservation *Interventions*.
	Unit cost of an *Intervention*.
	Fixed cost of *Administration*.
	Fixed cost of *Management*.

The preservation *System*, *Administration*, and *Management* are considered *fixed* costs, as they are incurred regardless of level of adoption. Thus they are represented in Eq. (1) by single terms, , , and , respectively. All other components are *marginal* costs, which scale with use, each represented by two terms: a unit cost and the number of units consumed or allocated.

2.3 Pay-as-you-go price model

The pay-as-you-go cost is based on an annual billing cycle, The per-unit *Owner*, *Streams*, and actual *Storage* cost components are specific to a given *Owner*. All other cost components – submission *Steams*, preservation *System*, *Servers*, content *Consumers*, *Analysis and planning*, *Interventions*, *Administration*, and *Management* – are considered "common good" costs; that is, they are generally applicable and beneficial to all content *Owners* equally, and as such, are properly apportioned across all *Owners*. In Eq. (2) the terms for these components are collected together and divided by , the number of *Owners*.

(2)

where:

	Pay-as-you-go cost of a given *Owner.*
	Number of submission *Streams* attributable to the *Owner*.
	Number of units of *Storage* attributable to the *Owner*.

2.3.1 Cost categories

It may useful to distinguish subsets of costs on the basis of the putative source of funding for the underlying activity or capacity. This requires splitting some of the terms in Eqs. (1) and (2) into constituent parts reflecting meaningful administrative distinctions. For example, the following breakdown of costs may reflect the administrative reality for many preservation service providers.

1. Common good.
    a. *Development*. Annualized costs of major technical development of preservation System and its subsidiary Services, generally funded through external grants.
    b. *Operation*. Annual costs of operating the *System*, supporting content *Consumers*, and preservation *Administration* and *Management,* generally funded by central University and campus funds.
    c. *Infrastructure*. Annual costs of operating the *Servers* underlying the *System*, generally funded by central University and campus funds.
    d. *Intervention*. One-time costs associated with a preservation intervention necessary to ensure ongoing viability of and accessibility to managed content. Due to the irregular necessity for preservation interventions, their cost may be best covered through special subventions.
2. *Owner*-specific costs.
    a. *Customer Relations Management (CRM)*. One-time costs of accepting and registering for use a new *Owner* and the submission *Streams* used by that *Owner*, generally funded by direct billing to the *Owner*.
    b. *Dedicated resources*. Annual or annualized costs of resources solely dedicated for use by an *Owner*, for example, *Storage*, generally funded by direct billing to that *Owner.*

The modified formulas for these six cost categories are derived in Appendix B.

CC BY

2.4 Paid-up price model

The paid-up price model is based on a one-time, up-front payment, , sufficiently large to pay for all subsequent preservation activities over a term of years, assuming an annual investment rate of return, , or surplus funds and a compounding discount (or markup) factor, , applicable to the cost of the activities.

(8)

where:

	Paid-up price.
	Pay-as-you-go price.
	Investment rate of return, as a decimal (2% = 0.02).
	Discount (or markup) factor, as a decimal.
	Term, in years.

See Appendix C for the derivation of Eq. (8).

The determination of the discount factor, , is highly dependent on the nature of the underlying cost component. it is useful to assign a different factor to each of the cost categories defined in Section § 2.3.1, based on the following assumptions:

1. The cost of staff-dependent components, for example, *Owner* and *Consumer* CRM or *Analysis and planning*, tends to *increase* over time on the basis of cost-of-living and merit salary increases, but this increase may be mitigated by increases in staff *productivity* resulting from improved tools and techniques.
2. The cost of technology-dependent components, for example, *Servers* and *Storage*, tends to *decrease* over time, at least is has historically through steady increases in semiconductor and storage media density. These two conditions are known popularly as Moore's law and Kryder's law.

CC BY

Appendix A Prior work

A number of international efforts have studied the question of long-term preservation costs; most notably, the Nationaal Archief of the Netherlands in 2005; the LIFE[2] project (Life Cycle Information for E-Literature) work on a Generic Preservation Model (GPM) in 2008; the KRDS project (Keeping Research Data Safe) in 2010; the Princeton DataSpace initiative in 2010; and the Danish National Archives and Royal Library CMDP (Cost Model for Digital Preservation) in 2011 [. (See [] for a summary of these and other relevant activities.)

All of these models analyze preservation costs throughout the full lifecycle of preserved assets. Both the Nationaal Archief and LIFE project work employ very fine-grained analysis of cost components and are based on representative *actual* costs rather than the *nominal* costs employed by TCP. The LIFE model includes cost components for content creation and acquisition, which are considered out of scope in the TCP analysis. Furthermore, LIFE scales preservation action costs by the number of expected objects. TCP assumes that there is no per-object marginal cost; instead, marginal costs are associated only with the various *Content Types* of which objects are members. The KRDS model is specific to the research data lifecycle, but its findings appear to be applicable to other contexts. Unlike the Nationaal Archief and LIFE models, KRDS assumes a discounting function that annually decreases the aggregate cost of preservation service. The DataSpace model is also based on a discounting function, but its analysis covers only the costs associated with preservation storage, which are defined on a pay once, store forever (POSF) basis. The OAIS activity-based CMDP work concentrates on post facto measuring of preservation costs, rather than on forecasting, although it is possible using the framework. While DataSpace is explicitly concerned with “forever” pricing, neither it nor any of the other models assume the benefit of an annual investment return in offsetting a portion of ongoing costs. Most, if not all, of the individual cost components of these models can be mapped to OAIS environmental and functional entities or TRAC criteria, facilitating common points of reference and comparison.

A number of institutions, both commercial and non-profit/academic, now offer long-term preservation services. (Or at least persistent storage; the important distinction between the two is not always apparent from the description of these service offerings.) Carbonite is representative of commercial preservation service offerings. Its cost is $599/year for up to 500 GB, with further increments of 100 GB available for $89/year. The Princeton DataSpace repository charges a one-time fee of $6.00/GB. The DuraCloud service has a variety of price points ranging from $1,500/TB/year to $6,900/TB/year for basic through enterprise service. Note that these are the baseline prices for the first TB, with a lesser

rate for subsequent use and further price decreases for content over 10 TB in size. The LifeTime Library at the University of North Carolina offers students permanent storage "and associated services" for 250 GB, apparently with no associated fees. The USC Digital Repository offers a paid up license for 20 years of preservation service for $1000/TB. In general, however, little information is available explaining the basis of their business models.

CC BY

Appendix B Pay-as-you-go cost categories

It may useful to distinguish subsets of costs on the basis of the putative source of funding for the underlying activity or capacity. This requires splitting some of the terms in Eqs. (1) and (2) into constituent parts reflecting meaningful administrative distinctions. For example, the following breakdown of costs may reflect the administrative reality for many preservation service providers.

1. Common good.
    a. *Development*. Annualized costs of major technical development of preservation System and its subsidiary Services, generally funded through external grants.
    b. *Operation*. Annual costs of operating the *System*, supporting content *Consumers*, and preservation *Administration* and *Management,* generally funded by central University and campus funds and subventions.
    c. *Infrastructure*. Annual costs of operating the *Servers* underlying the *System*, generally funded by central University and campus funds and subventions.
    d. *Intervention*. One-time costs associated with a preservation intervention necessary to ensure ongoing viability of and accessibility to managed content, generally funded by ???. Special assessment or subvention???
2. *Owner*-specific costs.
    a. *Customer Relations Management (CRM)*. One-time costs of accepting and registering for use a new *Owner*, generally funded by that *Owner*.
    b. *Dedicated resources*. Annual or annualized costs of resources solely dedicated for use by an *Owner*, for example, *Storage*, generally funded by that *Owner.*

(B.1)

(B.2)

(B.3)

(B.4)

(B.5)

(B.6)

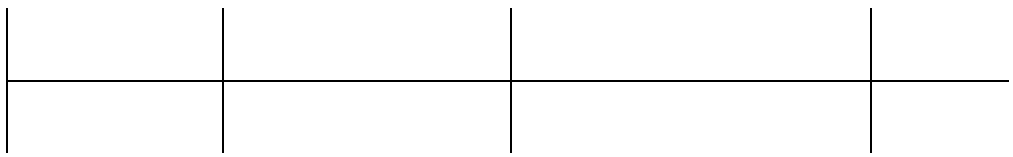
where:

	Pay-as-you-go common good development cost.
	Fixed cost of *System* development.
	Pay-as-you-go common good operational cost.
	Fixed cost of *System* operational support.
	Pay-as-you-go common good infrastructure cost.
	Pay-as-you-go common good *Intervention* cost.
	Pay-as-you-go *Producer* CRM cost.
	Pay-as-you-go Producer dedicated resource cost.

Appendix C Paid-up price model derivation

The paid-up price model is based on a one-time, up-front payment, , sufficiently large to pay for all subsequent preservation activities over a term of years, assuming an annual investment rate of return, , or surplus funds and a compounding discount (or markup) factor, , applicable to the cost of the activities. This can be illustrated by the cash flow diagram (CFD) in Figure 4.

Income



Expense

Figure C.1 – TCP paid-up price cash flow diagram

At beginning of the first year ( ) there is the one-time payment and expense . At year end there is the additional income generated by investing the surplus funds, , with a return rate of , and the following years expenses, which is the initial year cost adjusted by the compounding discount factor.

Since the paid-up price model is intended to be ultimately revenue neutral, at the end of the term the net financial position, equal to the previous year's net plus the current year's income minus expenses, should be 0. Thus, at the end of year :

(C.1)

Solving Eq. (C.1) for gives the one-time price.

(C.2)

(C.3)

(C.4)

(C.5)

(C.6)

(C.7)

An equation of the form of Eq. (C.7) is known as a *power series* and has the general solution derived in Appendix D.

(C.8)

(C.9)

Appendix D Power series

An equation of the form:

(D.1)

which can be represented in compact form as:

(D.2)

is known as a *power series*. Its general solution is derived as follows: Eq. can be scaled by a factor of :

(D.3)

Subtracting Eq. (D.3) from (D.1) yields:

(D.4)

Most of the terms cancel, leaving:

(D.5)

which can be rearranged to form the general power series solution:

(D.6)

(D.7)

for (D.8)

An *infinite* power series, one for which can be solved by observing that as so long as satisfies the convergence criterion :

for (D.9)

Various other equations in the TCP model are derived by substituting appropriate terms for and into Eqs. (D.8) and (D.9).

Appendix E Annualizing term costs

Many preservation costs are incurred for an activity or capacity with an effective lifespan of more than one year, for example, preservation analysis or servers or storage. In the TCP model these costs are *annualized* on an inflation-adjusted basis over their effective lifespan. For example, suppose a preservation service provider procures a server for $5,000 with the intention of replacing it after five years. The *nominal* annualized cost is $5,000/5 years = $1,000/year. But this does not take into account the inflationary effect on the annualized cost, since future dollars are worth less than current dollars. The *net present value*, , of series of nominally-equal annual cash flow amounts, , subject to an inflation rate, , is:

(E.1)

Eq. (E.1) is a power series that can be solved by substituting c for , for , and for in Eq. (D.8) of Appendix D.

(E.2)

which can be simplified to:

(E.3)

(E.4)

(E.5)

In the TCP case the NPV amount is known, so Eq. (D.5) can be solved for the unknown adjusted annualized cost .

(E.6)

In the example above, the adjusted annualized value is $1,039.99, assuming 2% inflation over 5 years.

Appendix F Open Archival Information System (OAIS)

The ISO 14721 Open Archival Information System (OAIS) reference model provides a useful reference for the development of a comprehensive archives "consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" [].

The simplified OAIS entities and interactions used by TCP are depicted in Figures F.1 and F.2.

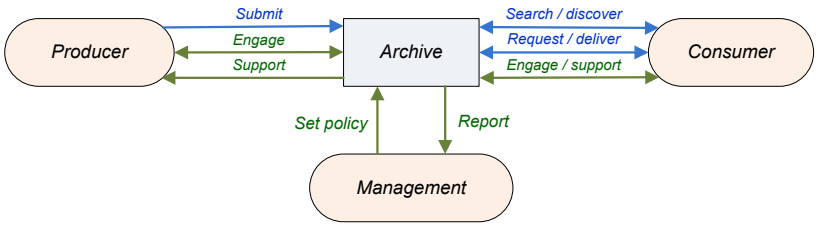


Figure F.1 – High-level OAIS entities and interactions

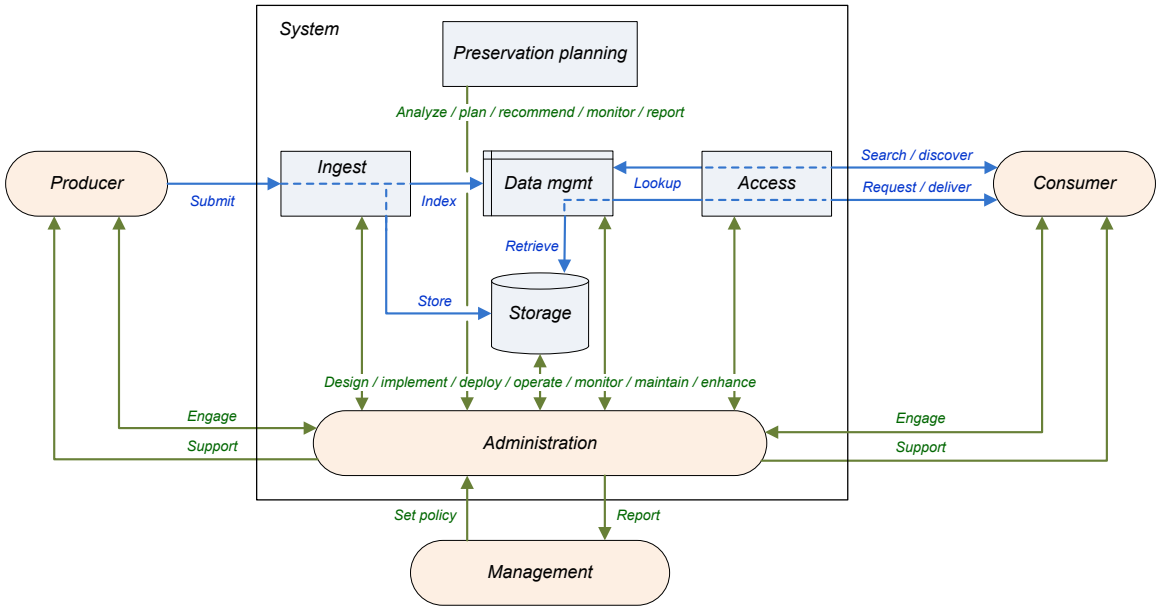


Figure F.2 – Intermediate OAIS entities and interactions

For purposes of accounting for and allocating the full economic cost of long-term preservation, the TCP model (see Figure 1) distinguishes between a narrowed *System* entity, composed of the *Ingest, Data management*, and *Access* services, and the *Servers* or infrastructural platform on which the *System* operates and the *Storage* that the preserved content occupies.

References

[1] Ayris, Paul, Richard Davies, Rory McLeod, Rui Miao, Helen Shenton, and Paul Wheatley (2008), *LIFE2 Final Project Report*, <http://eprints.ucl.ac.uk/11758>.

[2] Beagrie, Neil, Brian Lavoie, and Matthew Woollard (2010), *Keeping Research Data Safe 2* <http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>.

[3] Charles Beagrie (2011), *Keeping Research Data Safe (KRDS)* <http://www.beagrie.com/krds.php>.

[4] Danish National Archives/Danish Royal Library (2011), *Cost Model for Digital Preservation (CMDP)* <http://www.costmodelfordigitalpreservation.dk/>, accessed 2011-09-02. See also, <http://www.costmodelfordigitalpreservation.dk/contact/cmdp-1---preservation-planning-and-digital-migrations/documentation/Kostenmodel_in_Excel_versie_1.2_%28final%29.xls> and <http://www.costmodelfordigitalpreservation.dk/contact/cmdp-2---ingest-and-archival-storage/documentation/Tool_Version_2.xls>.

[5] DuraSpace (2012), *DuraCloud Subscription Plans* <http://www.duracloud.org/pricing>.

[6] Goldstein, Serge J., and Mark Ratliff (2010), *DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data*, <http://arks.princeton.edu/ark:/88435/dsp01w6634361k>.

[7] Hole, Brian, Li Lin, Patrick McCann, and Paul Wheatley (2010), "LIFE^3: A predictive costing tool for digital collections," *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, September 19-24, 2010 <http://www.life.ac.uk/3/docs/Ipres2010_life3_submitted.pdf>.

[8] Intel Corporation (2005), *A Conversation with Gordon Moore: Moore's Law* <ftp://download.intel.com/museum/Moores_Law/Video-Transcripts/Excepts_A_Conversation_with_Gordon_Moore.pdf>.

[9] ISO 14721 (2003), *Space data and information transfer systems – Open archival information system – Reference model*. See also, CCSDS 650.0-B-1 (2002), Blue Book <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

[10] Kejser, Ulla Bøgvad, Anders Bo Nielsen, and Alex Thirifays (2011), "Cost model for digital preservation: Cost of digital migration," *International Journal of Digital Curation* 6(1): 255-267 <http://www.ijdc.net/index.php/ijdc/article/viewFile/177/246>.

[11] LIFE project (2011), *LIFE: Life Cycle Information for E-Literature* <http://www.life.ac.uk/>.

[12] Nationaal Archief (2005), *Costs of Digital Preservation*, <http://www.nationaalarchief.nl/sites/default/files/docs/kennisbank/codpv1.pdf>, accessed 2011-09-02. See also,

<http://www.nationaalarchief.nl/sites/default/files/docs/kennisbank/kostenmodel_in_excel_versie_1.2_final.xls>.

[13] OCLC/Center for Research Libraries (2007), *Trusted Repository Audit & Certification: Criteria and Checklist* <www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf>.

[14] Rosenthal, David S. H., Daniel C. Rosenthal, Ethan L. Miller, Ian F. Adams, Mark W. Storer, and Erez Zadok (2012), "The economics of long-term digital storage," *Memory of the World in the Digital Age: Digitization and Preservation*, Vancouver, September 26-28, 2012 <http://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf>.

[15] University of North Carolina (2011), *Life Time Library thought to be the first nationwide*, September 15, 2011 <http://sils.unc.edu/news/2011/SILS-lifetime-library>.

[16] University of Southern California (2011), *USC Digital Repository: Pricing* <http://repository.usc.edu/pricing/>.

[17] Walter, Chip (2005), "Kryder's law," *Scientific American* 293 (July 25): 32-33 <http://www.scientificamerican.com/article.cfm?id=kryders-law >.

[18] Zeller, Jean-Daniel (2010), "Cost of digital archiving: Is there a universal model?" *8th European Conference on Digital Archiving*, Geneva, April 28-30, 2010 <http://regarddejanus.files.wordpress.com/2010/05/costsdigitalarchiving-_jdz_eca2010.pdf>.