Cost Aspects of Ingest and Normalization

Ulla Bøgvad Kejser
The Royal Library
Postbox 2149
1016 Copenhagen K
+45 33 47 47 47
ubk@kb.dk

Anders Bo Nielsen
Danish National Archives
Rigsdagsgården 9
1218 Copenhagen K
+45 33 92 83 26
abn@ra.sa.dk

Alex Thirifays
Danish National Archives
Rigsdagsgården 9
1218 Copenhagen K
+45 33 92 23 69
alt@ra.sa.dk

ABSTRACT
The Danish National Archives, and The Royal Library and the State and University Library are in the process of developing a cost model for digital preservation: Each of the functional entities of the OAIS Reference Model are broken down into measurable, cost-critical activities, and formula are being tailored for each of these in order to create a generic tool for estimating the short and long-term costs of digital preservation. This paper presents an introduction to the subject of the costs of digital preservation and describes the method used to develop the Danish Cost Model for Digital Preservation (CMDP). It then describes how the OAIS functional entity, Ingest, has been included in the model. For institutions basing their digital preservation strategy on migration, a major cost pertaining to Ingest is *normalization*, a digital migration from production to preservation format and structure, which is often quite complex in comparison to the subsequent migrations within the archive. The paper accounts for three aspects of migrations, which are decisive for the costs: the required migration quality, when in the lifecycle the first migration takes place, and how often subsequent migrations are executed. Lastly – with view to increasing the model's precision – existing cost data from submission projects have been used to test the CMDP and the results of this test are described.

Categories and Subject Descriptors
H.3 m [**Information Storage and Retrieval**]: Miscellaneous.

General Terms
Measurement, Documentation, Economics, Standardization.

Keywords
Activity based costing, Cost model, Ingest, Migration, Normalization, OAIS Reference Model, and Preservation.

1. INTRODUCTION
The digital preservation field lacks economic models, i.e. models which account for costs and benefits, to enable justification of investments [2]. In recent years several projects have worked to define cost benefit models, such as the KRDS project [1] and the DANS cost model for digital archiving [7]. Also, some projects have developed cost models *per se*, such as the cost model created by the National Archives of the Netherlands [15], the LIFE Costing Model [9], and the NASA-CET [11].

In Denmark the Ministry of Culture has funded a project to set up a model for costing preservation of digital materials held by national cultural heritage institutions. The project has been undertaken by The Royal Library, The State and University Library, and The Danish National Archives, and consumed a total of 2 Full-time equivalents (FTE). In the first phase of the project (2009) a methodology for a cost model was developed, and designated the Cost Model for Digital Preservation (CMDP) [12]. The CMDP is based on the Open Archival Information System (OAIS) Reference Model [3] and on activity-based costing [5]. Furthermore, the work draws upon general costing principles defined in the International Cost Model Manual [13]. The CMDP is designed to be generic in order to enable calculation, estimation and comparison of the costs of digital preservation across memory and research institutions holding different types of digital materials. So far the model only addresses costs of digital preservation by the migration strategy. With time we envision to enable costing of the emulation strategy.

For developing the CMDP, we used the OAIS model to identify functions and divide them into delineated activities. We then identified the cost parameters (variables) related to the individual activities and operationalised them as formulas in a spreadsheet. Thus the CMDP spreadsheet tool represents modules based on the functional entities in the OAIS Model. In the CMDP costs are stated as the time it takes to perform an activity multiplied by the wage, plus the costs of any provisions. The CMDP only accounts for so-called cost critical activities, defined as activities that take a minimum of one person week to complete. A person week is set to 32 effective working hours, but as other variables in the spreadsheet, such as wages, it may be changed by users depending on local requirements. The CMDP includes all direct expenses of establishing and operating the preservation system as well as indirect costs, such as general administration (overhead). Eventually the model will also take financial adjustments, e.g. inflation, into account. While this is the ideal goal, the task is hard, and it may well be necessary to scale down the ambitions.

The spreadsheet and other documentation are available from the project web site[1].

In the first phase of the project we operationalised costs of the functions under the functional entity Preservation Planning and focused on the costs of the migration strategy. We also operationalised functions from related OAIS functional entities, which sustain Preservation Planning, especially functions under the functional entity Administration.

In the second phase of the project we have addressed the costs of the activities within the OAIS functional entity Ingest and related functions from Administration. To improve the identification of Ingest activities we also analyzed the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [4], which provides a detailed description of the interactions that take place between the OAIS roles, Producer and Archive. Finally, to account for the costs of normalizations, we have improved the formula for digital migrations developed in the first phase of the project.

We have used OAIS terms as far as possible and these are, as in the OAIS standard, indicated by initial capitals, e.g. Ingest. As in OAIS we use the term Archive to denote any organization devoted to long-term preservation.

In the remainder of this article we present the results of the second phase of the project describing cost aspect of Ingest and in particular cost associated with normalization: In section 2 we present our analysis of the Ingest functions and the identified cost dependencies. In section 3 we analyze format obsolescence and different cost drivers in digital migration, including migration quality, timing and frequency. In section 4 we describe how the costs of migrations have been modeled in the CMDP, including the cost of monitoring and executing migration actions. We describe the results of testing the CMDP on empirical cost data in section 5, and conclude in section 6.

2. INGEST OF DIGITAL INFORMATION

As a first step in identifying activities related to the Ingest of digital information into an Archive, a flow diagram was prepared based on an analysis of the functional descriptions in the OAIS standard (see Figure 1). Note that the activity Generate SIP (Submission Information Package) is not part of the standard (see explanation below). The flow analysis also helped avoiding that critical activities were overlooked or accounted for more times.

In addition to the OAIS standard we consulted the PAIMAS standard. The strength of PAIMAS is that it includes a checklist for defining a Submission Agreement, specifying all the details about a submission necessary for ensuring long-term preservation of the information. PAIMAS also describes activities related to the transfer of data and the validation of the transfer.

2.1 Submission Projects

PAIMAS divides a submission project in four phases:

1. The purpose of the preliminary phase is to determine whether a submission project is feasible and financially viable. The phase comprises the first contact between Producer and Archive, the provisional definition of the project's objective and context, a draft description of the digital information and its structure, and the writing of a draft Submission Agreement.
2. The formal definition phase negotiates the Submission Agreement between the Producer and the Archive. It describes the design of the SIP and the digital information to be submitted. Also it determines legal and contractual terms as well as security, and describes how transfer and validation of the transfer are to take place. Finally, it sets up a timeframe for the project.
3. The transfer phase ensures that the SIP is transferred from Producer to Archive, and that the Archive's initial processing of the information takes place according to the Submission Agreement.
4. The purpose of the validation phase is to ensure that the transfer of the digital information is validated according to the requirements outlined in the Submission Agreement.

Definition of a formal Submission Agreement does not necessarily occur as part of a submission. This depends on the nature of the submission and the power balance between the Producer and the Archive. In some countries an Archival Act can mandate archives to specify SIP designs, and in this case the balance of power is in favor of the Archive. In other scenarios, the Archive has to accept SIPs from the Producer as they are. This is typically the case within the library and research sector.

Even if no Submission Agreement is formally required it may still be important for the Archive to analyze the PAIMAS checklist for the Submission Agreement and determine how these issues will be handled. As such, the Submission Agreement constitutes an important part of any Archive's policy and strategic planning documentation.

2.2 Ingest Flow and Cost Dependencies

Below we describe activities under Ingest in detail and the identified cost dependencies. No specific costs of are reported in the article since they often depend on several preconditions, such as type of material, volume and format complexity. To calculate actual costs please consult the spreadsheet.

The cost of the core preservation system, i.e. the system, which e.g. manages notifications and the reception and transfer of information, is assumed to be accounted for in the Common Services functional entity of CMDP. This module has however not yet been modeled in CMDP.

2.2.1 Negotiate Submission Agreement

In the OAIS Model the Submission Agreement is negotiated by the Producer and the function Negotiate Submission Agreement under Administration. The agreement must cover all parts of the submission project, including a data submission schedule and an assessment of the required resources to support the submission.

The costs of negotiating a Submission Agreement are first and foremost dependent on the balance of power between the Producer and the Archive, the diversity and complexity of the data, how well the data are documented, and the size of the submission project.

2.2.2 Generate SIP

If an Archive bases its preservation strategy on migration and receives SIPs in production formats, which are not regarded as suitable for long-term preservation, it is common practice to

[1] www.costmodelfordigitalpreservation.dk

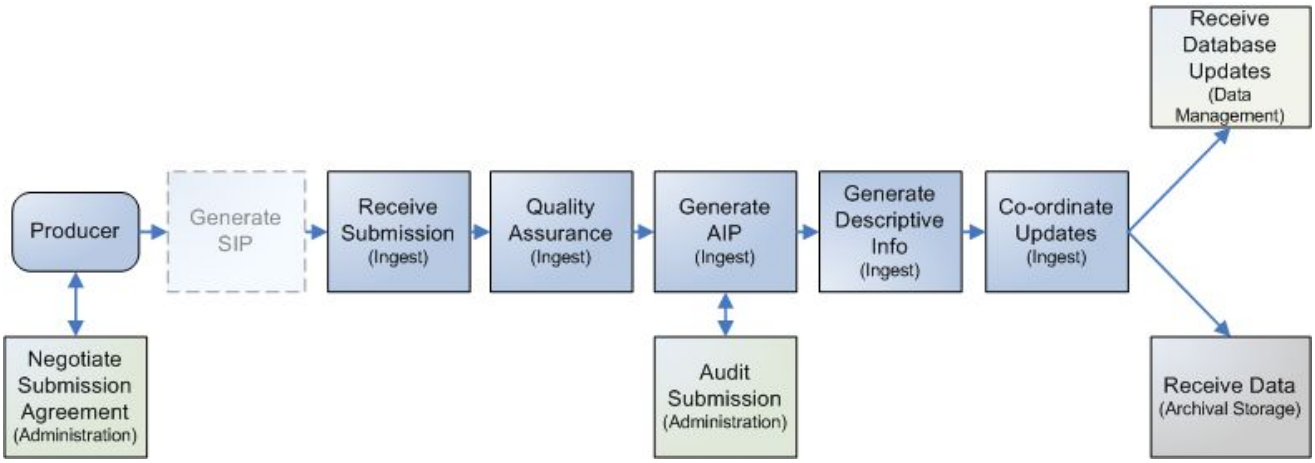

Figure 1 The flow between the OAIS functions from Producer to Archive on submission of digital records

normalize the information at Ingest, i.e. to migrate it from production to preservation formats and structures. Often normalizations are more complex and thus more costly than migrations from one preservation format to another. In OAIS, normalization is carried out by the Ingest function Generate AIP (Archival Information Package).

However, if the balance of power is in favor of the Archive, it can require the Producer to submit SIPs with normalized data in validated submission formats and/or enrich the packages with metadata before the SIPs are transferred to the Archive. As has also been noted as part of the five year review process of the OAIS standard [8], such preparation of the SIPs is not explicitly accounted for in the OAIS Model. In order to be able to calculate these costs, we have added the optional function Generate SIP to the CMDP.

If Ingest includes normalization this entails considerable costs, irrespective of whether these costs are carried out by the Producer or the Archive. It is the balance of power between the Archive and the Producer that determines who pays for the cost of normalizations. The cost of normalization and dependencies are described in detail in section 4.2.

2.2.3 Receive Submission

When the Submission Agreement has been concluded, the Producer transfers the SIPs to the Archive, where they are received by the Ingest function Receive Submission and placed in temporary storage. The transfer may be by movable storage media such as DVD or hard disk, or via a network.

In the CMDP we assume that reception of SIPs is an automated process, and thus not cost critical. We have also excluded costs of providing temporary storage for receiving SIPs because this extra storage capacity is likely to be reused for other activities.

2.2.4 Quality Assurance

The SIPs are then checked for errors by the Quality Assurance function, typically by a check-sum control. If the packages are in order a confirmation of reception is sent to the Producer. If there are errors the Producer is informed, so that the packages can be corrected and transferred once again. It is important to notice that in the OAIS Model this function only verifies the integrity of the data. It checks neither the authenticity nor the intellectual content, which in OAIS is managed by the Audit Submission function under Administration (see section 2.2.6).

This quality assurance process is assumed to be automatic and thus only entails costs for the establishment and maintenance of the quality assurance system as well as the potential error handling.

2.2.5 Generate AIP

The Generate AIP function transforms SIPs to AIPs, and may entail normalization. The function may also request that the functional entity Data Management provides additional information necessary for a full description of the package.

As the costs of generating, SIPs those of generating AIPs are described in detail in section 4.2.

2.2.6 Audit Submission

The Audit Submission function is part of the functional entity Administration. It checks whether the generated AIPs fulfills requirements and sends an audit report back to the Generate AIP function. If any errors or defects are identified the Producer is noticed and can then transfer the SIPs again. The validation phase specified in the PAIMAS standard corresponds to the Audit Submission function as well as the Quality Assurance. Audit typically comprises ensuring that the information packages are complete, that integrity is maintained, and that they fit the data model, including that the agreed data formats have been used and their syntax have been maintained.

Depending on the requirements defined in the Submission Agreement, the audit of AIPs can be cost intensive, because they are often manual.

2.2.7 Generate Descriptive Information

The Generate Descriptive Information function subsequently extracts Descriptive Information, i.e. primarily metadata used to search and retrieve the packages, from the AIP and other related sources, and sends the information, via the function Co-ordinate Updates, to Data Management.

The extraction of Descriptive Information is assumed to be an automatic process, and therefore not cost critical. Note that institutions may use considerable resources for providing metadata to the objects at Ingest. OAIS does not explicitly include such data qualification, and therefore it may be accounted for under Generate SIP in the CMDP.

2.2.8 Co-ordinate Updates

The Co-ordinate Updates function then sends the AIPs to Archival Storage, which confirms receipt and assigns an ID for the AIP package when storage has been executed and verified. Co-ordinate Updates includes this ID in the Descriptive Information and sends it on to Data Management.

The co-ordination with Data Management and transfer to Archival Storage is assumed to take place automatically and these activities are therefore not regarded as cost-critical in the CMDP.

3. COST DRIVERS IN MIGRATION

Below we analyze format and system obsolescence and its influence on when migrations should take place. Thereafter we describe three important drivers of cost in digital migration actions, namely migration quality, timing and frequency.

3.1 Format and System Obsolescence

Virtually irrespective of its format data in themselves are of little value, as they require a system, which can interpret the data format in order to reproduce the data content in an understandable form. The obsolescence of formats is therefore dependent on the obsolescence of the software that is to interpret the format.

With regard to data preservation, it is not sufficient that a program can interpret data in their current form (input format), as it must also be able to write data in a suitable contemporary format (output format). Especially in earlier times it has been emphasized that there was often a viewer for a format, which was therefore not obsolete. This is not a tenable argument; however, as the result is dependence on new generations of the viewer, and in addition the data cannot be further processed in other systems. The answer thus merely leads to the new question of when the systems in question that can read data in its present format (input format), and write the data in a contemporary output format, are obsolete?

For as long as new generations of systems are developed that can read data in its present format, and write the data in a contemporary output format, there is no real problem of obsolescence. This does require, however, that the system is tested, and that the reproduction is acceptable. When a generation of the system is developed in which the data format in question can no longer be interpreted, or just cannot be written in a suitable contemporary format, it is necessary to use the previous generation of the system to do this, thereby becoming dependent on its lifetime.

More and more formats can be interpreted by systems that are several generations younger, although naturally there are limits. It is therefore necessary to use the previous generation of the program to read the formats and save them in a contemporary output format. As for most new generations of programs and data formats there are a number of functionalities and derived data that are not supported in the newest generation.

The lifetime of systems does not end on the same day that a new generation of the system is born, or a competing system takes over the market. The lifetime of systems is dependent on the costs of their use and maintenance. For as long as a system can run on contemporary hardware and be integrated with contemporary systems the costs of its use and maintenance are manageable. Thus, neither the system nor the data formats it can interpret and write to are obsolete.

A known example of obsolescence is the BBC Domesday Project: In 1986 the BBC published an extensive modern multimedia edition of the famous Domesday Book that describes England in the 11th century. The BBC Domesday edition consisted of letters, maps, images, statistical data, videos, etc., stored on two interactive laser discs, LaserVision Read Only Memory (LV-ROM). In 2002, it was feared that the discs would become unreadable due to the technological obsolescence of the data storage medium and it was necessary to use migration, emulation and re-digitization in order to preserve the data. This was technically possible with great difficulty, and the high costs were a clear indication that the formats had become obsolete.

“The lesson of this digital preservation project is that if you have enough time, individual skill, dedication and imagination then almost anything is possible, provided that you don't leave it too late. If you start counting the cost this may seem an expensive project, but then the value of the record is high too - and that applies equally to the original Domesday Project. There is of course a great need to preserve other electronic records in a routine and predictable manner, and this rescue project is not a suitable model to be followed in such cases. The National Archives is working on ways to make this possible in future” [6].

This is despite the fact that from the outset the project’s creators were aware of the preservation risk and had in due time submitted data and documentation to an archive that did not handle the matter satisfactorily.

“The deputy editor of the Domesday Project, Mike Tibbets, has criticized the UK's National Data Archive to which the archive material was originally entrusted, arguing that the creators knew that the technology would be short lived but that the archivists had failed to preserve the records effectively [16].

Do we always have to rely on existing systems to be able to read data in a given format? In practice yes, since even with exhaustive documentation of the format it is normally a very demanding task to develop a system, to read data in one format and write it in another. The exception is the very simple formats for which, at a modest cost, it is possible to develop systems that can read data in one format and write it in another. Examples include TIFF, UTF-8 or XHTML.

3.2 Migration Quality

As for many other costs, the quality level of migrations is decisive to the level of costs. Migration quality is determined primarily by the choice of the output format, and by the error tolerance on migration of data from the input format to the output format.

3.2.1 Selection of Output Format

High quality in terms of an advanced output format, which enables preservation of a wide range of functionalities, rather than a simpler output format will entail significantly higher costs. This is because on migrating data from input format to output format programs must handle how all data in the input format is migrated

to an equivalent place in the output format, and it must be controlled that this has taken place (see below). For example, migration from one word processing format to another word processing format will result in higher costs than migration to a simple format in the form of a graphic bitmap format, as the word processing format contains far more information than a graphic format. This is a general observation, since in practice the situation may be that the system that migrates data from the input format to an advanced output format is far superior to the system that migrates data to a simple output format. The choice of output format is also essential to determining how *often* migration should be performed (see section 3.3).

3.2.2 Selection of Error Tolerance

With regard to error tolerance on migration of data, high quality in the form of a low error tolerance will bring about significantly higher costs than a high error tolerance. This is because a low error tolerance will typically require extra funds for the provision, operation and further development of the system for the migration. In addition, it will be necessary to use extra resources for error control, and especially error correction. Irrespective of the choice of error tolerance there will normally be higher costs for the error handling of an advanced output format than of a simple output format. This is because there is more chance of something going wrong, and it is more expensive to correct the individual errors.

Selection of output format and error tolerance can furthermore be combined, depending on the purpose of preservation and the data content. Note that an advanced output format thus does not necessarily entail a low error tolerance, just as a simple output format does not necessitate a high error tolerance.

3.3 Migration Timing

An important factor with regard to the costs of migration is when in the archival lifecycle, migration should be performed. There are different tactics for when it is best and least expensive to migrate, including to which output format.

3.3.1 Migration to Standardized Format

One tactic is to migrate data to a contemporary standardized format, as seldom as possible. The argument behind this tactic is that by migrating to a contemporary standardized format the number of migrations is reduced, and thereby the risk of unintended changes. The reason is that the lifetime of a standardized format is expected to be significantly longer than for other formats, as several systems will be able to read data in the format and write it in another. In addition, the standardized format should make it less expensive to provide, operate and maintain systems for actual migration, due to the larger supply available.

On the other hand, the number, market penetration and system support of contemporary standardized formats is estimated to be modest. It is therefore necessary to either select simple output formats, or to perform migration almost as frequently as if the next generation of the input format had been chosen as output format.

3.3.2 Migration to the Latest Format

Another tactic is to continuously migrate data to the most recent output format. The argument behind this tactic is that it adopts the situation of other IT users with a need to migrate data from the previous generation of the format to the latest as correctly and inexpensively as possible. This makes it possible to benefit from the systems for the latest generation of the format, which must be assumed to be the best for reading the immediately preceding generation of the format.

On the other hand, the frequent migrations are cost intensive and increase the risk of unintended changes. Moreover, the programs for the newest generation of the format are not always the best to interpret the previous generation [10]. Sometimes it is necessary to wait for the following generation to achieve better reproduction. In addition, suppliers and users generally seem more interested in creating new data in new formats, rather than reading older data in older formats in the new generations of the systems, that nothing particular is done to facilitate migration. It is thus difficult to find systems that handle mass migration of the previous to the current generation of the format.

3.3.3 Migration on Demand

A third tactic is called migration on demand and entails that if the data are in a relatively common and documented format the data are retained in the original format and not migrated to another format until the data are requested. The argument behind this tactic is that it is estimated that the number and variation in the use of data formats is continuously narrowing, and that market penetration, openness and documentation are widening. The probability that in a few years it will be possible to read a previously relatively common and documented format is therefore so high that there is no reason to perform migration before then. This saves a large number of intervening cost intensive and hazardous migrations.

On the other hand, the risk is considered by some to be too high, i.e. the probability that after a number of years there will, after all, not be any system that could interpret the format. In addition, depending on the output format, it is often an advantage to migrate shortly after the data are created, as many formats are not isolated, but depend on external data, for example fonts in the system, or references to images or other data outside the format that may have been altered after a number of years. These are external dependencies of which the encapsulation requires systems that have to be acquired, operated and further developed. Some standard programs, such as MS Word 2010, now support partly embedded fonts.

3.4 Migration Frequency

The immediate answer to how often migrations should take place is as seldom as possible, while bearing in mind the risk of obsolete data. This is because each migration entails a risk of losing information when data are migrated from one format to another, and because each migration entails costs.

On the basis of the current situation our tentative estimate of when a format is obsolete is eight to 20 years after its introduction on the market.

Twenty years is based on the furthest horizon we dare estimate within digital preservation. Eight years is based on the time within which we estimate that it will generally still be possible to run a program that can read data in its input format and write it in a suitable, contemporary output format.

3.4.1 Format Lifetime Parameters

It is extremely difficult to estimate the lifetime for a given format between the extremes of eight and 20 years, but we assess the vital parameters to be market penetration, complexity and documentation of the format. Lifetime increases with widespread

use, low complexity and good documentation. The three parameters are mutually dependent, which does not make the estimate easier. Simple, well-documented formats are often widely used, and simple formats are often well-documented.

In this context market penetration concerns the number of users, but especially the number of different systems that use the format. IT is a market with considerable network effects, and the aim is to develop programs that can fully read a competitor's format, but only write in their own formats; otherwise it is necessary to compete on the competitor's home turf, or on an equal footing.

Complexity is dependent on the number of types of information in the format, including the functionality in the system that is reflected in the format. Highly complex formats are often replaced more quickly (than formats of low complexity) by new generations of the format, as producers or users require even more functionalities. As stated, formats of very low complexity can be independent of existing systems because on the basis of the documentation, if it is good, it will be possible, without prohibitive costs to develop a system to interpret the format.

Documentation concerns the description of the structure and use of the format. A characteristic of good documentation is that it gives others besides the original creator of the format a feasible opportunity to develop systems that can interpret the format. It will at times also be necessary to have partial documentation of the system in order to understand how to interpret the format. For documentation to be good it must first of all be accessible, and secondly include the entire structure and use of the format, and finally be explanatory, i.e. intended to ensure that others besides the original developers can understand the format.

4. COST OF MIGRATION ACTIONS

There are numerous costs related to migration, of which the most important are the provision, operation and further development of systems for:

- Ongoing monitoring of which formats are obsolete, and of which the content must be migrated to other formats.
- Actual migration of data from one format to another, including control that the data is not changed unintentionally.

The following cost-estimates are based on own experience and a review of the literature on this subject [1]. With regard to the further development of the migration cost formula, we have been inspired in particular by the guide: Software Development Cost Estimating Guidebook [14].

4.1 Costs of Monitoring

Costs must be defrayed for the provision, operation and further development of a system for identification and registration of all formats for all data, stating the precise version of each data unit.

In practice this entails that on ingest of data in the preservation system all data are analyzed, so that its formats can be identified and registered, and so that all data in a given format can be retrieved when it is transferred to another format.

This task can be handled by the Producer if the Archive can get the Producer to undertake the task, and trust the result, but in practice most preservation institutions will handle this themselves. Identification should in practice be followed by validation and partial characterization. This is because far too much data does not comply with its format, and that many formats are so rich in content that it can be necessary to have information on their characteristics, i.e. which parts of the format contain data.

4.1.1 Provision of Monitoring System

Provision of such a system currently requires that it has to be developed, although there are partial solutions in the form of JHOVE2[2], PRONOM and DROID[3]. We estimate that the costs of provision of the core of a modular system that via specific modules for the individual formats can perform reasonable identification, partial validation and a small degree of characterization will be 12-24 person months.

The costs of the development of the individual modules depend on the formats' complexity and documentation, and are estimated to be respectively exponentially increasing and diminishing. We estimate that the cost per format will be from a few person weeks for simple formats to several person weeks for advanced formats.

Going beyond what we unclearly call reasonable identification, partial validation and a small degree of characterization, we estimate that there will be a highly exponential increase in the costs. It has, for example, still not been possible to achieve a complete validation of PDF/A. It is currently necessary to use validators from several suppliers to cover as many areas as possible. It will not be possible to avoid incorrect identification or incorrectly formatted data. In practice, it must be hoped that the programs to migrate data to other formats are relatively error tolerant. It will not be possible to avoid a few errors without very high costs.

4.1.2 Operation of Monitoring System

We assume that monitoring takes place by manual review of the list of formats used and comparison of their development in the market, in order to assess whether some formats are becoming obsolete. Work is also taking place on the establishment of a joint international format register, the Unified Digital Format Registry (UDFR)[4], which will be able to streamline monitoring. Monitoring of the market means that for each format there is one or several system(s) that must be registered and stated as necessary to interpret the format. These systems' lifetimes must also be assessed, including whether the format is supported in the newest generation of the system.

The task of monitoring is highly manual, and we estimate that the cost is proportional to the complexity of the format. On this basis it is estimated that monitoring will take from a few person days to a few person weeks, and that it will most frequently have to take place every second year for a given format.

4.1.3 Maintenance of Monitoring System

Besides general maintenance, the maintenance of the system, for example in connection with a new operating system, also includes the development of new profiles for identification, validation and characterization of any new formats that the Archive might use.

[2] https://bitbucket.org/jhove2/main/wiki/Home
[3] www.nationalarchives.gov.uk/PRONOM/Default.aspx
[4] www.udfr.org/

4.2 Costs of Migration

In terms of costs the migration of data from one format to another can be divided into provision, operation and maintenance of migration systems:

4.2.1 Provision of Migration System

We assume that a migration system has the following modules to handle the required tasks:

A general module that on the basis of central registration of data and their format can retrieve the data in an information package (a SIP or AIP) of which the format is estimated to be obsolete, and unpack this data.

A general module to manage all information packages and data retrieved in the obsolete formats, as well as their status, throughout the migration process. For each body of data in a format the module must request the specific module created for each format, register the result, and if successful send the migrated data in its new format for repackaging with the unaltered data from the package, so as to create new packages. To ensure efficiency the module must be able to parallelize its requests.

Specific modules for each format that ensure that the data in the format is migrated with the system considered to be the most suitable for the process and in the required quality. These programs will normally be the same as were registered in conjunction with the monitoring of the format's obsolescence. To be able to automate migration the module must be able to control parts of the program's behaviour, for example so that it is not stopped by enquiries from the program. If an advanced output format is selected there may also be a need for further management of the program in order to migrate all the required information to the output format.

The costs of developing the above system are considerable, and reuse of others' solutions is an obvious alternative. We do not know any turnkey solutions, but a number of sub-solutions, such as Apache Hadoop[5] or Berkeley Boinc[6], might be used.

We estimate that development of the general modules takes 12-24 person months. The costs of the specific modules are not necessarily proportional to the number of formats, if a series of formats use the same program for migration. The test of the correct functioning of the module with a given format is, however, proportional, and the cost can therefore be almost proportional. Reuse of others' solutions is an obvious path to take, but we do not know of any such solutions. For each format, primarily the advanced formats, where there is a need, there are often full or partial solutions, such as Apache POI or Microsoft Open XML Format SDK[7], that can manipulate the running of a program or directly access the format. We assess, however, that the cost of directly accessing the format in the case of advanced formats, such as ODF or OOXML, exceed what is feasible for an individual preservation institution. The institutions must therefore await development in a wider community if the quality is to exceed that offered by turnkey programs.

We estimate the cost per format to be exponential to the format's complexity, and vice versa in terms of error tolerance. Furthermore, we estimate that the development of a module for a simple format with a low error tolerance will take a few person weeks, while an advanced format with a low error tolerance will take several person weeks.

[5] http://hadoop.apache.org/
[6] http://boinc.berkeley.edu/
[7] http://msdn.microsoft.com/en-us/library/bb448854.aspx

4.2.2 Operation of Migration System

The costs of operating the system are primarily related to error handling, which depends on how reliably the system has been developed to operate. In this respect the costs of development and subsequent error handling are often inversely proportional, and it is not easy to calculate the optimum distribution.

Error handling comprises actual operational interruptions in the areas for which the system has not been developed to operate reliably enough. It also includes the identification of errors in the individual modules, when a format cannot be migrated as expected. Finally, error handling concerns errors that the system does not know that it makes, and which can only be detected via subsequent random sampling. In other words, handling errors that it is known will arise; errors that are assumed to arise; and errors that are not expected to arise. When the errors have been identified it is necessary to decide whether they are to be corrected, and if so, how.

Depending on the migration quality selected, primarily the complexity of the output format and the migration's error tolerance, we estimate that monitoring per format per TB (Terabyte) takes from one person day to a few person weeks. Furthermore, we estimate that error correction takes up to ten times longer than monitoring.

Even though the costs can be compiled per format, there are still economies from migrating several formats simultaneously, for example on packaging and unpackaging, storage, and error handling. As formats do not die on the same day that they are declared to be obsolete, several obsolete or virtually obsolete best practice is to gather formats for simultaneous migration.

4.2.3 Maintenance of Migration System

Maintenance of the system comprises general maintenance, for example in connection with a new operating system, and the development of new modules for new formats.

5. TEST OF CMDP ON COST DATA

A questionnaire was sent to a number of public Danish authorities in order to collect information on their actual consumption of time and resources to produce information packages of data from IT systems in connection with submission to The Danish National Archives. The data collected have been used to test and adjust the Ingest module in the CMDP. If the authorities used an external supplier to prepare the information package, cf. Generate SIP, they were also requested to submit a copy of the contract for the assignment in order to obtain a full overview of the costs.

The questionnaire was sent to 34 authorities, of which approximately half replied. The responses received point in many directions and show that the authorities found it difficult to understand the questionnaire and compile the consumption of resources. Based on the responses received a tentative conclusion for large submission projects (>160 person hours) is that project management costs approximately 13% of the total submission project. The identification and the description of the digital objects and their references accounts for approximately 16%. Normalization accounts for approximately 66% and the testing of

the information package accounts for approximately 5%. The responses concerning the time spent on the physical submission are not included in the study as the responses showed that the question was not understood correctly. Furthermore, in the case that the authorities have used consultants, a high price is not necessarily equivalent to high earnings for the consultant, as the price/earnings ratio is not equal for all consultants.

6. CONCLUSIONS

In overall terms, we believe that the developed method of identifying Ingest costs is viable, although the cost model is not yet sufficiently detailed to give accurate results for all types of records: All empirical data originates from ingest of archival records, which means that the model is currently best suited to estimate the costs of this particular type of material.

An important conclusion from the survey on submission projects was that the normalization of formats is by far the highest Ingest cost, namely around two thirds of the total costs. The fact that normalization also entails cost-sensitive choices such as migration frequency, timing, quality, error tolerance, emphasizes that this particular cost requires very special focus when considering the precision of the cost model. Likewise, the study indicates that the balance of power between Producer and Archive has great influence on the costs, and their distribution, so that this is an essential parameter in the model.

The study also confirmed a former key finding: The choice of the digital object (the format), its complexity and volume as the basic calculation units, makes the model potentially generic and thereby capable of calculating the costs for various digital collections. In order to achieve accurate results for all types of digital materials there is, however, a need to expand with several parameters for each object type, for example number of objects.

Likewise, more work is required to increase the precision of the model. By default the CMDP has a number of estimates such as format complexity, lifetime and thereby migration frequency, which are dependent on the actual preservation scenario, highly uncertain and subject to debate. In order to address this problem, the model makes it possible to state other values than those proposed default.

Generally, our work shows that preservation institutions depend to a great degree on being able to use standardized solutions, as it would be very expensive for them to develop a number of tailored tools corresponding to the number of types of ingested data.

The implementation of the model in the spreadsheet has proved to be problematic. The requirements of transparency and precision cannot be fulfilled simultaneously. In a new version of the model, with greater precision, it will therefore probably be necessary to sacrifice some of the immediate transparency and state the formula in code. The lack of an actual user guide and user interface to the spreadsheet is another deficiency, as the model in its current form is very difficult for external parties to use.

Currently, we have funds for 1½ man-month for developing the cost module for Archival Storage, which is a somewhat easier task as we have longer experience with these functions and more empirical data is available.

If the precision of the CMDP is to be increased, the remaining modules of CMDP are to be developed, and the model expanded to account for different preservation strategies and different digital collections, additional work and funding is needed. For the purpose of further development of the model, the project has stayed abreast of the international development of economic models for digital preservation. It is our hope that this focus will lead to formal or informal cooperation with other stakeholders in the future, as it is assessed that both the interest in and the necessity of greater certainty in this field are generally considered to be substantial.

7. ACKNOWLEDGMENTS

Thanks to the Danish Ministry of Culture for funding this study.

8. REFERENCES

[1] Beagrie, N., Lavoie, B., Woollard, M., 2010. Keeping Research Data Safe 2, Final Report, Charles Beagrie Limited, www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

[2] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. Sustainable Economics for at Digital Plant: Ensuring Long-Term Access to Digital Information, Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Consultative Committee for Space Data Systems (CCSDS). 2002. Reference Model for an Open Archival Information System (OAIS), 650.0-B-1, Blue Book (ISO14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] Consultative Committee for Space Data Systems (CCSDS). 2004. Producer-Archive Interface Methodology Abstract Standard (PAIMAS), CCSDS 651.0-M-1, Magenta Book. http://public.ccsds.org/publications/archive/651x0m1.pdf

[5] Cooper, R., Kaplan, R.S., Maisel, L.S., Morrissey, E. & Oehm, R.M. 1992. Implementing Activity-Based Cost Management: Moving from Analysis to Action. New Jersey. Montvale, Institute of Management Accountants.

[6] Darlington, J., Finney, A. & Pearce, A.. 2003. Domesday Redux: The rescue of the BBC Domesday Project videodiscs, Ariadne Issue 36. www.ariadne.ac.uk/issue36/tna/intro.html

[7] Data Archiving and Networked Services (DANS). Costs of Digital Archiving vol. 2. www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2

[8] Higgins, S. & Boyle, F. 2006. Response to CCSDS's comments on the OAIS Five-year review: recommendations for update, The Digital Curation Centre (DCC) and The Digital Preservation Coalition (DPC). www.dpconline.org/events/previous-events/427-oais-5-year-review-follow-up

[9] Hole, B., Lin, L., McCann, P. & Wheatley, P. 2010. LIFE3: A Predictive Costing Tool for Digital Collections, In: Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects, Austria www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf

[10] Karjalainen, M. 2010. Large-scale migration to an open source office suite: An innovation adoption study in Finland, Academic Dissertation, Faculty of Information Sciences of the University of Tampere. http://acta.uta.fi/pdf/978-951-44-8216-8.pdf

[11] NASA Cost Estimation Toolkit (CET). http://opensource.gsfc.nasa.gov/projects/CET/CET.php

[12] Kejser, U.B, Nielsen, A.B., Thirifays, A. 2011. Cost Model for Digital Preservation:Cost of Digital Migration. In: The International Journal of Digital Curation, Issue 1, Vol. 6, pp. 255-267. www.ijdc.net/index.php/ijdc/article/view/177

[13] OECD. 2004. International Standard Cost Model Manual to reduce administrative burdens. www.oecd.org/dataoecd/32/54/34227698.pdf

[14] Software Technology Support Center (STSC) Cost Analysis Group, U.S. Air Force. 2010. Software Development Cost Estimating Guidebook www.stsc.hill.af.mil/consulting/sw_estimation/SoftwareGuidebook2010.pdf

[15] Slats, J. and Verdegem, R.. 2005. Cost Model for Digital Preservation. Proceedings of the IVth triennieal conference, DLM Forum, Archive, Records and Information Management in Europe. http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf.

[16] Tibbets, Mike, 2008, ACM Committee on Computers and Public Policy, Forum on Risks to the Public in Computers and Related Systems, Vol. 25: Issue 44. http://catless.ncl.ac.uk/Risks/25.44.html#subj