The Digital Black Hole

Jonas Palm
Director, Head of Department of Preservation
Riksarkivet/National Archives
Stockholm, Sweden
Jonas.Palm@riksarkivet.ra.se

Digital projects may seem easy to plan and fun to conceive. The sky is the limit: the possibilities are seemingly endless and once material is digitized, its potential for use appears both exciting and cheap. It can't get much better, can it!

In a comic story from the 60's the Walt Disney cartoon character Gyro Gearloose invented a machine that could answer all the questions. Eventually Gyro gives up running the machine because he can't come up with enough questions to all the answers. This story can be used as an analogy for today's enthusiasm for digitzation projects. In the excitement about the solutions digitization offers, the right questions about costs are often not asked, especially about long-term costs for keeping the digital files alive. This enthusiastic attitude is risky, for the conversion process to create the digital files may well be quite expensive to start with, and these investments may turn out to be wasted if planning for the future is ignored and no structural funding for maintenance is secured.

Without such long-term planning, digitization projects can come to behave like black holes in the sky. Scanned information, which in the analog world could be accessed simply by the use of our eyes, is suddenly stored in an environment where it is only retrievable through the use of technology, which constitutes a constant cost factor. The more information is converted, the more the costs for accessing it go up. The digital black hole has got its firm grip on the project. It will go on swallowing either money or information: the funding must be continued or the input will have been wasted. If funding starts to fade, the information may still be retrieved but after a while it will no longer be accessible due to corrupted files, or obsolete file formats or technology. Then the digital information is lost for ever in the black hole.

The course of a typical digitization project may well be compared to the life cycle of a star. Stars are born and eventually they die. Fig. 1 shows this life cycle. The analogy becomes obvious when the phases in the life of a star are replaced by the stages in the life of the average digitization project (Fig.2)

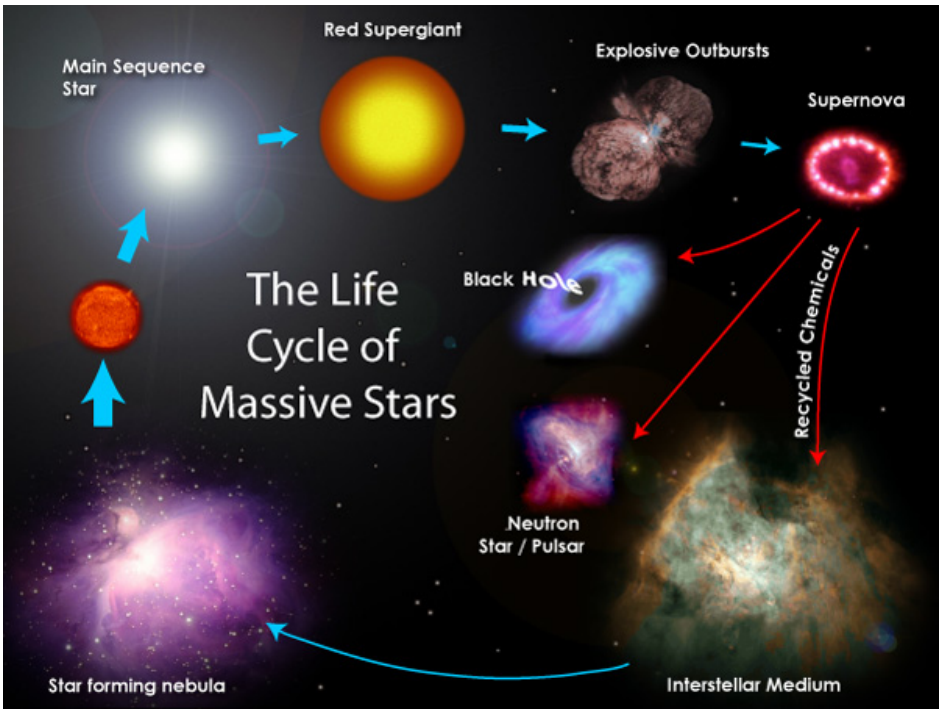

Fig. 1. The Life Cycle of Massive Stars (published on www.star.ucl.ac.uk/groups/hotstar/research.html).

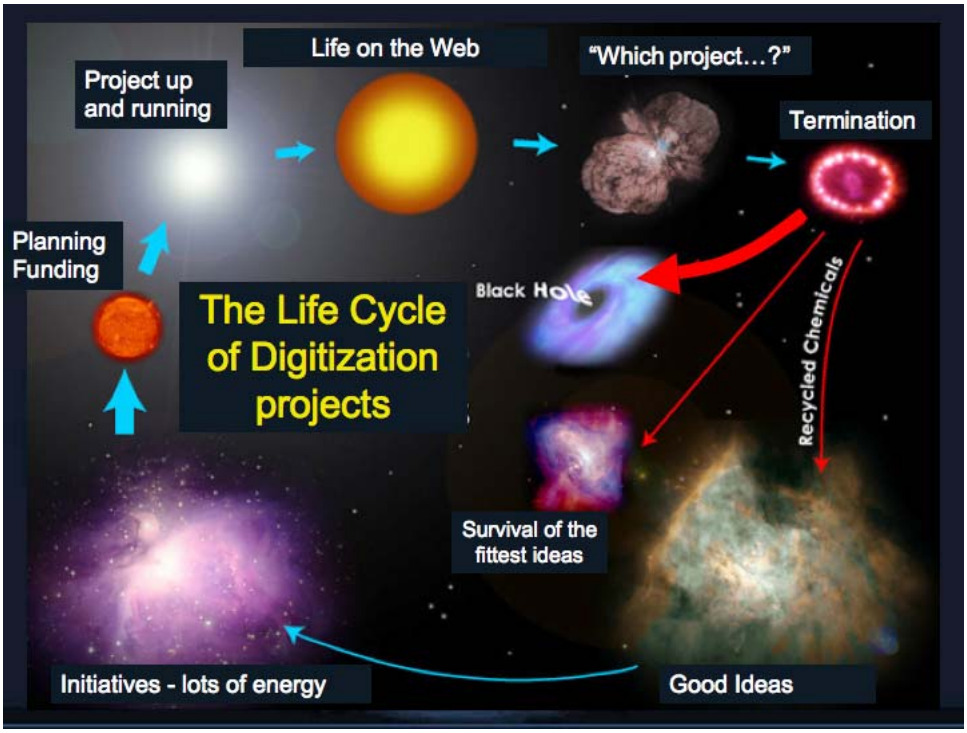

Fig. 2. The Life Cycle of Digitization Projects (modified from Fig. 1 by the author).

Good ideas are always around, like an interstellar medium. In the initial stages a lot of energy is drawn together and transferred into next phase, of planning and securing project funding. Then the project is underway, things begin to take shape, digitization is started. When all information is digitized and organized on a website it grows into a Super Giant, bright and strong and visible to the whole world. But then new projects begin to be

developed, other interests get in the way, our project begins to be neglected and starts to collapse. The organizers eventually decide to terminate it - and that's the end of another good project. Not everything necessarily dies with the project: the most important information may well survive, following a Darwinian process of the survival of the fittest, and some of the old good ideas will meet new good ideas to form new projects.

Whereas a cycle with a relatively short life expectancy may be perfectly acceptable for minor projects that are of interest only for a limited period of time, for larger projects it is too costly not to plan for a Life for the Files beyond the horizon. In such cases the choice between starting the project or not in fact depends on the willingness to plan for the future: a projectmay or may not be launched, but if it is , this decision entails a long-term financial commitment.

This article presents an analysis of costs for digitizing and long-term storage at the Riksarkivet (National Archives, RA) in Stockholm, Sweden. It is presented as an example as the actual costs may obviously differ between institutions and countries due to differences in costs for premises, salaries etc. Still, the model for the estimation of costs has a wider relevance and can be used to make similar calculations in other situations.

Costs of long-term storage

The National Archives (Riksarkivet, RA) in Stockholm have increasingly been receiving records in digital form since the 1970s; in 2005 they received about 25 Tb (terabytes). To be able to secure this data for future use and research the Riksarkivet invested in a major data storage system, an HSM-system (Hierarchical Storage Management System) two years ago. Such a system is based around a storage robot - in this case a tape cassette system - connected to servers and computers. The system is built to (a) detect and correct data errors in stored digital information and (b) be able to migrate the data to the next generation of mass storage system and so on. Data that is to be used is first copied from tape to a server, so that information in the storage robot is never actually used. The costs for such a system do not lie in the storage media (of which the cost amounts to around 5%-10% of the total) but in the rest of the system - hardware, software, support, maintenance and administration/operation.

RA receives basically two kinds of digital information - digitally born information and digital copies of traditional documents and records. The digitally born information iconsists of state agency records; the digital copies from records are from RA's collections in an effort to open up and improve access for anyone interested. The digitally born files are fairly small in size, since they mainly consist of databases. The files of digitized records, though, are almost exclusively image files and thus represent much more information and are eventually more expensive to handle. This digitization activity is a result of national effort to make governmental agencies to some extent available 24 hours a day.

Three years ago discussions started at Riksarkivet on the costs and problems of long-term storage of digital information. The question was whether, once materials were digitized, it was cheaper to maintain the digital files over time, or instead rely for long-term storage on images on microfilm produced from the digital files with the use of COM (Computer Output Microfilm). In both cases the originals would be kept as well. The starting point of the

discussion was that numerous files were produced in digitization projects that served various goals, but that it was not clear whether these files should then also be kept alive over a long period of time.

Two articles triggered this discussion. The first one was by Steven Puglia (National Archives and Records Administration) on the costs of digital imaging projects[1] The second was an article by Stephen Chapman (Weissman Preservation Center, Harvard University Library) on the costs of repository storage.[2]These articles showed clearly things were not as simple as many had thought. It was expensive to preserve digital files.

At the Riksarkivet we made calculations based on the costs of the hierarchical storage management system that we use for storage of digital information. When the outcome of our calculations was compared with Chapman's results, they coincided well, as can be seen in Fig. 3. In both cases the costs of storing the same amount of information were compared: an average book of 332 pages (1) in its original format stored in an air conditioned repository, (2) as a microfilm stored in a climate controlled vault , (3) digitized as 600-dpi bitonal images, and(4) digitized in 300 dpi images in grayscale (8-bit). The grayscale images take up more space and hence are more expensive to store, even though storage space by itself is now very cheap and not the main cost factor. Storage costs include the system needed to manage and preserve data, which covers checks on data integrity, backup procedures, checks for restoring information, automatic transfer to new tapes etc.

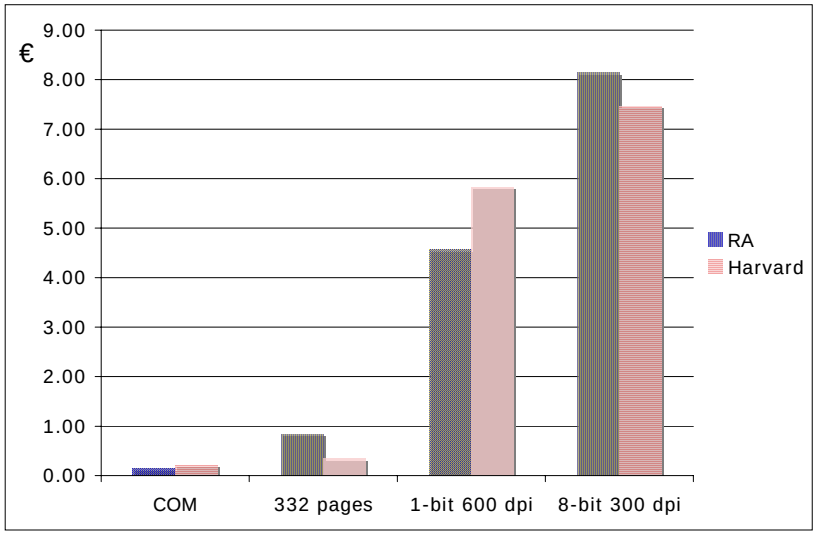

Fig. 3. Comparison of costs for storage by RA and Harvard University Library, of identical information in film format, original text on paper and two different digital file formats.

The costs of digital storage are much higher than generally believed because much more is involved than most people realize. In the discussions of these issues it has been

[1] Steve Puglia, 'The Costs of Digital imaging Projects,' in *RLG News*, Oct. 1999.

[2] Stephen Chapman , 'Counting the Costs of Digital Preservation: Is Repository Storage Affordable?' *Journal of Digital Information*, Vol. 4(2), Article No. 178, May 2003.

suggested from within the industry that as storage increases, the economic load increases faster. The fact that the capacity of storage media doubles each year results in the misconception that prices of storage are rapidly decreasing. For the short term - typically less than five years - this is true since not much has to be done to keep files accessible, but over the long term the costs of management will keep going up. Jim Gray, head of Microsoft's Bay Area Research Center put it like this:

> …But the real cost of storage is management. Folks on Wall Street tell me that they spend $300,000 per terabyte per year administering their storage. They have more than one data administrator per terabyte. Other shops report one admin per 10 TB, and Google and the Internet Archive seem to be operating at one per 100 TB. The cost of backup/restore, archive, reorganize, growth, and capacity management seems to dwarf the cost of the iron. This stands as a real challenge to the software folks. If it is business as usual, then a petabyte store needs 1,000 storage admins.[3]

In general costs for hardware are still decreasing, and storage media are now so cheap they may be of very little importance in the overall discussion. There is a difference though between storage media costs and computer costs (Fig. 4 and 5). While the price of computers in terms of capacity has dropped considerably, at the same time the amount of data that computers deal with and hence the capacity required for processing files has increased a lot. This is not necessarily a matter of dealing with more information –it often means just handling more options. This becomes obvious if one compares the cost of a single 2Tb hard disk drive – 450 euro,- - with the cost of a typical 2Tb back-up hardware system which may cost from 10 times as much – 4500 euros and upwards. As with the HSM-system the major cost is not the storage media but the surrounding hardware and software.

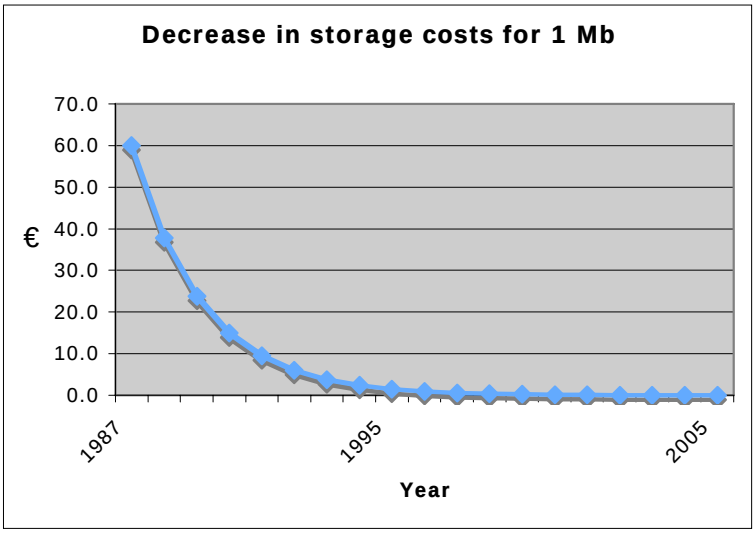

Fig. 4. Decrease in storage costs for 1 Mb information on magnetic storage media.

[3] Interview in *ACM Queue* Vol. 1 (4), June 2003.

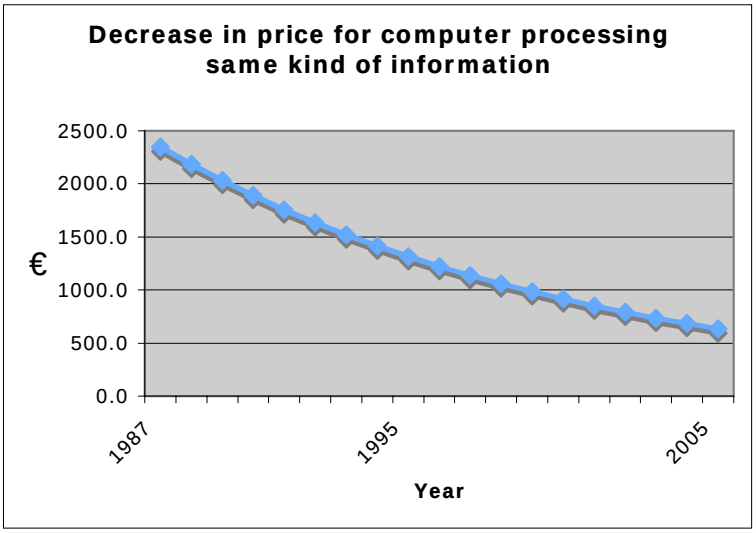

Fig. 5. Decrease in price for computer, processing same kind of information.

Large-scale systems for long-term storage most likely are subject to another pattern of price development. It is generally accepted these systems have a lifespan of approximately 5 years. When a system is new the price is at its peak. It will decrease until the next generation is introduced. Then a new peak in prices will occur: prices will go up again, though not to the same level as at the beginning of the cycle. In our calculations we assumed there will be a slight decrease in price between each generation of about 25% (Fig. 6). This is *just an assumption, as is any attempt to foresee the future of digital storage more than 5-10 years in advance.* Still, assumptions like this were made in order to get an idea of the economic conditions in a future situation.

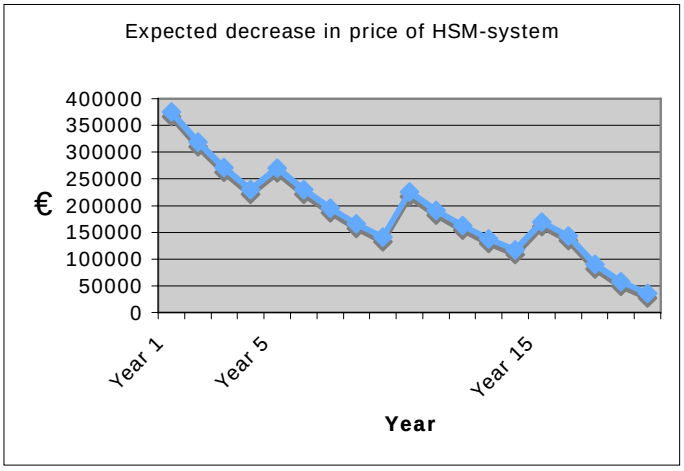

Fig. 6. Expected decrease in price over a longer period for large-scale storage systems.

The HSM-system at RA, with a tape cassette robot for long-term storage of digital records, has a capacity when fully used of 200Tb, and is set up to be able to expand with 40Tb/year. (As mentioned above the increase so far has turned out to be only 25Tb a year.) It was installed in 2003 and has been running for about 18 months. The costs for the system itself and for operating it are shown in Fig. 7.

Specification of costs	1st year	2nd year	3rd year	4th year	5th year	5 years
1 HSM storage system						
price 2003 + 3% interest/year						
€406,643 spread over 5 years, 81,328/year	94818	92379	89939	87499	85059	**449694**
Staff for operation, 0.6 fte, €40,000/fte incl all costs	24000	25200	26400	27660	28980	**132240**
Premises 100m^2, €126 per m^2	12600	12915	13237	13568	13908	**66228**
Service/support	22700	28900	28900	28900	28900	**138300**
Total storage costs	154118	159394	158476	157627	156847	**786462**
Annual storage cost per Gb	3.85	1.99	1.32	0.98	0.78	
Average storage cost per Gb for 5 years						**7.86**
Storage medium 40 Tb/yr	17930	11295	7116	4483	2824	**43648**
Staff for input, 0.4 fte €40,000/fte incl all costs	16000	16800	17600	18440	19320	**88160**
Yearly Input Cost (staff, storage medium)	33930	28095	24716	22923	22144	**131808**
Cost of input per Gb	0.84	0.7	0.61	0.57	0.55	**0.66**
Total cost per newly added Gb	4.69	2.69	1.93	1.55	1.33	
Average total cost per GB for 5 years						**9.18**

Fig. 7. Costs for HSM storage system at RA, Stockholm, Sweden. Costs are quoted in € (EUR).

Figs 8 and 9 show how costs for equipment will decrease while the costs for salaries and premises will rise. Normally, costs for support and updates would have shown a rising curve as well, but according to the contract between RA and the vendor, the cost is evened out in the contract over 5 years.

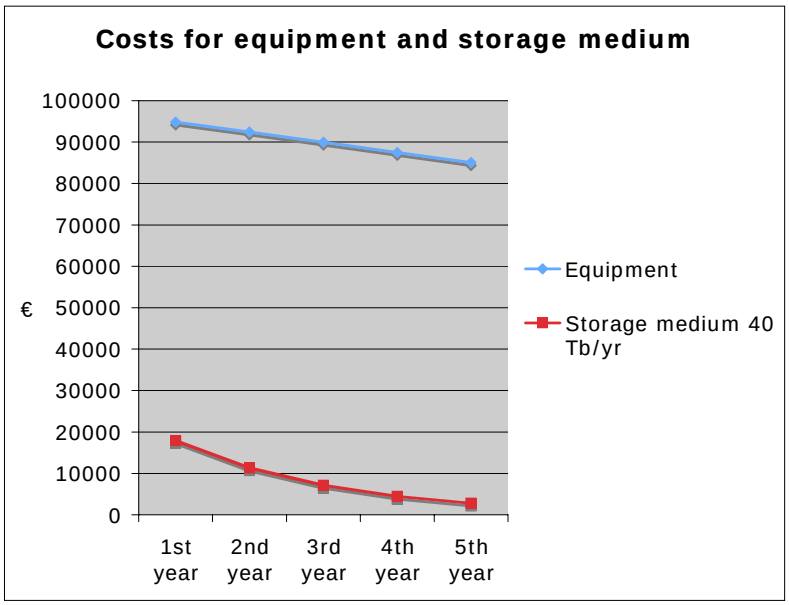

Fig. 8. Costs for hardware for RA's HSM storage system.

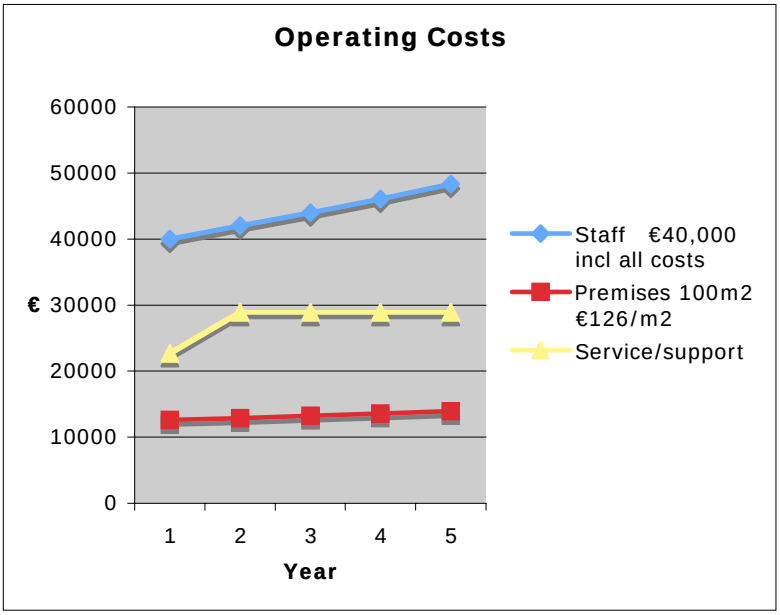

Fig. 9. Operating costs for RA's HSM storage system.

When the costs are divided into technology, staff and premises (Fig. 10) it turns out that the cost of labor accounts for 39% of the total. This will increase in the years to come, as salaries will go up and as more staff will be needed to manage the system as it grows. Not all staff would have to be highly qualified, but as salaries in Sweden are not as differentiated as in some other countries, this will not make much difference to the calculations.

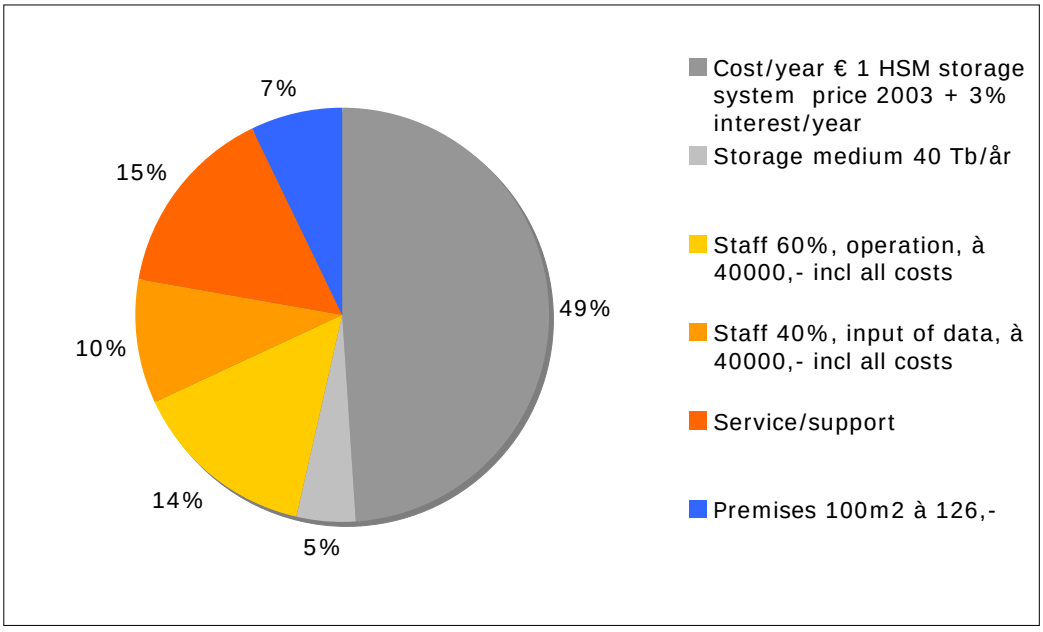

Fig. 10. Division of costs of RA's HSM storage system.

If one attempts the impossible, maybe even the ridiculous, and tries to make forecasts beyond a period of 10 years, the only thing which is certain is that salaries will escalate as well as the general cost index. The calculation made for the Riksarkivet is based on the

assumption that the vast majority of the digital information stored will be passive. Costs for personnel are related to access activity, and the situation at RA requires only limited staff for keeping the system running as opposed to companies, banks and Google (the examples mentioned by Jim Gray of MicroSoft above). Even so, costs for personnel and premises at RA will continue to rise, and Fig. 11 shows that total costs of staff, support and premises will exceed equipment cost by more than 12 times in 30 years' time. The costs for storage medium is barely visible in this graph and then only the first decade.

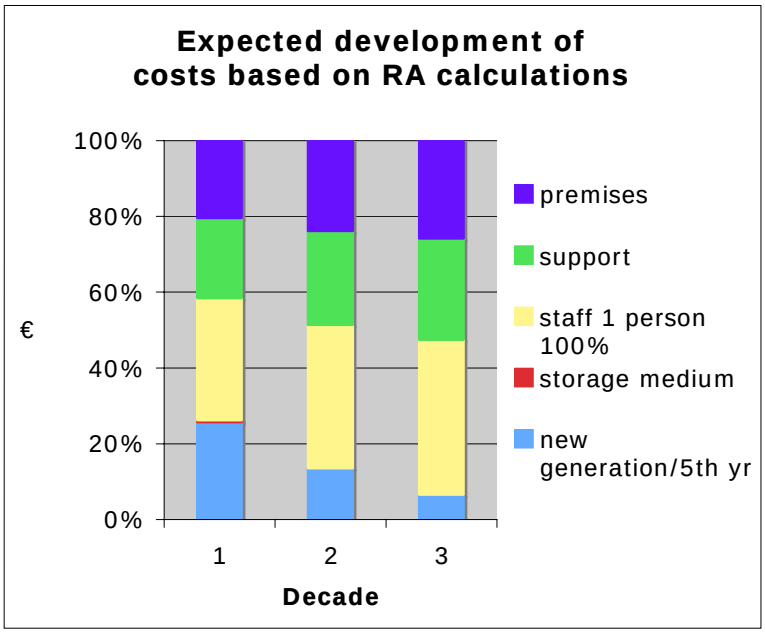

Fig. 11. Expected costs development at RA.

The costs of long-term storage are dependent on the rate of activity: the more the information stored is used, the higher the administrative costs. When use of the information goes up, there is also more need for external servers from which information is accessed. If one includes in the calculations made so far a factor for costs associated with a future rate of activity as at RA in Stockholm the following picture merges.

When the negative economics of scale are taken into account and as well as increased activity of the stored information, the costs of staff for operating and managing the system most probably will increase at RA to be at many times the cost for equipment. With growing staff, costs for housing and hence for premises will also rise. As it is hard to predict development of support costs they have been kept constant, but most probably they will increase with the size of the system. (Fig. 12)

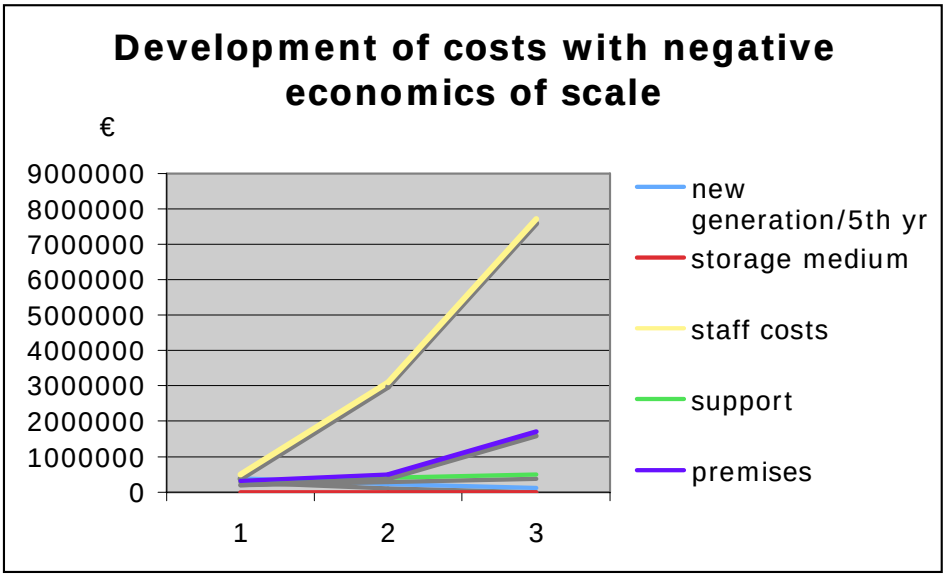

Fig. 12. Probable effects of negative economics of scale at RA in the long run.

Digitization

The cost scenarios for storage may give some food for thought. As they show the level of long-term financial commitment that is required to store digital files –let alone keep them alive through a constant process of refreshment and migration- they underline the need for careful consideration of the initial investment of digitization. The risk that materials end up in the grasp of the digital black hole is all the more serious because the costs of digitization itself are high as well. Digitization includes different activities, such as selection, creation of descriptions and metadata, project management, and the actual conversion (scanning or capturing with a digital camera). The costs of the actual scanning are by now reasonably well known.

Scanning quality is dependent on equipment, process specifications, knowledge of material to be scanned, and handling. Choice of equipment is related to the material to be scanned, specifications are related to properties and quality of the original information, and knowledge of the material to be digitized is essential in quality control, handling and setting up a responsible workflow. In digitization of images and sound expertise on content and carriers is an absolute necessity to ensure optimal capturing of information contained in the originals.

For digitization of paper materials, some cost calculations were made at the Riksarkivet in 2005. The Riksarkivet has its own scanning facility, MKC (Medie konverterings centrum, Media Conversion Centre) with around 80 employees in 2005. The objects scanned are records, bound and in sheets, and large-format maps and drawings. All figures below are based on their information.

At MKC 5 million images are scanned each year as 1-bit 600 dpi files in A4 format. The costs for each scanned file is approximately 0.10 euro. The records are scanned in an automatic feed scanner, The distribution of costs for creating a digital image file can be seen in Fig. 13. A third of the cost goes to scanning while preparation, quality control, extras and administration all four account for an equal, major part of the cost.

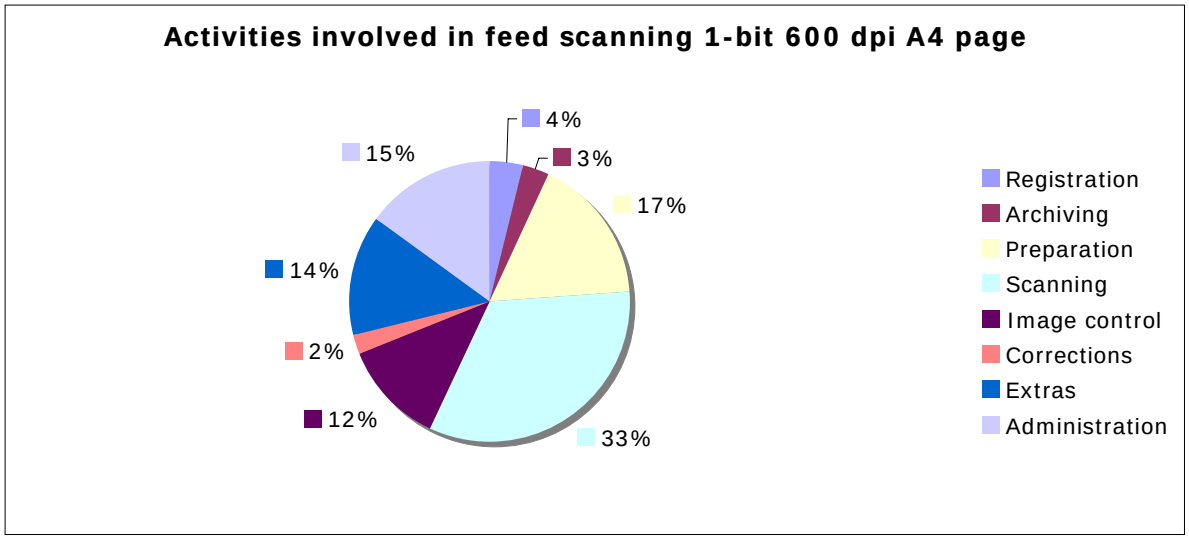

Fig. 13. Distribution of costs at RA's scanning facility MKC, Fränsta, Sweden.

Scanning of large-format drawings and maps is done at in 8-bit grey-scale at 297 dpi, in manually fed scanners. The cost for creating each file is approximately 0.61 euros, with 1,321,000 image files created each year. The costs for this kind of scanned file are distributed as shown in Fig. 14. Here the cost of the scanning itself accounts for almost twice as big a share of the total (65%). Administration is the second largest cost factor, while the rest are more or less equally distributed.

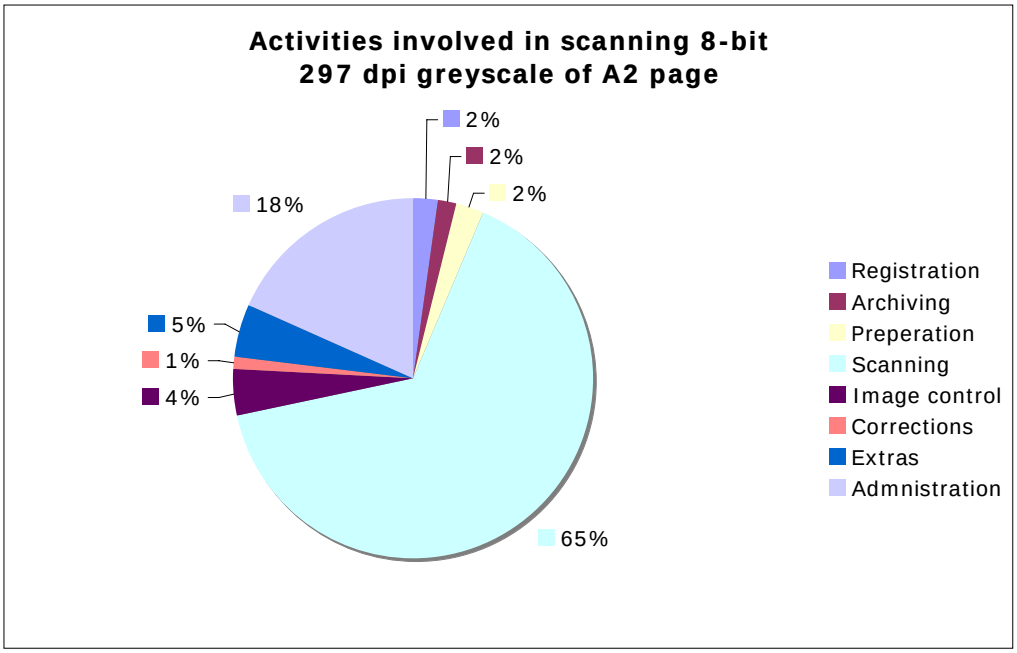

Fig. 14. Distribution of costs at RA's scanning facility MKC, Fränsta, Sweden.

When it comes to digitizing audiovisual information, it is quite a different story. This is both very time consuming and creates huge amounts of digital information. It is also the only case in which there is no other possibility to preserve materials for the future than by

digitization. In other words, whereas for preservation audiovisual materials must be digitized, the result will be enormous quantities of digital data that must be stored and preserved for the long term.

In 2004 the Swedish Ministry of Culture published the report *Preserving sounds and images*[4] which discusses strategies for the preservation of the collections of the State Audiovisual Archives (Statens Ljud och Bild Arkiv, SLBA). The collections consist of 4,5 million hours of audio and video, 30% audio tapes and 70% videotapes. If this was to be digitized in slightly 'compressed' or restricted formats - considered to be of a sort of minimal quality by many - like CD (16bit 44100 khz sampling rate) and DVD (MPEG 2), in total this would amount to 8 Petabyte (i.e. 8,000,000 GB). If 'true quality', the state of the art at a given time, was to be achieved, the amount of data would be even higher. And since technology improves quickly in these issues the definition of true quality is fluent, to say the least. Since the collections include many different formats and types of recordings, different digitization processes couldl be used. For instance, it is suggested that 1/4" tapes with speech could be digitized at twice the original speed of recording. For materials of this kind this would ensure sufficient quality; as it concerns a large number of these tapes the savings in time would be considerable. Even at its most efficient, however, the entire operation is estimated to take 10 years at a cost of 90 million euros.

The report states that 'Due to condition and technical circumstances transfer should be made within the next ten years'. This kind of material must be digitized in the near future to be preserved, since the original media are deteriorating continuously and it, as well as equipment, becomes obsolete and difficult to maintain in working order.

Distribution of costs are not specified in detail in the report. Most of the costs will go to the conversion itself, since many machines could be set up to run simultaneously operated by only a few staff. Preparation and extras would probably be the second largest cost factors. With audiovisual material one has to take into account the costs of maintenance of the analogue equipment and of adjusting it for optimal signal extraction. This is specialist work that can be time consuming. A comparison of production costs per Gb for AV compared to other materials is presented in Fig. 15.

[4] *Bevara ljud och rörlig bild* (SOU 2004:53), Swedish Ministry of Culture, 2004

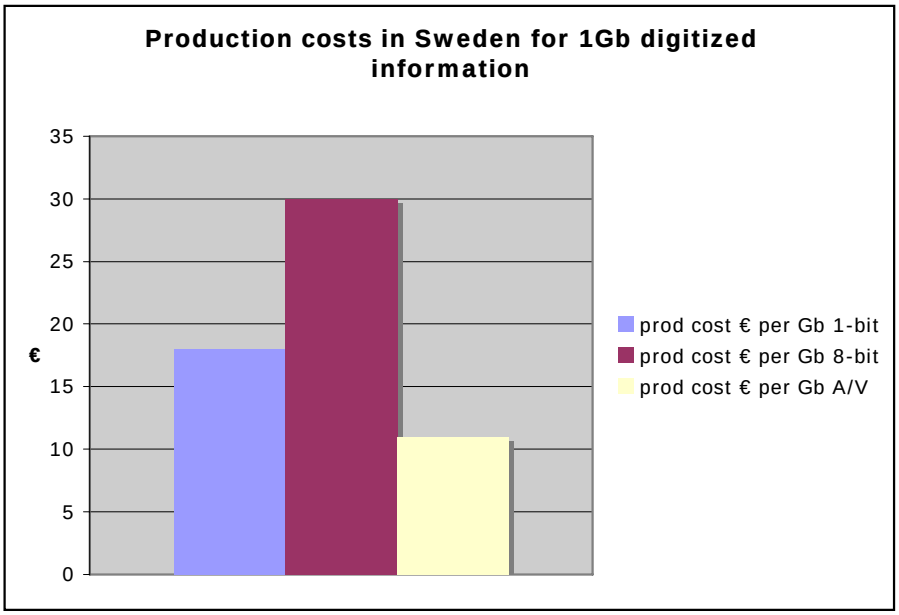

Fig. 15. Production costs for different file formats - 1 Gb 1-bit and 8-bit image files and audio visual files

Conversion of all this material would probably result in the annual production of of around 800,000Gb of digital information. Fig. 16 compares the amount of image files produced per year at MKC with the estimated annual production of audiovisual files. This huge investment in digitizing has to be matched with adequate arrangements for preserving the work done for the long term.

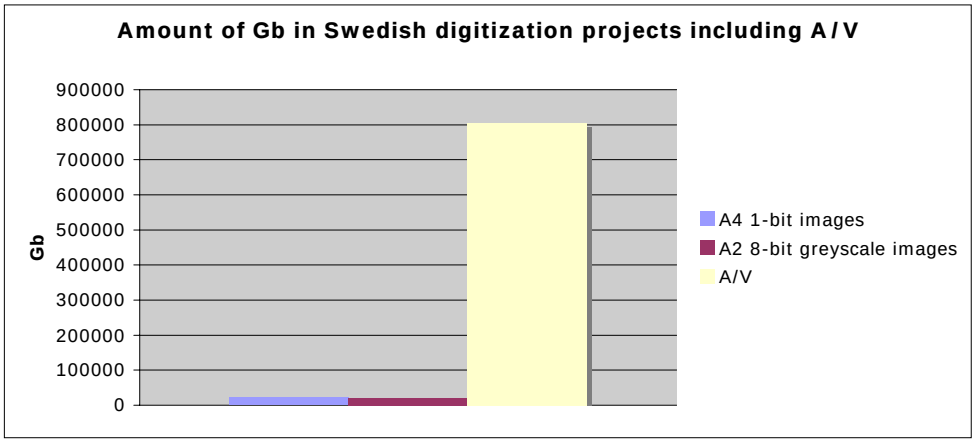

Fig. 16. Comparison of the annual amount of image and audio visual files to be created in some Swedish digitization projects at RA/MKC and for the SLBA in coming years

However, when one considers the costs of storage alone, it is clear this requires structural funding at a level that at present one can hardly expect to be forthcoming.

For audiovisual materials, this creates a quandary: as there is no other choice but to convert to digital format and to preserve the digital files, provisions for conversion and long-term storage will somehow have to be made, while it is doubtful whether funding for such large projects and continued maintenance can be secured. For paper materials, there is still a possibility of digitizing for access and to rely for preservation on the originals or on

microfilm. The decision to create digital images, use them to produce COM (which then becomes the preservation format) but not make the commitment to keep the digital files for the future may in financial terms be a sensible one. Digital collections may fulfill a certain role only for a limited period and there is not always a need to keep them forever, especially as they can relatively cheaply be rescanned from COM, should the need arise.

RA is currently studying whether it is feasible to use COM in an effort to improve the strategy of microfilming, which has a long record for securing information on materials in bad condition. Instead of just microfilming RA is considering to transfer image files to COM together with metadata for searching. (If one moves in the other direction and first produces microfilm from which image files are made, the microfilms lack these data for searching). The digital images can be used directly, but with the COM there is not the same necessity to preserve them as would otherwise be the case. In the future the films could be (re)scanned very quickly and be available in the digital domain as well as searchable

Whatever strategy one chooses to follow, the essential point to consider before undertaking large-scale digitization is the level of long-term financial commitment that can realistically be secured and to develop a preservation strategy accordingly. Estimations of costs that cover all aspects should be part of the planning process to limit the risk that a project ends up as yet another digital black hole, as so many others have done.