Controlling the cost of digital preservation[1]

By Herman Uffen and Tamar Kinkel[2]

Background

For some time now, heritage organisations and other institutions have been in agreement about the need for the long-term preservation and accessibility of valuable digital resources. What is much less clear, however, is how much cost this would involve. In many cases, not enough consideration is given to the long-term cost of curating digital collections, often because very little is actually known about this aspect. As a result, such management costs are frequently not included in the ordinary operating expenses of the institutions in question. Moreover, many heritage institutions tend to assume that the long-term costs are very high, partly due to the exponential rise in the volume of material, whether digitised or born-digital. In many cases, incidental revenue (i.e. project income) is used to pay for long-term management costs. But is this assumption correct? And how can we control the cost of curating digital collections in the long term? We have designed a cost model for analysing and controlling the cost of digital preservation.

The cost of long-term accessibility: developing a cost model for digital preservation

The aim of the model is to make it easier to control the costs (i.e. both fixed and variable) of the long-term accessibility of digital heritage. In order to do so, it is absolutely vital to have a clear picture of the cost structure and the cost drivers, not only as they stand now, but also with a view to budgeting for future costs. A uniform financial framework is needed, so as to facilitate comparison. Institutions, policy-makers and funding-providers can then use this information in order to reach better informed decisions on investments in and the use of facilities for enabling long-term access.

This project builds on Dutch and international schemes that have been launched in the past to help institutions obtain a clearer picture of the cost of long-term access to their digital material. One of these was a project entitled *Collaboration to Clarify the Casts of Curation* (4C), which resulted in the development of the CCEx tool (Curation Costs Exchange), which institutions can use to analyse their own expenditure and compare this with the level of expenditure incurred by other institutions operating in the same field. The 4C project formulated the following vision of the future:[3]

---

[1] This article is the translation of the article published in the S@P jaarboek 19, 2028. Margriet van Gorsel, Erika Hokke, Bart de Nil, Marcel Ras (ed.), *Preserveren. Stappen zetten in een nieuw vakgebiek*, pp 163-173.

[2] The article is a result of the project "towards a Dutch cost model for Digital Preservation". This project has been started in 2015 by the National Coalition on Digital Preservation and continued by the Dutch Digital Heritage Network since 2018. The project aims to develop an activity based cost model for digital preservation. Project lead is by Herman Uffen and Tamar Kinkel from BMC Consultancy. More on the project is to be found on the website of the DDHN (in Dutch language).

[3] 4C, 2015. *Investing in Curation*; *A Shared Path to Sustainability*, 20 February 2015, pp. 1-26. http://www.4cproject.eu.

"In five years' time (2020) it will be easier to design or procure more cost-effective and efficient digital curation services because the costs, benefits and the business cases for doing so will be more widely understood across the curation life cycle and by all relevant stakeholders. Cost modelling will be part of the planning and management activities of all digital repositories."

The developers of the cost model for digital preservation examined a number of existing models[4] and sought to create consistency with the CCEx tool. The model is designed to help institutions to:

1. analyse the costs of digital preservation (in part to gain better control over these costs) and their component parts;
2. monitor these costs so as to be better able to keep them in check;
3. compare the costs of long-term accessibility with those incurred by other institutions, so as to learn and gain better control over these costs.

The model makes it easier for institutions to make certain policy decisions, including decisions on collection policy, on the use of staff and other resources, on partnerships with other institutions and on the necessary infrastructure.

Activity-based cost model

The cost model for digital preservation is an activity-based cost model focusing on activities that need to be performed as part of a process of long-term preservation and access. Providing long-term access to digital information is an active process involving the following activities:

- selection/pre-ingest
- ingest
- processing
- documentation
- archive
- access and distribution
- user support

Each activity also involves a number of overarching process activities:

- metadata
- preservation management
- infrastructure
- ICT

What exactly do the stages of the process involve?

- **Selection/pre-ingest** consists of activities that have to do with assessing the value of data, data criteria and data selection. This stage may also involve the provision of data guides, training,

[4] http://www.4cproject.eu/summary-of-cost-models/.

consultancy and communication with data suppliers on issues such as data formats, data management, data plans and data rights.

- **Ingest** is the process of receiving, reading and verifying the quality of data, categorising incoming data up to and including the entry of the data in question in an archive (but excluding processing). The ingest process may be either manual or automated, and often involves manual process stages such as quality checks.
- **Processing** involves the creation of operational product flows, non-operational products and reprocessing, process control (i.e. production planning, monitoring, etc.) and the generation of products. It also involves reprocessing new versions of existing products in accordance with a set timetable or a specified reprocessing capacity. It may also include the integration of different data sets and data mining, in which software is used to search through different data sets and combine data that satisfy certain criteria.
- **Documentation** involves developing or refining incoming data and product documentation (such as user manuals and catalogues) and user experiences, read software, systems information, maintenance and replacement. It does not include metadata.
- **Archive** means entering data in the archive, curating and managing data, and the long-term preservation of data, metadata and documentation in an archive. Incoming data may include data supplied from external sources or data generated or adapted by the institution itself.
- **Access and distribution** consists of retrieving data as requested from the archive, performing authorisation checks (if necessary), making partial selections, resampling, reformatting and converting data formats, reprojection or repackaging, and using electronic or physical media to disseminate the end product.
- **User support** involves assisting users contacting support staff, for example by answering questions, accepting orders, manning a help desk (where staff wait for users to call with their questions), reaching out to potential future users and training or educating current or potential future users. User support staff also possess the skills required to assist users in selecting and using data products.
- **Metadata** is descriptive or contextual information on a data object.[5] Activities relating to metadata may also have a bearing on other activities. Metadata is a continuous process in the long-term digital preservation of data. Metadata has been included as a separate activity on account of its importance and the amount of work it involves.
- **Preservation management** involves services and features for monitoring the digital environment, making recommendations and preparing sustainability plans for preserving access to and the use of digital objects and ensuring they remain accessible for use and reuse in the long term, even if the original digital environment becomes obsolete.
- **Infrastructure** consists of developing and operationalising data and information systems and the facilities needed in order to achieve the institution's objectives with regard to digital preservation, including the development and operationalisation of the data infrastructure, i.e. hardware, software and back-up facilities.
- **ICT** consists of sustaining engineering, i.e. maintaining and improving software applications (including bespoke software) and all post-implementations development work. Engineering support focuses on the institution's internal operations. Technical coordination, including engineering, has an external focus and involves supporting the institution as part of a system of institutions working in collaboration with each other.

Our decision to incorporate these process stages and overarching process activities was based on a study of a number of existing cost models for digital preservation.[6&7] Among the aspects

[5] The term 'object' includes a persistent identifier as a characteristic feature.
[6] http://www.4cproject.eu/summary-of-cost-models.
[7] Keijser, U.B., K.H.E. Johansen, A. Thirifays, A.B. Nielsen, D. Wang, S. Strodl, T. Miksa, J. Davidson, P. McCann, J. Krupp & H. Tjalsma, 'Evaluation of Cost Models and Needs & Gaps Analysis', 4C, (30 June 2014), pp. 1-99.

we examined were the variables and activities included in these models, their aims and their complexity. The cost model for digital preservation is based on the findings of this study. Additionally, the activities, sub-activities and related definitions are based on the OAIS[8] model, with further refinements being added in the wake of discussions in committees set up by the institutions participating in the project. The same committees also tested the model.

The cost model for digital preservation generates information on the cost structures and the cost drivers, and identifies the financial cost of the curation and preservation of, and the maintenance of access to, digital heritage. If the model is used over a period of years, it can reveal the impact of these factors on cost, i.e. the cost correlations. The correlations can then be used to generate a statistical cost forecast.

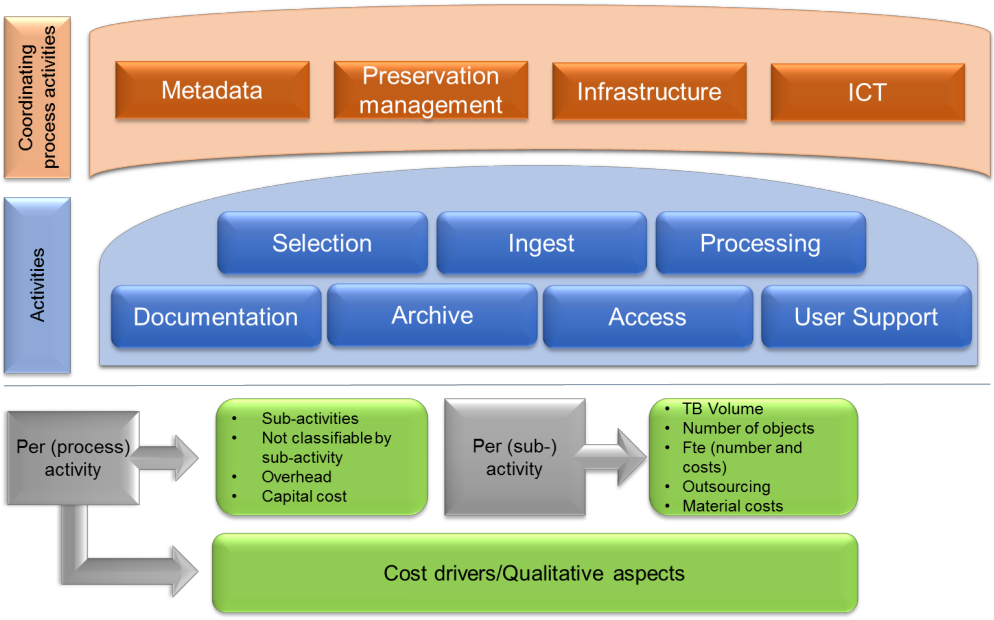

Figure 1: Dutch Digital Heritage Network Digital Preservation Cost model

Initial results of trial: first stage

A beta version of the cost model for digital preservation was trialled by ten institutions in 2016 and 2017. This exercise was the first stage of the road map drawn up by the Digital Heritage Network as part of the development of the cost model (see section entitled 'What next? Using a road map as a compass'). The testing stage generated a number of interesting findings about the cost of digital preservation.

At present, institutions have little or no information on costs. The use of the model generates information on costs and hence raises awareness within the organisation in question.

[8] The Open Archival Information System (OAIS) Reference Model was designed in order to identify the functions required for sustainable curation, and to formulate a common set of terms for this purpose. The model identifies six functions:
- ingest
- data management
- archival storage
- administration
- preservation planning
- access

Lavoie, B., 2014, *The Open Archival Information System (OAJS) Reference Model: Introductory Guide* (2nd edition), DPC Technology Watch Report 14-02, October 2014, pp. 1-37.

The first finding is that institutions have little or no awareness of the cost of digital preservation. Most institutions are in the early stages of financial awareness; only the relatively large and more experienced institutions have a fairly clear picture of the cost in relation to processes and the relevant cost drivers.

A second finding is that the cost model helps to paint a clearer picture of the internal costs. The use of the model by the institutions taking part in the trial helped to raise their awareness of their primary process and their financial structure. The model also offers management information that can be used to gain better control over costs.

Costs in the process of digital preservation

Most costs are incurred at the start of the process of preservation and access, i.e. during the selection, pre-ingest and ingest stages. If the suppliers or data originators can perform better in selecting, structuring and describing the incoming data, the level of cost incurred by the institutions curating the data could well be lower.

It is clear that a large proportion of the costs relate to the overarching process activities. This is bound up with the (clearly evident) fact that many institutions are still in the process of investing in long-term access and the relevant infrastructure, thereby creating depreciation costs. Most institutions go it alone; very few seek to form alliances with other bodies either within or beyond their own sectors.

Staff costs are the main cost component. At the majority of institutions taking part in the trial, staff costs accounted for more than 50% of aggregate expenditure. The fact that staff costs take up such a large share of expenditure underlines the importance of acquiring and sharing information on how staff costs can be controlled and making optimum use of staff resources.

Refining the cost model

An analysis of the sectors and the institutions taking part in the trial reveals wide variations in terms of objectives, context and background. In order to properly interpret the figures in the cost model, we need to factor in these different objectives, contexts and backgrounds and take them into account in making comparisons between institutions. The fact is that the cost drivers have their roots in these objectives, contexts and backgrounds, which therefore explain the relative level of cost. While comparisons have to date focused mainly on the similarities between institutions,[9] it would make sense also to look at the differences within and among sectors and to take these as learning points for improving collaboration among institutions.

For the time being, the number of institutions making use of the cost model in practice is not large enough to enable comparisons to be made between different institutions in terms of the cost of long-term access and to calculate correlations and regression (i.e. the relationship between costs and cost drivers) and identify their effects. This means that it has not been possible to fully exploit the model's potential at present.

[9] Van der Nat, J. & M. Ras, *Samen bouwen aan een netwerk van landelijke voorzieningen: Eindrapport van het onderzoek naar een landelijke infrastructuur voor duurzame toegang tot digitale informatie* (NCDD, April 2015) pp. 1-80.

Eye's experience with the cost model

This section looks specifically at the experience gained with the cost model by the Eye Filmmuseum in Amsterdam.[10] Not only was Eye closely involved in the development of the model, it also took part in the trial and was one of the first institutions to adopt the model in practice.

During the period in which Eye made use of the model, it found itself confronted by rising costs of digitisation and of curating and providing access to its digital heritage collection. These rising costs – a trend encompassing a number of years – came over and above the cost of curating the analogue collections.

In 2016, the Dutch Ministry of Education, Culture and Science asked Eye to undertake an analysis of its activities. This analysis was performed by the Boston Consultancy Group (BCG)[11] and was designed to restore the balance between the underlying, long-term trend in income and expenditure (together with non-recurring, i.e. exceptional, income and expenditure items) and, in doing so, to analyse the impact of the burgeoning digitisation process. The analysis produced information on the 'buttons needing to be pushed' (= the cost drivers) so as to take the right strategic action (i.e. make or buy decisions, see the section entitled 'What next? Using a road map as a compass').

The cost model for digital preservation helped Eye to undertake this exercise. It proved to be an interesting project, in that it enabled Eye to identify the costs affecting the entire ingest, storage and access chain. The model both helped to refine the calculations already made by Eye in the past and, as a result, to assist BCG in its analysis.

Although Eye's experience was that, although the cost model for digital preservation generated valuable information, it also pinpointed a number of issues to which solutions needed to be found:

- The organisation's existing processes were not based on activity-based costing (ABC); staff and associated costs were not attributed to activities.
- The financial accounts were not consistent with the ABC model.
- Projects and their related costs needed to be associated with both the direct costs (either recurring or non-recurring) and the funding of activities (i.e. exceptional items such as grants and ordinary income such as visitor earnings).
- Both the terms and definitions used in the model and those used by Eye needed to be reworded to make them consistent with each other.
- Overhead expenses (including the cost of premises) needed to be attributed to activities.

[10] www.eyefilm.nl.
[11] Boston Consultancy Group, *Financieel Onderzoek Eye* (Ministry of Education, Culture and Science & EYE, September 2016) pp. 1-54.

- In its current form, the model does not allow the future cost of access to objects to be calculated.

The following picture emerges if Eye's staff costs are allocated to individual activities in the cost model:

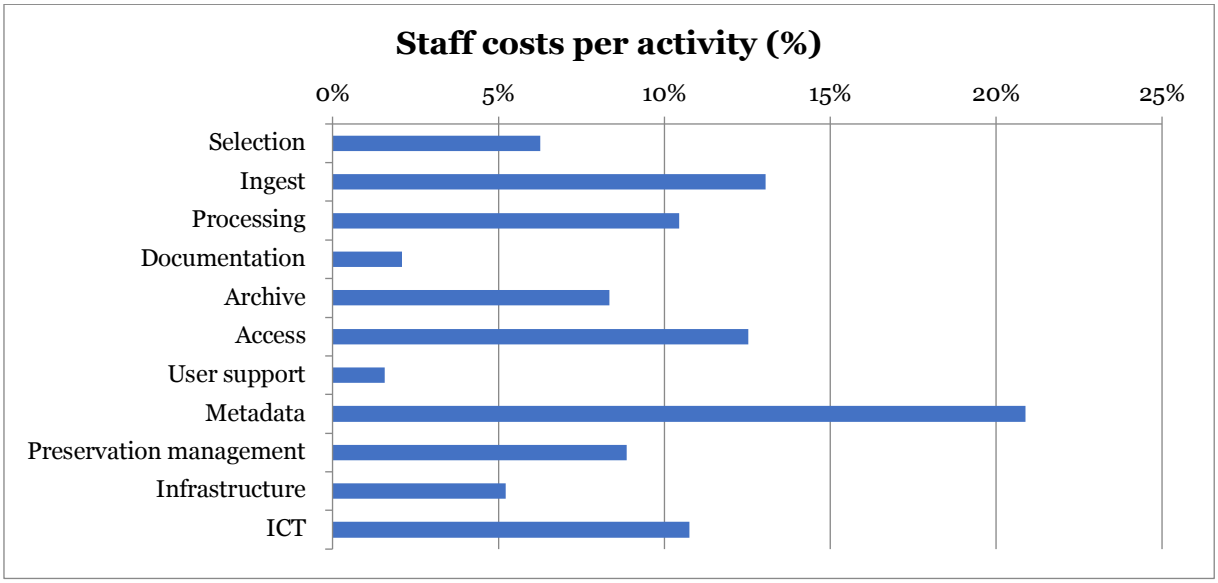

Figure 2: Allocation of Eye's staff costs to individual activities in the digital preservation cost model (as %)

Two activities in the digital preservation process account for a relatively large proportion of Eye's staff costs: ingest and access. Eye has made large investments in the front end of the process in order to maintain the highest possible standard of quality in curation and access to films. As a result, ingest is an activity that takes up a lot of time, given that various manual activities (taking up staff time) need to be performed at regular intervals in order to enrich the files. This also means that 'metadata', as a process activity, also absorbs more staff resources than the other activities. Eye spends a great deal of time and energy on metadating films, i.e. describing their contents, creative makers, production details and rights status for the purpose of access and curation. Where possible, technical metadata are added to the database by means of an automated process. Eye also invests in the back end of the process, i.e. in the cost of opening up access to digital data.

If material costs are also included, the following picture emerges of the aggregate staff and material costs (excluding external staff employed on temporary contracts, infrastructure, and ICT) in relation to each process stage and overarching activity at Eye:

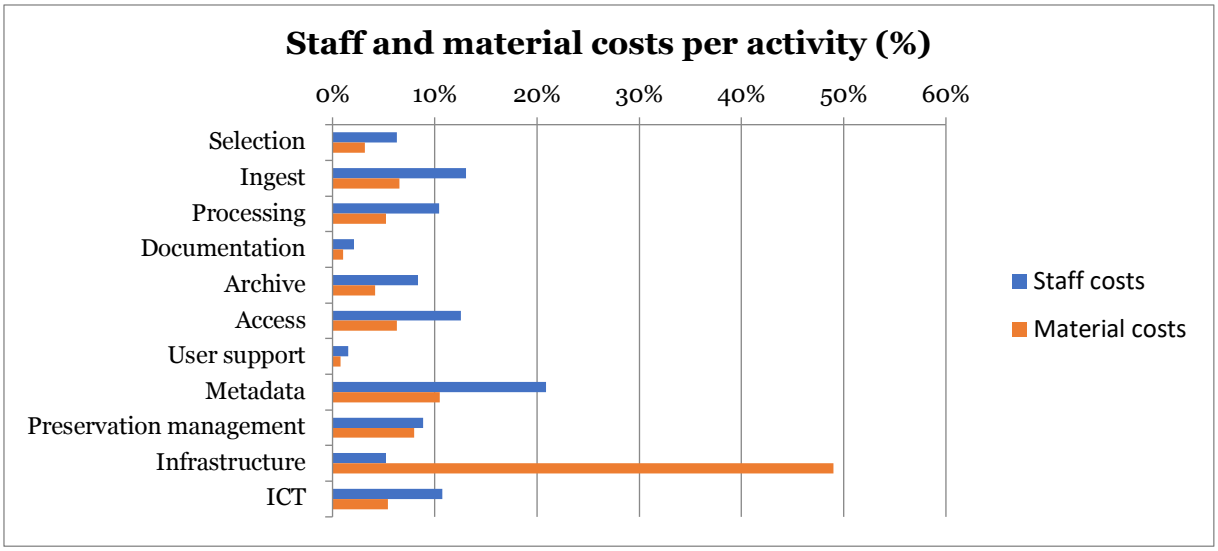

Figure 3: Aggregation of staff and material costs at Eye for each process stage and overarching activity in the digital preservation cost model (as %)

Figure 3 shows that the staff costs for the infrastructure are relatively low, but that the material costs are relatively high. The latter include hardware, operating system software, data transmission and software licences. This means that infrastructure and metadata are the two activities with the highest level of cost in the process of digital preservation. At Eye, metadata and the technical infrastructure account for the bulk of these costs. In other words, these are the processes that need to be optimised to gain better control over the costs in question.

The following figure shows the cost per object in relation to each process stage and overarching activity:

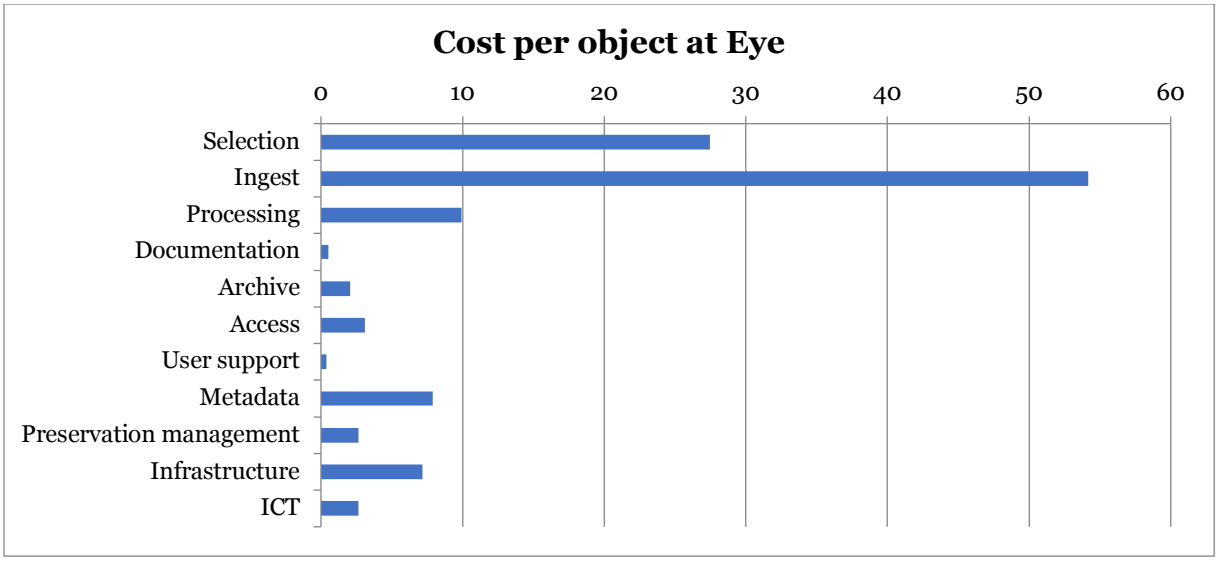

Figure 4: Cost per object in relation to each process stage and overarching activity

The above picture is due to the fact that the costs of the process stages starting from documentation apply to more than one object (to the entire collection, in fact). This is not the case with selection, ingest and processing. The three latter activities are used for additions to

the collection, which means that they are more costly on a per-object basis than the other activities in the digital preservation process.

In terms of the percentage distribution, 56.8% of Eye's costs relate to overarching process activities and 43.2% relate to individual stages in the digital preservation process. ICT (8.5%) and infrastructure (23.2%) account for 31.8% of the aggregate cost, with metadata accounting for a further 16.6%. Ingest is the most expensive of the individual process stages, accounting for 10.4%. The whole front end of the process, i.e. pre-ingest, selection and ingest, together accounts for 15.4% of the aggregate cost. The ratio of staff to material costs at Eye is 58.9% to 41.1%. Storage costs (in documentation and archive) absorb only a small proportion (8.3%) of the aggregate cost.

The cost model has given Eye a good idea of how the costs are distributed and has generated useful information for future decisions on the funding of activities. The model also allows Eye to take carefully considered strategic decisions on whether to perform certain activities itself, to subcontract them to external suppliers, or to perform them in partnership with third parties. This applies to storage, for example.

What next? Using a road map as a compass

The adoption of the prototype by the ten institutions represents the completion of the first stage in the development of the cost model for digital preservation.[12] We have now refined the model on the basis of the experience gained in the trial and on the basis of the resultant conclusions (see above).

The next step is to ensure that the model is used by more institutions as a means of analysing – and hence controlling – the cost of digital preservation. The more widespread use of the model will also give the Digital Heritage Network access to more information on the cost of digital preservation. This, in turn, will enable us to perform long-term analyses and exploit the model's full potential. As far as the latter point is concerned, linking costs with cost drivers is key here. In other words, what is the effect on cost when a button is pressed or a dial is turned (i.e. a policy is adopted or modified)? Information on the linkage between costs and cost drivers can be used as evidence in support of strategic decisions.

The following stage in the road map is to match supply and demand. The model provides users with information on the similarities and differences in cost structures in the various sectors, based on an analysis of each stage in the process. This information can be used as the basis for decisions as to whether institutions should perform certain activities themselves, subcontract them to external suppliers, or join forces with third parties, and also for aligning supply with demand. Finally, the information can also be used to challenge institutions and/or suppliers to come up with the right propositions so that they can solve certain issues,

[12] *NCDD Onderzoek naar de kosten digitale duurzaamheid* (BMC Onderzoek, January 2017) pp. 1-68. See in particular chapter 5, entitled 'Roadmap Dutch Cost Model for Digital Preservation', outlining a road map for the development of the cost model for digital preservation. The road map consists of the following four stages:

1. development of a Dutch cost model for digital preservation;
2. expanding the user base and corroborating findings;
3. matching supply and demand;
4. groundwork for social cost-benefit analyses.

irrespective of whether these apply specifically to their own sectors or more widely. The 4C road map[13] refers to this stage as 'Who should do what? – Market efficiencies'.

The final stage in the road map paves the way for undertaking social cost-benefit analyses and calculating the social return[14] of digital sustainability in various sectors. In other words, what is the added value for society and is it worth investing in this when the decision is taken? Clear information on costs and the underlying cost drivers forms the basis for social cost-benefit analyses. This fourth and final stage represents the 'dot on the horizon', i.e. the distant point that the Digital Heritage Network aspires to reach in the future, working in close collaboration with its affiliated institutions.

What does your institution stand to gain from the cost model?

Initial findings from the development of the cost model for digital preservation indicate that institutions stand to gain a great deal from analysing, controlling and actively managing the cost of digital preservation. The prototype model is a means of gaining a better understanding of, and hence achieving better control over, the costs. It is designed specifically to link up with the policy decisions taken by the institution in question (i.e. the cost drivers) on digital preservation. The model also makes it possible to compare institutions with each other and, in doing so, to identify similarities and differences in terms of cost drivers and costs. The resultant dialogue generates input for intensifying cooperation on digital preservation, both among institutions and within and between different sectors. The ultimate aim is to further improve the way in which the digital heritage is collected, curated and rendered accessible, and hence open it up to users.

One thing is clear (in addition to the need for further research): the importance of raising our understanding of the cost of digital preservation will only increase in the years to come.

[13] 4C *Investing in Curation; A Shared Path to Sustainability* (20 February 2015), pp. 1-26.
[14] The 'social return' is the added value held by a project for society as a whole. It is an indication of how much the project costs in relation to its social benefits (i.e. why should an organisation undertake such a project in the first place?). The social return can be used to inform strategic decisions, for example on whether an institution should perform certain activities itself, subcontract them to external suppliers, or join forces with third parties. A social cost-benefit analysis is a method of calculating the social return. It involves undertaking a systematic, coherent analysis of all the various effects caused by a project and comparing these with a scenario in which the project did not take place, i.e. a no-action alternative. As with the costs, the benefits of the project are expressed in euros. The same applies to aspects that are not immediately expressible in monetary terms, such as noise, a nice view or a sense of security. If the benefits are found to be greater than the costs, the project may be said to be conducive to social well-being. The weakness of a social cost-benefit analysis lies in its theoretical nature and the limited involvement of stakeholders. Moreover, in practice, such analyses tend to depend heavily on key indicators and theoretical assumptions, thus undermining the impact of their findings. Finally, they are labour-intensive and costly in their full-blown form (see also Ministry of the Interior and Kingdom Relations, *Werken aan maatschappelijk rendement; Een handreiking voor opdrachtgevers van MKBA's in het sociale domein* (October 2011) pp. 1-73).