Report on Knowledge Exchange workshop
Cost models for keeping knowledge: economic models for digital preservation
Copenhagen, 11 June 2012

Executive summary
Twenty seven participants attended the workshop organised by the Nordbib programme and Knowledge Exchange (KE) on cost models for keeping the results of scientific and scholarly research. Six projects from four countries presented the work they had done on costing the preservation of knowledge. Several points arose when comparing the models. It is notable that models are being developed in various institutions in parallel. On the one hand it is valuable to share and to prevent double work taking place, on the other hand institutions have diverging demands regarding the functionality and coverage of a costing model. There is quite a variation in the level of detail included in cost models. It would be valuable to have a simple lowest denominator model as a standard, comparable to OAIS. This would allow comparisons of costing across institutions, countries and disciplines. This generic model would also help inform decisions involved in storage, for example whether to keep it in-house or store data in the cloud.
To show the value of data storage and preservation the benefits should be communicated to the stakeholders, this will require different arguments for different stakeholders and on different timescales. It is important to aim at price models that enable the development of services in a modular fashion.

Introduction
Hans Pfeiffenberger kicked off the workshop and welcomed all the participants. The cost of preserving research outputs is quite significant and therefore forces us to think hard about the way to arrive at clear calculations as well as about the benefits of keeping this knowledge. There is also the need to consider selection processes as a component of cost reduction. He introduced the speakers that each presented a project on cost models in the four countries.

Heiko Tjalsma from DANS presented their cost model on running a data centre. Two major models were developed: an Activity Based Costing model (ABD) and a balanced score card. The balanced scorecard looks at how the processes have an impact for the users and how this relates to the strategy of the institute. The costs of DANS can be split into direct costs (related to data archiving) and indirect costs (other staff, fees for IT). Labour is the most important part of the cost. All auxiliary costs can be distributed over primary costs. In the example of archeology 21% is direct cost.
The aim of this exercise was to provide accountability, to show what the cost is of archiving a dataset. As funding changes, DANS is looking into simple models which are easy to communicate to funders, for example by only charging for larger datasets.

Paul Wheatley presented the LIFE models. In these models the activities were mapped according to the lifecycle of preservation. It proved hard to identify the purpose of the costing, to define what the figures would be used for. The maturity amongst projects was very varied.
It was difficult to decide at what level of detail the modeling should take place. A very high level model contains little information, a detailed model is complex to develop, maintain and interpret. At present a lot of

---

duplication is taking place, both in studies and events and coordination is required. Paul has listed a series of initiatives on a wiki and please add 3 if you have a project to share.

Ulla Bogvad Keyser of the Danish Royal Library presented the CMDP model which has been developed by the Danish National Archive the national library and other libraries. The aim was to assess current and future costs to be able to use resources more effectively. The project team went through the activities and placed these in the model in order to identify dependencies. Labour and acquisition costs were estimated. It proved very difficult to estimate future costs. The resulting model can be used to assess different solutions for storage but does not take price developments into account. The challenges they ran into was the lack of empirical cost data, the lack of resources for development of economical models for digital curation and the lack of consensus on strategies for development of economical models. They agreed that collaboration between cost projects would be valuable.

Sabine Schrimpf presented the Digital Preservation for Libraries (DP4Lib) cost model. The reason to undertake this project was that services offered to third parties should be priced and it was unclear how much these costs were. The goal was to design a framework recording all costs and allocating them. They used existing models but added their own twist by using the ‘Cost by service’ approach. DP4Lib used the OAIS model but grouped it on a higher level. Although they set off to provide a simple model, the spreadsheet became just as complex as existing models. There are sheets for each of the phases (ingest, curation and access). DP4Lib are also collaborating in the Nestor working group on costs and in the APARSEN work package on costs and benefits of data collection and modeling.

Neil Beagrie presented the Keeping Research Data Safe (KRDS1+2) reports. Twelve institutions contributed to this work. The partners felt that one generic spreadsheet on costs was not the main priority and so they focused on benefits and impact. Three stages in the process were identified; the pre-archive part (e.g. metadata creation) was special as it is not included in all cost models. The projects delivered various outputs including fact sheets and a user guide. One lesson learned was that creating a model for people did not appear to be the solution, they should go through the process themselves.

The KRD-projects are also taking research data into account and are considering the whole cycle of service costing. The projects also provided a benefits analysis toolkit with a brainstorming tool to see which benefits are relevant to specific projects. In order to identify who benefits there is a framework which will allow the user to map the benefits in three dimensions.

Yvonne Friese presented the work of the Nestor working group on costs. This working group consists of 6 members from 5 institutions. One of the goals was to implement the DP4Lib model in other institutions. Using (a shared) cost model would offer the benefits of:
- budget planning
- achieving reliability in planning
- keeping control of costs
- improved accountability

However, sharing information on costs can prove difficult. Software vendors do not allow the publishing of the costs of licences and institutions are also not always willing to make some of their costs public. The whole process of cost calculation proved costly in itself. It has now provided guidelines for all German speaking institutions looking into calculating the costs of storing knowledge.

After these various perspectives the participants split up into four break out groups to discuss the issues from four different angles. After lively discussions the groups reported back on the outcomes.

3 Please share your project on the wiki at: http://wiki.opf-labs.org/display/CDP
**Report back**

**Challenges and opportunities**
Drivers are imperative for organisations to work on preservation. One challenge faced by institutions is to be more transparent in their costs. There are also barriers in using cost models. The uptake is not great at present and there is not a lot of experience in implementing them. Large consortium initiatives can be useful as there will be an enhanced sense of obligation within them to engage and use cost models which will help to drive their development.

In the discussion the question was raised whether one cost model for all was realistic. Different types of organisation needed different levels of detail on cost information. It was agreed that international collaboration would be useful although there were some question marks whether costs are comparable across countries. An international cost exchange could be set up with a submission template to enter information. It might be tricky to make this open as confidentiality might be required, but information short of actual explicit costs could also be usefully shared, e.g. information about the processes used to determine costs; or the definition of cost categories.

**Overlaps and gaps**
The use of terminology in the models can be confusing and act as a barrier to the implementation of a model. It might be useful to map the various models, a task which is currently undertaken by APARSEN. This will show whether a model covers the full lifecycle or only parts of it. The mapping will lead to an insight into the major overlapping parts of cost modeling and thus to an identification of a common denominator, and it will also make it easier to identify where certain models have gaps. The question was raised to include user requirements. This could be helpful in understanding how costs can be distributed over the different stakeholders. Ultimately, the task is to arrive at pricing models in order for institutions to offer or buy services for the preservation and management of digital collections. It is important to consider this task as a collective task in the future in which a number of actors have to be involved. Especially the costs for long-term archiving will have to be covered. Clarity on the costs will also help funders to develop funding schemes.

**Tasks for international collaboration**
There are several possible tasks to collaborate on internationally. Activities could focus on coordination and the sharing of data from projects. Organisations could also look into the lowest common denominator across cost models, what is the minimum set of factors that should be included in a cost model? Another approach would be to test the models using real costing data; this would mean that this data would have to be shared.

A next step could be organising a workshop to get a small group of experts to work on a high level abstract model. They could start costing preservation and build on the model from there. It is worth noting that other parties are already working on linking costing to an OAIS type model. Chris Rusbridge and Brian Lavoie (formerly members of the Blue Ribbon Task Force for Sustainable Digital Preservation and Access) have done some preliminary work on an Economic Sustainability Reference Model. This work requires more development and ultimately needs broad community buy-in. In general the various models have started to converge lately and people are building on others work.

Cost items must be balanced against the benefits. It would be good to focus on the positive benefits of the work. At present a lot of focus is laid on the negative impact of not archiving, it would be good to show the benefits of actually archiving.

**Benefits of digital preservation**
When considering the benefits of digital preservation it is important to keep an eye on the stakeholders involved. For example, funders and data creators will appreciate very different benefits and therefore need different messages to convince them. It would therefore be valuable to collect a range of different success stories on the (open) accessibility of data.

The Sloan Digital Sky Survey (SDSS) was named as an example and others would be welcome. In the SDSS case, 50% of all articles resulted from the re-use of data. There may well be other examples where data can be capitalized on only after 10 or maybe 40 years of collecting.

It was noted that it is important to consider the timescales when benefits will become apparent. As governments are focusing on showing return on taxpayer investment, typically within 5 years, it is important to also be able to show short term benefits, even if they are working on setting up a legacy (‘in the case swine flu comes back in 50 years’).

As new opportunities arise from the preserved data it is worth looking into how to measure these new opportunities. This would help to persuade funders that these initiatives also need support. A different approach was taken in the KRDS project, which showed the risks of not storing data, which costs have been avoided by storing data.

It was raised that there is a difference between benefit and impact. There is a value getting users to register as it shows that other sectors are also using the information. When arguing the value of preserving
knowledge you will have to convince different audiences and here the difference between benefit and impact can be useful. It is worth noting that most of these and other benefits named are based on the open access to data.

Closing words
Jens Klump wrapped up the workshop by raising some final thoughts. When considering how to proceed with the data archive it might be necessary to include other perspectives, for example that of the data creator and user. The cost models discussed are at different levels of abstraction. Once you implement these models you start to see the differences. It is valuable to test the various models with different data. This could lead to best practices. We could work to a cost model, which can become a standard like OAIS, which then could be used to benchmark existing data. It is crucial to thereby provide clarity on the cost of storage, show a price tag to take to the customers. Benchmarking can show the difference between actual cost and what you charge. This can also be helpful in deciding what can be done in-house and what can be stored in the cloud. There can be other considerations to store knowledge in-house, but this can then be more clearly balanced with the associated cost.

At present the storing of research data is a complex innovation and in such a situation it is typical that it is not clear what is going to succeed. To encourage development it is important to allow projects to fail. To show the value of data storage the benefits should be communicated to the stakeholders, this will require different arguments on different timescales.