Bo-Christer Björk
Cloudlake Consulting Oy
Finland
7.7.2007

EVALUATION OF THE COSTING ACTIVITIES AND ECONOMIC MODELS FOR DIGITAL CURATION USING THE LIFE METHODOLOGY

Report commissioned by the LIFE-2 project consortium

Background

The LIFE project (Lifecycle Information For E-Literature) was carried out during 2004-2006 by a consortium consisting of The British Library and University College London Library Services . The project was joint venture funded by JISC under the programme area Institutional Management Support and Collaboration. The project has received favourable feedback, for instance during a workshop organised at the end of it, and JISC has agreed to fund a second phase during 2007-2008. The consortium has been strengthened by three associate partners (SHERPA-LEAD Consortium, SHERPA-DP and the Medical Research Council). In addition some funds were reserved for the use of an outside economic consultant for an evaluation of the life-cycle models that emerged as the key results from the first phase.

The LIFE-2 project consists of five work packages, and this report is part of the first of these. The objective of WP 1 is formulated in the LIFE^2 Project proposal as follows:

Validation of the economic modelling and methodology for the Lifecycle and Generic Preservation formulae developed in Phase 1 of the LIFE project, with technical and presentational development of the models.

Cloudlake Consulting Oy has been commissioned by the consortium to carry out this validation. The report has been written by Bo-Christer Björk. He is professor of Information Systems Science at the Swedish School of Economics and Business Administration in Helsinki, Finland. He has been conducting research concerning the scientific publishing process since 2000 and has published several peer reviewed journal articles as well as conference papers on the subject. He is often an invited speaker at international workshops in this area.

Baseline for this report

Written documents

McLeod, R. and Wheatley, P. and Ayris, P. (2006) *Lifecycle information for e-literature: full report from the LIFE project.* Research report. LIFE Project, London, UK. 122 p. http://eprints.ucl.ac.uk/archive/00001854/

Watson, James (2006) The LIFE project research review, Mapping the landscape, riding a life cycle, 96 p.
http://www.ucl.ac.uk/ls/lifeproject/documentation/review.doc

LIFE^2 Project proposal to the JISC capital programme: Repositories and Preservation, 10 p. Confidential

Meetings with LIFE-2 project participants in London 14-15.5

During this two day meeting the background of the project, the results of the first phase and the aims of the second were extensively discussed. The participants were

Paul Ayres	UCL
Richard Davies	BL
Paul Wheatley	BL
Rory McLeod	BL
Helen Shelton	BL
Bo-Christer Björk	Cloudlake Consulting

Summary of Life phase 1 results

The central objective of the first phase of the project was to find answers to the question:

- What is the long term cost of preserving digital material?

There were also some other objectives, but the above is central to the aim of this report. The method used to tackle this question was an extensive review of previous literature both on methods and empirical data concerning LCA (life-cycle costing). The results are collected in the report by Watson. The bulk of this state-of-the art report was on LCA in library environments although some material pertaining to other fields applying LCA were also included.

The results of the literature review were used to formulate the two equations which are key to the results and also the critical evaluation in this report:

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

Where **L** is the complete lifecycle cost over time 0 to **T**. The detailed descriptions of the meaning of the different terms can be found in the final report.

The model was then used as a basis for cost comparisons in three markedly different case studies. Among other aims one objective was to test if the model is generic, that is applicable to very different types of digital material.

The generic LIFE preservation model is as follows:

The preservation cost for n number of objects of the same file format, over a period of t years beginning at the present time, is:

$$\text{Preservation} = t * TEW + (t/ ULE + PON) * (CRS + UME + PPA + QAA)$$

For more details see the final report.

A number of questions can be asked concerning both formulas, such as:

- Are all relevant costs included
- How should inflation and depreciation be handled
- Are the models applicable over different types of digital material
- Are the detailed terms correctly modelled
- Is it possible to collect empirical data to populate the model or estimate the parameters with a reasonable degree of accuracy.

Context of where the Life-cycle costing models will be used

It is important to devote some time to discussing the circumstances in which the LIFE model is intended to be used, since the context also may influence the structure and applicability domain of the model.

The original intent of the LIFE project seems to have been strategic decision making and perhaps budgeting of preservation activities of digital collections, although this aspect is not very explicit in the report. In the first phase all three case studies dealt with born digital collections, but in the second stage collections which originally have been in paper format will also be included. If born on paper collections are also included in the scope then it would be logical to extend the range of the model to also include the modeling of analogue preservation (where the cost issue currently is better understood and the periodicity of migration activities is longer). And importantly it would be important to cost model the activity of digitisation from analogue to electronic. This is because in some instances the crucial decision for existing collections is to decide whether to convert the collection and to compare the long term costs of keeping the collection on paper only or in digital form only. In such a context it is also important to remember that the costs of providing access to a collection in digital format is likely to be much lower than in paper format (for material archived in a permanent long term storage facility).

In addition to decisions to go on with analogue and/or digital collections for existing material another is (in particular in the national library setting) to be able to choose between different possible new electronic collections such as the Web archiving or voluntary deposit of electronic material. Due to the enormous amount of material today being produced in digital format a national library must also make priority decisions about what material to collect based essentially on two different criteria.

- The importance to society of preserving the material in question
- The life-cycle cost of preservation

In the context of the institutional repositories of universities the situation is usually different. Here choosing what material to preserve is less important since the repositories currently are just in the beginning and contain relatively little material. Thus the default strategy is to try to preserve all the material. This is also made easier if the bulk of the material consists of publications of different sorts (reports, thesis, copies of journal and conference papers).

The suggestion is that in the report from the next phase of the LIFE project this discussion is included and perhaps extended and clarified by concrete examples.

Life-cycle model of preservation activities

In order to get a better understanding of the archiving and preservation processes underlying the above equations an IDEF0 model was constructed. IDEFO is a graphical modelling language, also known as SADT, and it is one of a number of available process modelling protocols for which efficient software is available (NIST 1993). An IDEF0 model consists of a set of activities, depicted by rectangular boxes, which are interrelated and may be arranged in a hierarchical decomposition. An activity needs some inputs and transforms these inputs into outputs by use of machines or people in the organisation. Controls constrain these activities by specifying which conditions are actually regulating the performance of an activity. The purpose of using IDEF0 is to the reveal the meaning of a particular activity and to show the kind of information, material or energy, which is conveyed through the interfaces (i.e. arrows) of activities in the process.

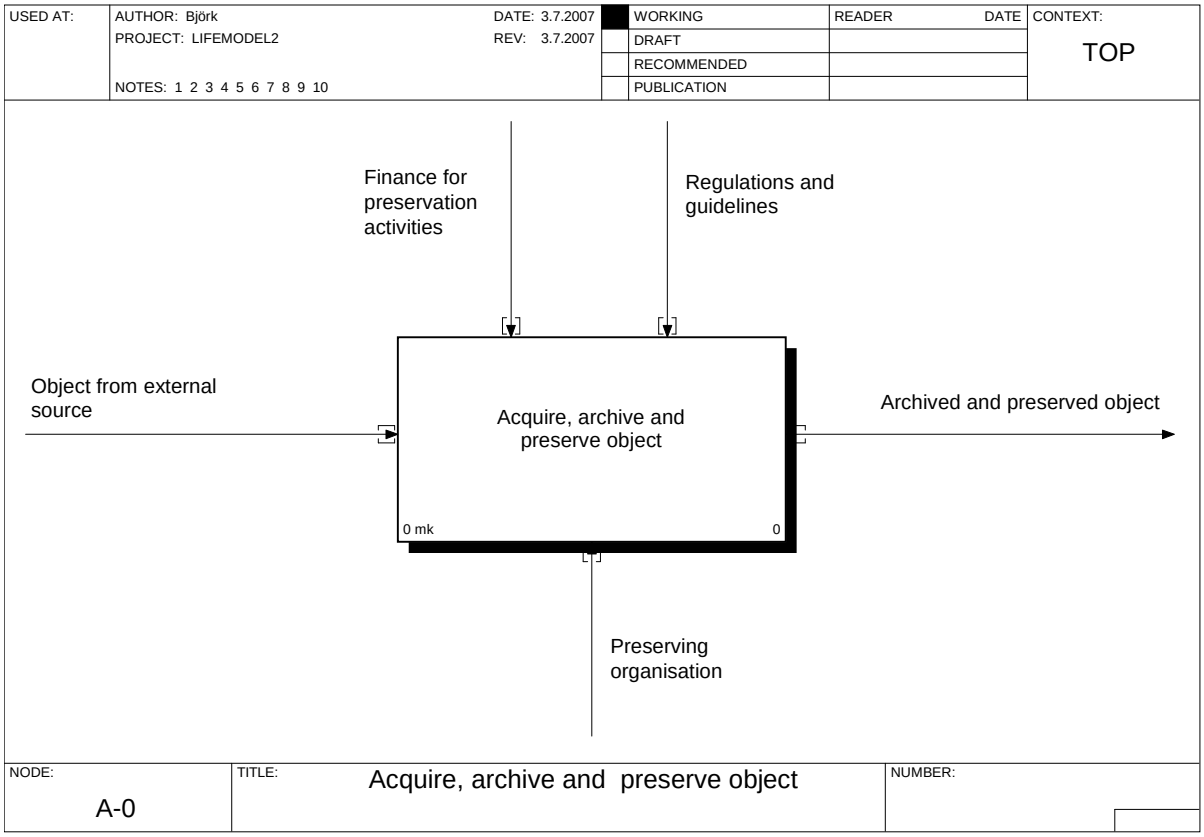
IDEF0 has been used by the author in developing a comprehensive model of the scientific publishing process (Björk 2007). The method has also been used specifically for modelling the preservation of digital objects in the InterPARES project (2005).

The method can be used in different ways. Sometimes modellers clutter the diagrams with so many details that they become very difficult to grasp for people who have not themselves participated in the modelling. It is often better to leave out some details and to make the models as simple and understandable as possible.

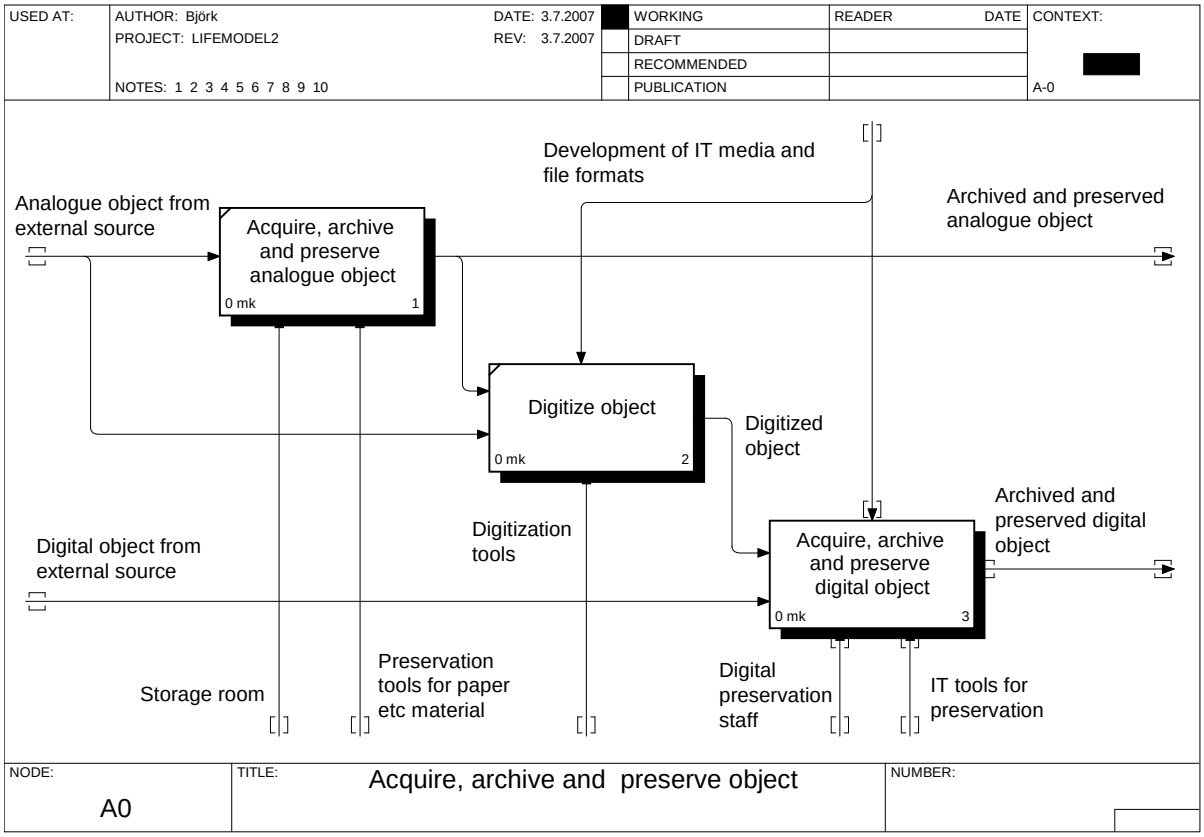
The model covers the same scope as the formulae that resulted from the LIFE phase one project. The different activities are however structured in a slightly different way.A first version of the model was developed for the meeting in London in May 2007 and shown to the LIFE-project participants. It was in particular pointed out that the Reference Model for an Open Archival Information System (OAIS) is the standard reference today and any modelling done within this work should strive to be compliant with the OAIS model (OAIS 2002). This has not however been done in the model presented below.

Below a slightly amended second version is presented. The central change is, based on feedback from Helen Shelton on the interim report, that the preservation of paper materials has been included as well as the digitisation of paper (or other analogue) material. This provides a more comprehensive context for the model.

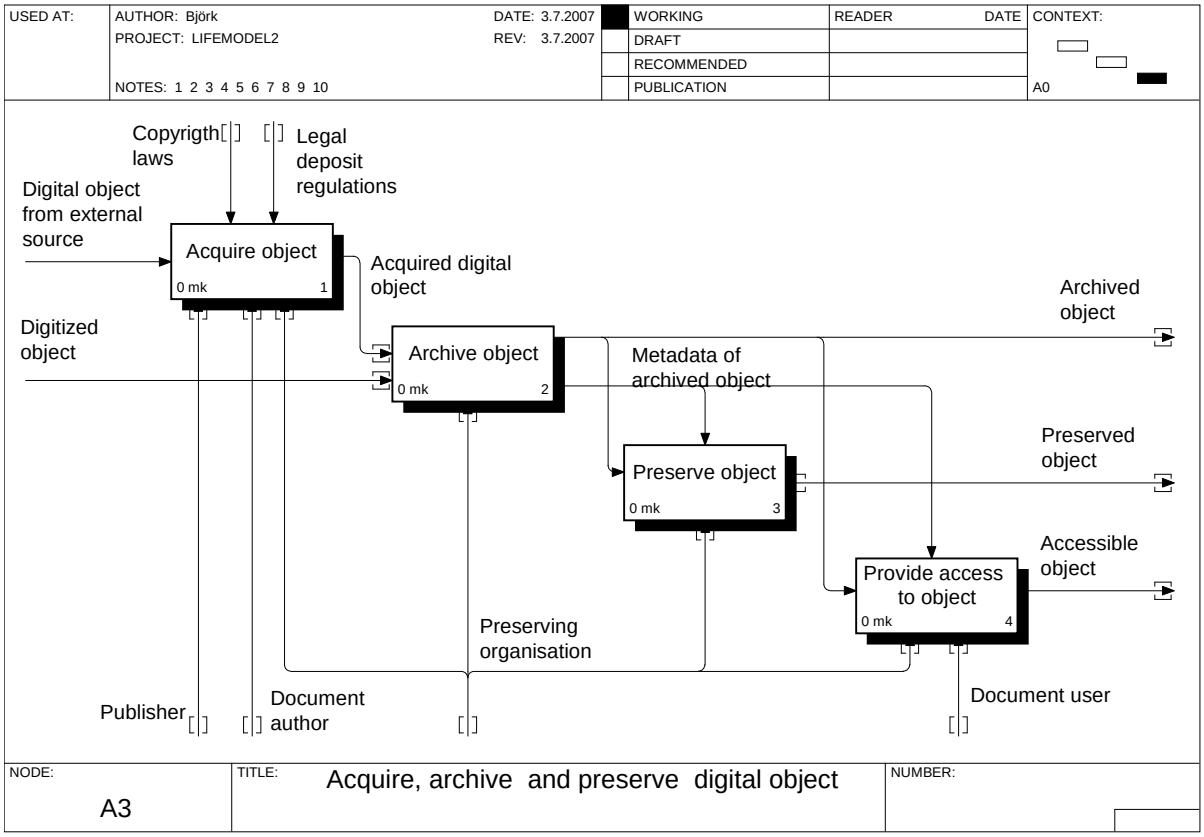
In the following the diagrams from the IDEF0 model are shown. Since the readers of this report are experts in the domain the diagrams are not explained in a more detailed manner.



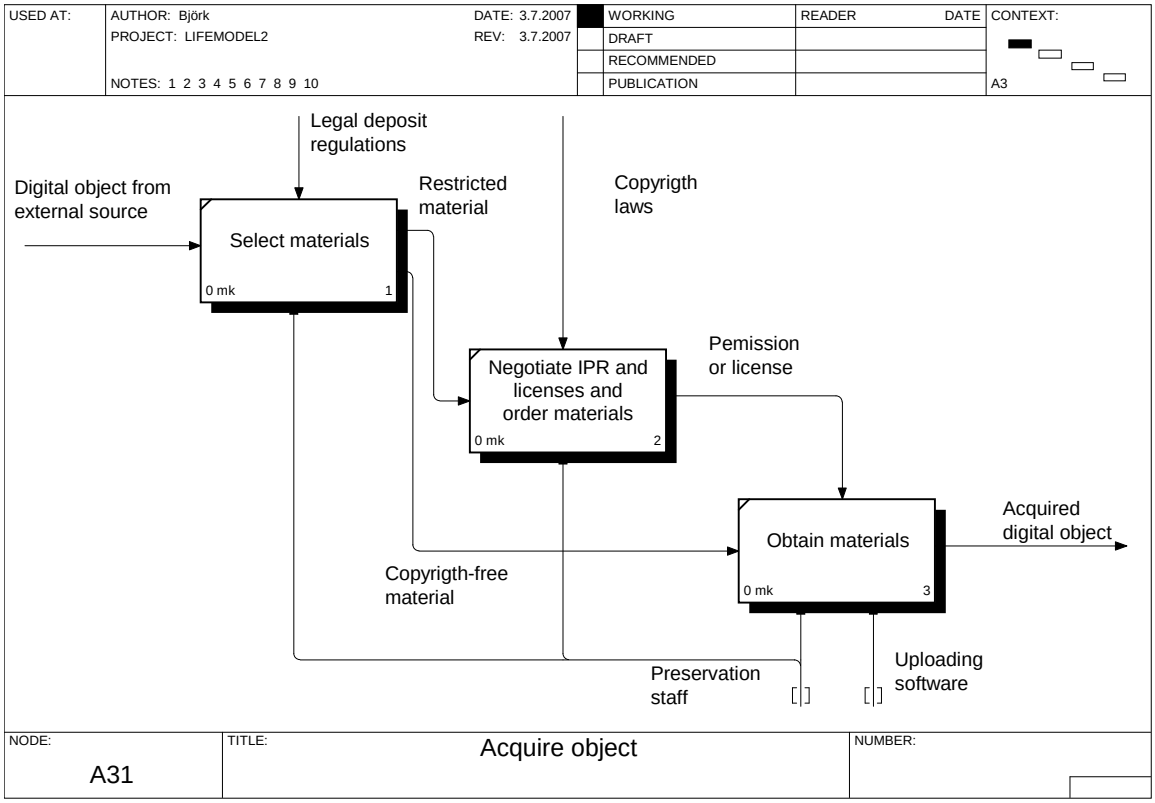
In this figure (A-0) the overall context is shown. An organisation responsible for long time preservation gets a constant influx of objects on different media (whether paper, digital or other) and archives and preserves these. The activity is constrained by the available finance and regulations such as legal deposit.



This figure (A0) is an important addition since the first version presented at the May meeting. It shows the two main streams of preservation activities; preservation of analogue materials (paper, film, photographs, phonograph records etc.) and preservation of digital objects. These two are linked via a digitisation activity. Digitisation can occur directly upon acquisition in which case there might not be any need to archive and store the analogue object, or it can occur at a later stage for an already existing collection. In this model the first two boxes have not been further detailed, but it is clear that in particular the decision to digitise objects (and possible stop preserving the paper equivalents) should be based on long-term cost calculations where the LIFE model comes into the picture.



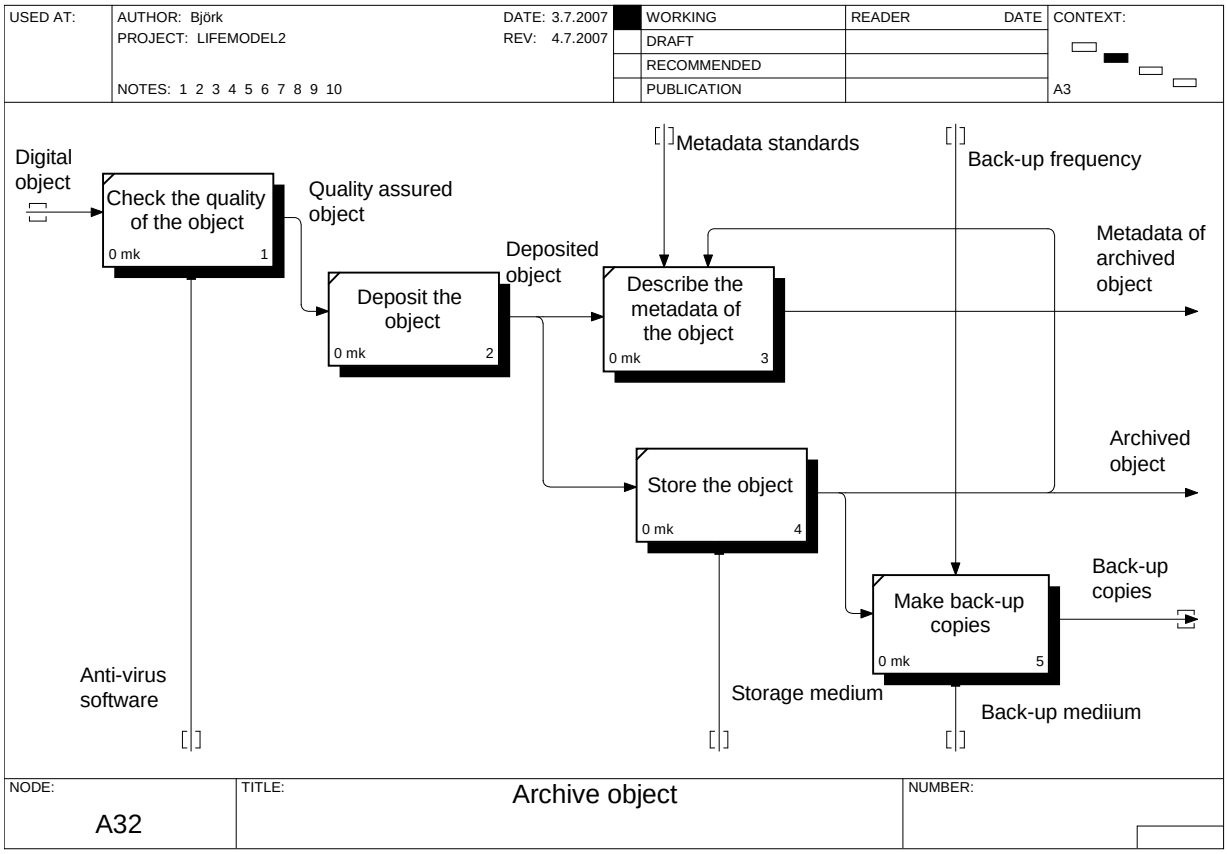
Here (A3) the model has been expanded for the case of digital materials (born-digital or digitized). By preservation is here meant the operations performed on archived objects to update storage media and file formats. The last activity, provide access to object for third parties, is not an absolute necessity but is often performed. For this reason it has been put as the last one in the diagram.



The Acquire object (A31) diagram splits into three subactivities. The first one is concerned with selecting the materials to be preserved. In the second phase of LIFE this is an important activity due to the vast amount of possible choices. In other cases such as the Burnley collection this is not an issue.

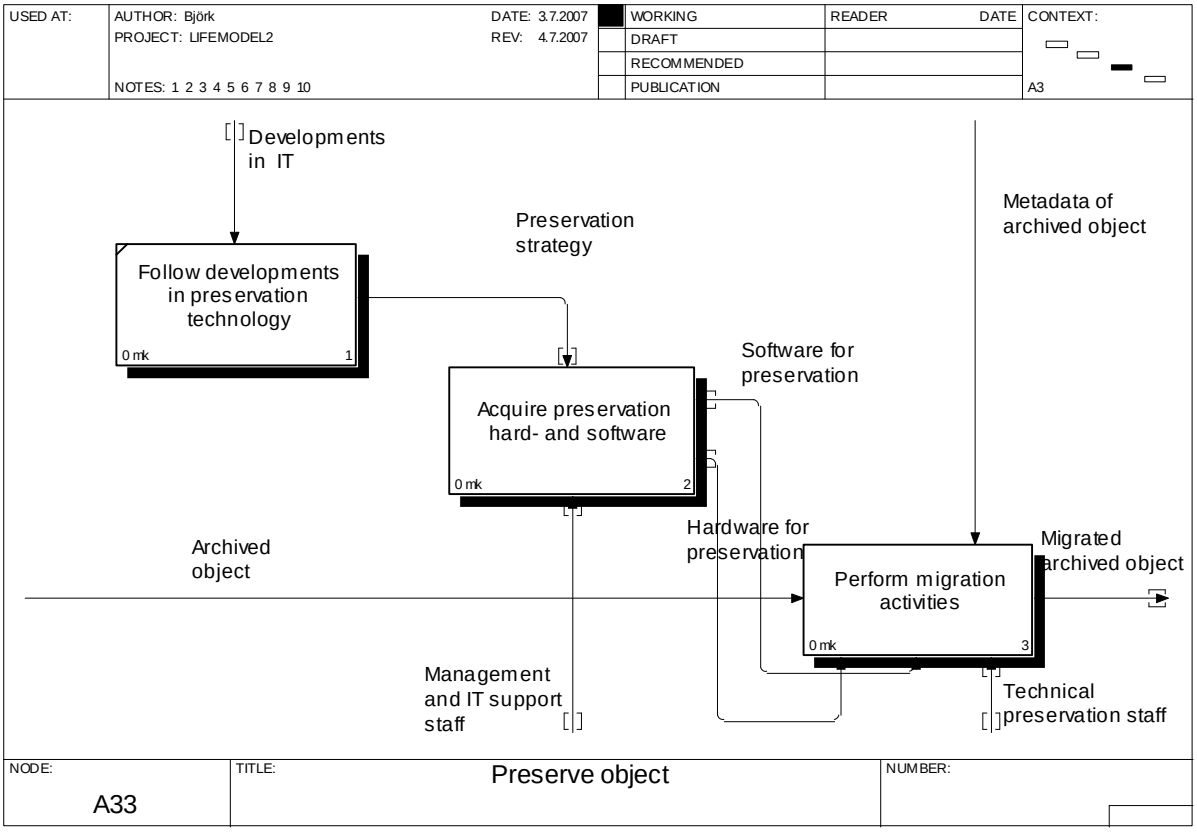
The second stage, which only is of importance for some collections, is negotiating for licences and in general dealing with copyright issues.

The third stage is the obvious one of physically obtaining the materials. In the case of hand-held digital media this can be quite labour-intensive.

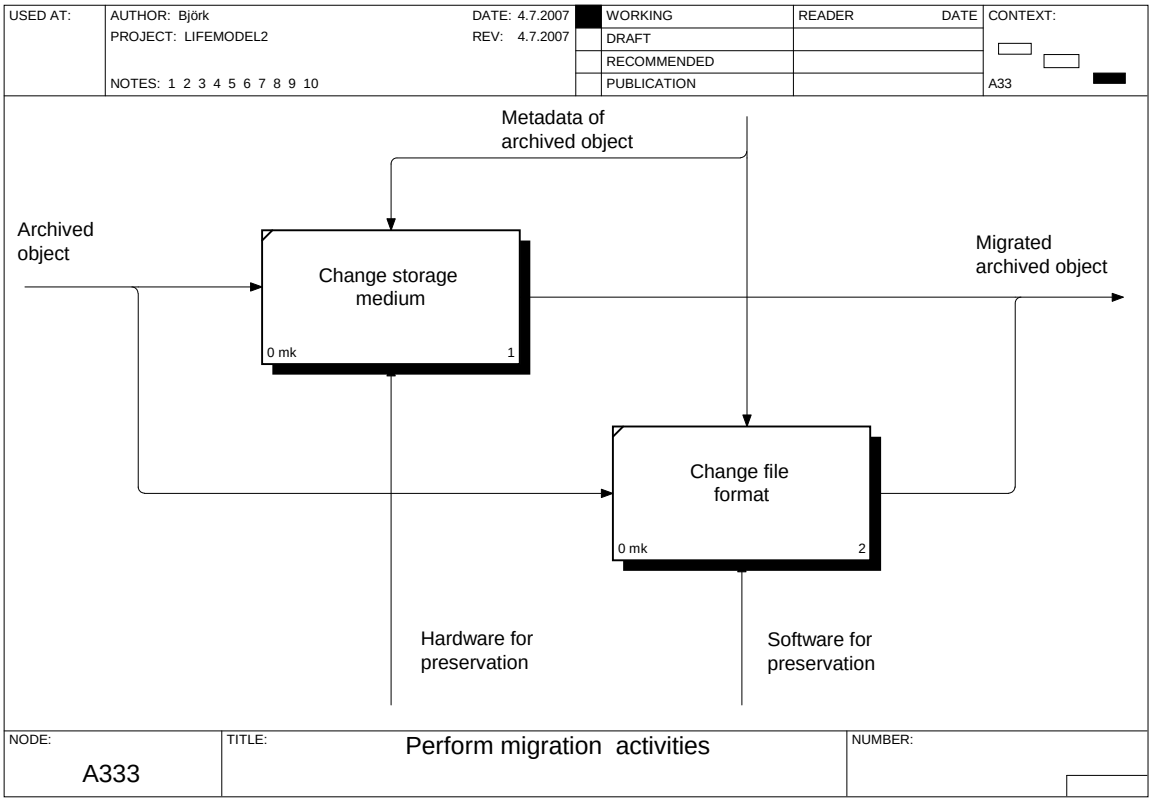


The archive object diagram (A32) contains five activities. It could possibly have been split into two levels by a further decomposition of the “deposit the object” and “describe the metadata of the object activities”. This is the part of the model where this author is most uncertain. There is for instance no activity in the model called “ingest” which again is quite visible in the LIFE model and report.

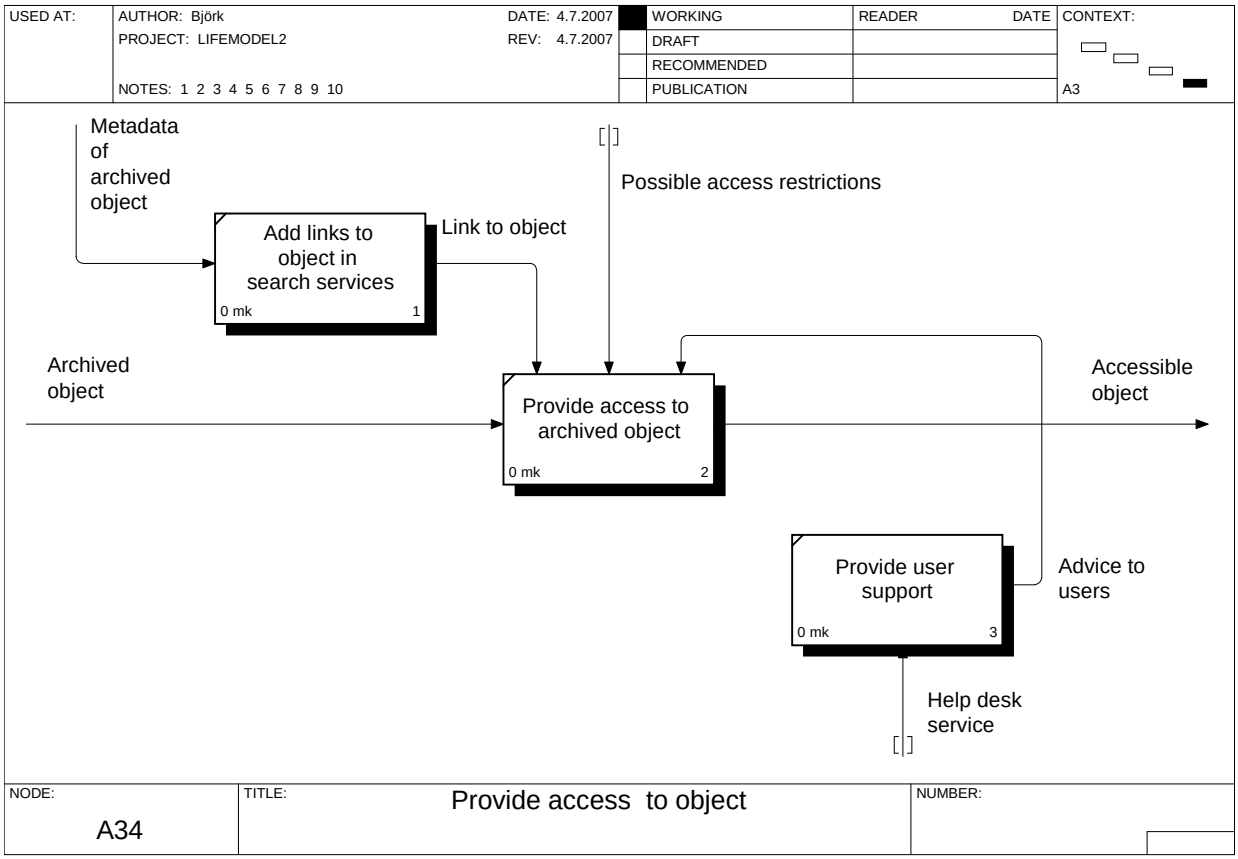
Store the object means here the storage over time on hard-disk, which costs money, rather than the very rapid act of making a copy of a file on the medium. Back-up copying has been added as a separate entity.



The preserve object activity (A33) has been broken down into three separate activities. The first one is a different label (according to the IDEF0 convention of using verbs for activities) for “Technology Watch”. In a cost breakdown according to the model (for instance an Excel sheet based on the activities) the second activity who be the right place to associate software licenses with and the third one the place for man-hours spent on conversions.



Perform migration activities has here (A333) been further broken down into two subactivities, reflecting migrations in storage medium or file format. This could equally well have been done at the higher level diagram A33. Migration to new physical premises was also considered but in the end not included. In a similar model for analogue preservation the physical storage facilities would be an important cost item.



This last diagram (A34) of the model closely follows the breakdown of the LIFE model and report. One aspect which was not discussed here is that also this part would need to be regularly updated as new web technologies etc. emerge.

This model could, if the project group so wishes, be further developed to be one-to-one with the cost equations and presented as part of the report of LIFE 2. It could also be made more complient with the OAIS 2002 model.

Applicability of LIFE model for different digital object categories

The aim of the Life project has been to develop a "generic" life-cycle costing model, which can be applied to many different types of digital objects. In phase 1 the model was tested on three different cases:

- Voluntarily deposited electronic publications at the BL
- Selective archiving of web sites by BL
- E-journals management by UCL

In the case concerning voluntarily deposited electronic publications at the British Library the costs for acquisition are quite low. One significant cost factor is the variety of formats. Although the material is comprised of a large variety of file formats, the three most common, txt, html and PDF account for 86 % of all documents. The second case is more uniform in structure. The third case is markedly different in that the costs are mainly concentrated to acquisition. The issue of long-term preservation of e-journals that a university subscribes to is even internationally unsolved. In the print world the same principles as in the digital LOCKS proposal (Lots of copies in different places keep the documents safe) have applied since most universities who have subscribed have archived the paper issue they have received. In the digital world this is not the case since a license is typically just a "passport" to access to material kept by the publisher. The primary responsibility now seems to lie with the publishers themselves or the national libraries. For instance the Dutch National library has taken on such responsibilities, but can only archive and preserve, not offer access to the materials outside their own premises.

In the second stage of the LIFE project additional case studies will be performed:

- e-print repositories run by the SHERPA-LEAD and the SHERPA-DP consortium
- Empirical data to be used in research in the form of medical records from a period of 60 years (MRC)
- Newspapers from the 17th and 18th century, in the form of digital surrogates created from microfilm

Some tentative conclusions concerning the applicability are:

It would be advisable to limit the models applicability to such material for which the organisation assumes the primary responsibility for long-term preservation and access provision. Thus for instance institutional repositories are very much in focus. Also specialised and internally rather uniform data and document collections such as the Burnley newspapers and the MRC records are very important. The case of university libraries procuring access to e-journals from publishers, where the materials are actually preserved by the publishers, is so special that it should be excluded.

This issue was also discussed during the meeting and it was acknowledged that journal collections, that a university subscribes to is a very special case, and should possibly be excluded from the application range of the model.

Another case of some importance consists of the teaching objects that are increasingly included in institutional repositories. There are several reasons against them being actively included in long-term preservation efforts:

Teaching objects are often complex objects consisting of primary material (i.e. powerpoints), which includes references to other material (for instance journal articles). Of these the material linked to is often the material more in need of long-term preservation, but where the preservation responsibility lies elsewhere. Authors tend to revise teaching material almost yearly and such material becomes obsolete rather quickly. Also many authors are quite "jealous" about their teaching material and want to keep it within restricted access (course participants only).

In rare instances there could be reasons to preserve also teaching material, but those could be handled as special cases.

All in all there seem to be two major application areas for the LIFE models:

- Institutional repositories, which span a range of object types that are likely to populate the IR of a particular university.

- Specialised collections of national libraries and similar organisations, which have a national and sometimes legal obligation to long-term archiving.

In the latter case it seems more sensible to apply the model to individual collections than to the totality of objects stored in say a national library. In such a case it is important for a national library, which works within budget restrictions, to be able to compare the long term preservation costs of different collections, in order to make informed priority decisions. This is in contrast with the Institutional Repository Case.

An important point which could have far-reaching consequences for the parameters of the model is how institutional repositories (which are numerous) are going to solve the preservation management issue. In contrast to national institutions such as the British Library, universities would gain very obvious benefits from sharing resources for preservation, for instance via consortia, outsourcing, using external service providers etc. (A discussion of strategies can be found in Hitchcock et al 2007). A good case is for instance the technology watch function included in the model. One can argue if there is a need for every university to duplicate this effort. A more sensible approach would be for certain service providers to assume the responsibility for issuing guidelines.

Some issues related to costing

The concept of cost

The concept of cost (p. 3) could be elaborated. It is evident from reading the report that the cost modelled is the cost to the organisation curating the collection, for instance a university or a national library. This essentially means that the cost has to be covered via the budget of the organisation.

For many of the collections modelled this is unproblematic. A good example is the Burnley collection for which the British Library is the only organisation assuming responsibility. For other collections or cases this is not the case. The example of the electronic journals used in the UCL case is an example. Here we have a case of an individual subscribing organisation performing digital preservation on a collection where the primary responsibility would be the publisher or/and a national library. Assuming that all subscribing universities in the UK would perform the same preservation function for such journals the total costs, financed through the government budget for higher education, would become considerable. In the long term also the idea of for instance LOOKS seems like very wasteful. The same comment could be made concerning the double preservation of copies of journal papers both by the publisher and the author's institutional repository in the so-called green route to open access.

Comparisons between single items and yearly volumes

One issue which makes comparisons between cases very difficult is the fact that the "atoms" or "molecules" studied are so different in size. For monographs the basic entity is the single monograph, but for serials and journals the basic entity seems to the title and the yearly volume. Documents to be used as data for research, such as old newspapers, medical records or the state of web sites at some point in time offer an even more difficult field. Somehow a common denominator should, however, be found. It would seem that the most sensible common denominator would be the document requiring separate meta-data and being handled as a single file from the viewpoint of the system. Computer storage is becoming so cheap that the size of a document has hardly any significance. Thus

- a monograph
- a copy of a conference presentation or article
- a data set from one research project
- a journal article
- a medical record

should be treated as the basic denominator for calculating and reporting costs. Compared to the earlier report and the cost comparisons this would in particular lead to differences concerning journals.

Handling of Inflation

The aim of life-cycle costing exercises is usually to make comparable costs occurring over very long periods of time. It would be very difficult to compare costs occurring in this year with costs occurring say 20 years from now. Looking at historical inflation rates the face value of costs occurring 20 years from now would be 1,5 to 2 times as high as current costs, although the real resource expenditure would be the same. Equally it would be almost impossible to predict the future inflation rate and even a mistake of 1 % per annum would make a huge cumulated difference. For this reason the standard procedure is to use real prices, that is inflation-adjusted prices, where prices from different years are divided by the general price index for the same year.

The proposal is that the project should deal with inflation by using real, inflation-adjusted prices. In practice this means using the price levels of year one. This does not, however, mean that it should not be allowable to take into account long-term trends in relative prices. Wages for instance tend to rise by perhaps 2 % in real terms whereas digital storage media, if measured in terms of Pounds/Gigabytes tend to become cheaper according to the famous Moore's law (in half per 18 months).

Handling of discounting

Different types of capital investments have very differing profiles in terms of how the expenditure and returns are spread out over the lifetime of the investment. A hydroelectric power plant has for instance a high initial investment but relatively low yearly running costs, whereas an oil-based power plant has high yearly costs due to the price of the fuel used. In comparing the life cycle costs of such investments economists use the technique of discounting. The assumption underlying this is that 1 pound of income (or expenditure) a year from now is worth less than 1 pound today. The reason for this that if we have 1 pound available today we could keep it invested for one year and obtain back 1 pound + the general rate of return on capital investments.

There are different ways for making different cost profiles (over time) comparable. The basic one is the Net present value, which converts all future costs or incomes to a value at year 1.

Thus the present value of a cost or income in year n =

$$PV = \frac{FV}{(1+r)^n}$$

where PV = present value
FV = value in year n
r = discount rate

Thus if the discount rate is 5 % 100 pounds in ten years time has a net present value today of 32 pounds.

There are a number of different methods for estimating the discount rate. In finance a typical method is to look at the cost of long-term finance (cost of capital). Often the yield of government bonds which is risk-free is used as a proxy for the discount rate.

Why is it problematic to apply discounting in this context?

Discounting is particularly well suited to private enterprises where the purpose of investments are to generate returns on investments and where there are usually lot's of alternatives. It can be used to compare capital intensive public investments where the time profiles of alternatives differ a lot, and where the products generate easily measurable revenues (or services which can be evaluated at monetary value). For cases such as universities and national libraries which in this context have an obligation to handle the preservation for perhaps centuries using depreciation would be highly problematic. The central issue is that the preservation and collection management budgets of these organisation tend to be relatively stable over the years (in real terms, inflation adjusted) and that they have no opportunities to invest potential savings today in bonds etc.

Also for most types of digital objects there would seem to be rather high initial costs in year 1 but after that relatively stable low yearly costs, in particular if we spread out the reoccurring changes in format etc evenly. And most importantly the initial costs occur rather evenly spread out over the years, since different parts of the collection are acquired in different years. Thus for instance in the paper world journal subscriptions have tended to be rather stable over the years (disregarding the serials crises) since there has been a constant influx of new issues to be catalogued, circulated and stored. The same applies to electronic materials.

The proposal is not to use discounting

Taxonomy of preservation objects

One aspect which emerged from a close reading of the report of the first stage and the description of the case studies is the need for a clearly defined taxonomy of digital preservation objects. Life-cycle costs differ a lot depending on the types of objects and often the preservation efforts of different institutions range over a number of different types.

In the following a number of possible classifications are listed:

TYPE

- Text
- Numerical data
- Image

FORMAT

- Paper
- microfilm

- digital

ORIGIN (FOR DIGITAL)

- Born digital
- Converted from paper

DIGITAL FILE FORMAT

- PDF
- WORD
- JPEG
- ET.C.

FORMAT LIFECYCLE STAGE

- Original format
- transformed format

ORIGINAL OR COPY

- Original publication
- Copy of original that has been published elsewhere

TYPE OF PUBLICATION

- Scientific Journal article
- Conference paper
- Monograph
- Website
- Popular newspaper

STRUCTURE OF DIGITAL OBJECTS

- Single document
- Multimedia collection (for instance website)
- Database

TYPE OF COLLECTION

- Institutional repository
- Global subject-based repository
- Specialized collection (of national library etc)

The consultant will not attempt to develop this taxonomy but it is suggested that such a taxonomy is developed for the report of LIFE 2.

Use of life-cycle costing in other industries

Two industries which to some extent are using life-cycle costing principles are the construction and facilities management industry and the IT users community (Total cost of ownership). Both are quite relevant for this study. They are mentioned in the literature review report from phase 1 of LIFE but given relatively little attention.

Detailed commenting of the LIFE-1 report.

Comments concerning the VDEP case study

The calculations for this case study highlight quite well some of the issues related to the LIFE model. Around 230000 objects have been deposited since the year 2000. A central issue is to try to understand how the total life-cycle costs related to this collection are spread over the different cost categories. In the following a ten-year scope is used since this is what is reported in the executive summary. The spread, which is particularly interesting, is over the six central categories of the model.

A problem in the calculation is that certain costs for serials occur only once for the whole serial and then separately for each issue. It would be useful to be able to make all average calculation for the individual issue, which should be compared to the individual monograph. In order to make such a comparison the average number of issues per serial over the whole collection should be available.

In looking at the actual figures reported on pages 40-49 it is a bit difficult to get the overall picture since the different items for which the costs are calculated are so different in size.

HAND-HELD MONOGRAPHS

For handheld monographs the three examples are stored on one CD-ROM each, but of quite different sizes (2.2, 17 and 587 MB). Thus the storage costs are proportionately low and the Ingest and Metadata costs dominate for the first two. For the third case the storage costs become the dominant term, quite mechanically. Preservation costs are almost negligible.

HAND-HELD SERIALS

The two examples here are quite different. The Belfast working papers has an extremely low volume (1 issue per every second year!) and has one file of less than 1 MB per year. The OAG Data is a very high volume one with 100 MB per issue and 24 issues per year. Consequently in the Belfast case almost all costs are in the Ingest/Metadata bracket whereas in the OAG the cost is relatively evenly divided between Ingest, Metadata and Storage.

ELECTRONIC MONOGRAPHS

Same pattern as before is repeated here in the form of two cases where one is small (1,6 Mb) and one larger (150 Mb). Consequently in the first the costs are concentrated to I and M, in the second also S gets an important share.

ELECTRONIC SERIALS

Two cases of which the latter has a very high storage requirement (almost 5 GB per year). In the first case storage is only 0,1 %. !!!As can be expected in the latter case storage costs dominate and account for 72 % of total costs of the ten-year period.

SUMMARY

In all the above the share of preservation (P) as a percentage of overall costs is rather low, although there are indirectly Ingest and Metadata costs associated with preservation activities at the 5 year and 10 year milestones. Also it was assumed that Acquisition costs and costs for providing access to the material are nil.

The construction of the averages for the four categories is a bit risky since it gives equal weight to each of the two or three cases used. To take the last example (E-serials) the cost in year 10 of the first case is 3,81 and of the second case 14,18. The mathematical average of these, 8,99, is the overall result reported in the executive summary. It should however be remembered that the first is a serial having four issues per year with storage demands of 0,256 Mb per issues whereas the second has 12 issues per year with 398 MB per issue. If we would calculate over all E-serials in the VDEP collection the question is we should go by title, issue or even size (in digital terms) in calculating the average.

This case also highlights the effect of the time span used for comparisons. In a relatively short time frame of 10 years (in the context of long-term preservation of for instance scientific papers) the initial Ingest and Metadata operations which are manual and individually performed for each item tend to dominate the cost equation. Only as the time frame is prolonged markedly the preservation (P) costs tend to gain in importance.

Summing long-term preservation costs over digital formats for which periodic preservation and migration activities are undertaken is problematic. For instance in the VDEP case there are 22 different formats shown in the table on page 38. Looking at the total cost over 20 years 610000 GPB it's relatively evenly broken down over the formats, reflecting the high overhead cost of technology watch, software etc which are not dependent on the number of entities. A mechanical application of the formulas leads to apparent absurdities such as a cost of over 30000 GBP for one entity alone (the "Li" format). It would seem clear that no organization would consciously budget this sort of sum for the preservation of one file. In fact it can be noted that the three most common formats (Txt, Html, PDF) constitute 85 % of all the files but only 7 % of the total costs over 20 years. The 12 least frequent file formats constitute 0,1 % of all files but contribute with 41 % of total costs.

In practice what is likely to happen is that during the first migration exercise the number of file formats would dramatically reduce to only a handful (one for text, one for images, one for database et), thus making the preservation at points 2, 3 … less costly. Also it is likely that the chosen standard formats would be well supported worldwide with migration tools (considering the number of PDF stored nowadays it would be realistic to assume that standardised migration tools would become available).

In the VDEP case study a format migration interval of 5 years was used (p. 34). This is in conflict with the later presentation of the preservation function where the interval tends to be 8 years or mode. The BLE is assumed to be 8 years and the ULE =BLE + 0.1 * t (pp. 93-93).

Comments concerning the Web Archiving Case Study

This case study deals with material (web sites), which as such is more uniform than in the VDEP case study. Internally the web sites can of course contain quite diverse file formats as documented in the table on page 58.

In contrast to the VDEP study the share of preservation costs (P) is very high in this case (62%). This may have to do with the mechanics of how the LIFE model is applied. It should also be noted that costs related to instances (rather than to the title alone) are completely dominating. The costs per title are estimated at only 21 Pounds where the total costs over the 20 years generated via multiplying over the instances is 13731 Pounds. Thus in order to validate the calculations a closer look should be taken at how Acquisition, Metadata and Preservation costs per instance have been arrived at.

Acquisition per instance (Obtaining Aq4)

This cost has been obtained by taking the current figures for work done during one month and dividing by the number of archived instances (141). The totally dominant term is called Magus enquiries and fixes (2373 Pounds) but it is not explained if this is related to work expenditure or invoicing from Magus (and if so how this is related to the instances). There is a sentence on page 56, which states “In some cases effort is expended by Magus who host the gather software”. Later on page 60 there is some clarification in the form of a table specifying the charges from MAGUS for hosting and support. The figures in the table are a bit unclear. Figures are given for each month but it is unclear what unit they are in. It is also how the totals are related to the average monthly figures. What is clear is that the Magus enquiries and fixes figure on page 56 is equal to the average monthly figure for support in the table on page 60.

Ingest (Quality assurance I1 and Holdings update I3)

Here the figures are relatively clear and based on substantial time spent by staff divided by the number of instances.

Preservation (P)

This as mentioned above is the dominant item in the total cost picture. Within this according to the table on page 61 the items technology watch and preservation tool account for slightly over half (52 %) in the 20-year time frame and even more in the first year.

The central issue here is contained in the sentence “The model is based around the preservation costs for numbers of files of specific formats”. As far as I can understand from the description the numbers of files of different formats where calculated for one months worth of instances and the figures were scaled to represent 1000 instances. Why the scaling was needed is unclear since the average cost per instance is used in any case in the later calculations.

On page 58 we find a description. 38 web instances were analyzed resulting in the table on page 58. This demonstrates that there were over 200000 files with a very uneven spread over 36 file types.

The most common format contributed	88 %
The three top formats contributed	97 %
Half the formats contributed	0,08 %

Assuming that the calculations of preservation costs were performed in the same way as in the VDEP study it is likely that each format contributed almost equally to the total cost perhaps with the exception of the first format (due to the large number of files).

I would put a big question mark about the realism of these figures. It would not seem realistic to assume that each exotic format would continue a separate existence over 20 years being converted to new equally exotic formats which would not coincide with others among the 36 different formats. My guess is that during the first migration one single target format would be chosen for generic information types (such as text, formatted text, image, video, audio, web page.

In looking at web pages for archiving it is interesting to note that although static web pages can be archived and studied in their entirety later it is impossible to do this later with dynamic web pages which contain dynamic content generated from data bases (for instance e-commerce sites). The vast majority of information on the web is nowadays of this type.

Sensitivity testing of the preservation model

One issue which was raised in the discussion in London was the testing of the preservation model by varying different parameters. The model as it stands contains several parameters where rough estimates of values have been made. Since some of these parameters are multiplied in the model it is possible that the overall result can be very sensitive to small changes in some of the parameters. This could perhaps be carried out in co-operation between this author and Paul Wheatley.

It is recommended that this is given high priority in the work.

My proposal for LIFE 2 would be to develop an alternative Preservation model which would take as its starting point that the first migration is the costlier one but reduces the number of file types to about half a dozen ones. Also the costs for technology watch and tool software licenses would need to be revised since it could be assumed that there would be a large community of archival authorities which could collaborate on these issue, open source solution developed etc.

Another option for voluntary deposit or selective preservation of websites would be to restrict the file formats of accepted material from the start to the more common file formats.

The suggestion would thus be that the preservation formulas should somehow be adjusted to take into account the fact that subsequent migrations would be restricted to a few popular and well-understood file formats.

On page 106 in the section 8.8 further work this problem is acknowledged:

“Preservation actions like normalisation will often result in a repository having to deal with fewer file formats as preservation activities progress. This is difficult but perhaps crucial to model.

REFERENCES

Björk, Bo-Christer, 2007, **A model of scientific communication as a global distributed information system**. Information Research, , 12(2) paper 307 paper 307. Available at http://InformationR.net/ir/12-2/paper307.html

InterPARES (2002a),The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, Appendix 4, A Model of the Selection Function, http://www.interpares.org/book/interpares_book_m_app04i.pdf

InterPARES (2002b), The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, Appendix 5, A Model of the Preservation Function, http://www.interpares.org/book/interpares_book_n_app05ii.pdf

NIST (1993). *Integration definition for function modelling*. FIPS PUBS, Federal Information Processing Standards Publications, National Institute of Standards and Technology.

OAIS (2002) Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, CCSDS 650.0-B-1, Blue Book, Issue 1, January, adopted as ISO 14721:2003 <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

Digital Preservation Service Provider Models for Institutional Repositories: Towards Distributed Services
by Steve Hitchcock, Tim Brody, Jessie M.N. Hey, and Leslie Carr, *University of Southampton*
doi:10.1045/may2007-hitchcock
http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html