デジタルブラックホール
The Digital Black Hole

スウェーデン国立公文書館
理事、プリザベーション部長
Jonas Palm

（訳：楢林幸一）

デジタルプロジェクトは計画しやすく、考えるのも楽しいかもしれない。デジタルは無限の可能性を秘めているように思える。資料は一度デジタル化されると、使い方の潜在力はエキサイティングでコストもかからないように見える。これ以上のものは無い！と思える。

1960年代のウオルトディズニー漫画に出てくる、ジャイロ・ギヤルースはどんな質問にも答えられる機械を発明したが、実際に使うのは諦めた。全ての答えを引き出すだけの質問が思い浮かばなかったからというのがその理由だった。この話は現在のデジタル化プロジェクトに対する熱意に重ねてみることができる。

デジタルが提供してくれるソリューションが素晴らしいために、コストについての当然の質問がなされないことがよくある。特に、長期間デジタルファイルを維持するコストについて、その傾向が見られる。デジタルへの過剰な熱意は危険が伴う。デジタル変換プロセスを開始するにはかなりの初期投資が必要である。その投資は、デジタルのメンテナンス経費を継続的に保証する構造的な資金調達手段が無かったり、将来計画が廃止されたりした場合は無駄な投資になってしまう。

このような長期計画の無いデジタル化プロジェクトは宇宙のブラックホールのようなものである。目で見れば読める単純なアナログの世界に比べて、スキャンされた情報は、技術を使わない限り検索できないという環境に置かれる。技術を使わないと検索できない環境を維持するにはコストがかかる。情報が電子化されればされるほどアクセスコストも膨らむ。デジタルブラックホールがそのプロジェクトを掴んで、金と情報を飲み込んでいくことになる。投資は継続しなければならない。さもなければ、電子化は無駄になってしまう。投資が先細りになってくると、情報はしばらくはアクセスできるとしても、いつかはファイルが壊れたり、ファイルフォーマットや技術が陳腐化してアクセスできなくなる。こうして、デジタル情報は永久にブラックホールの中に消え去ることになる。

デジタル化プロジェクトの進行は星のライフサイクルにも喩えられる。星は生まれ、そして死んで行く。 図1は星のライフサイクルを示している。星の一生の各段階を、図2のデジタル化の各段階に置き換えると、とてもよく似ていることわかる。

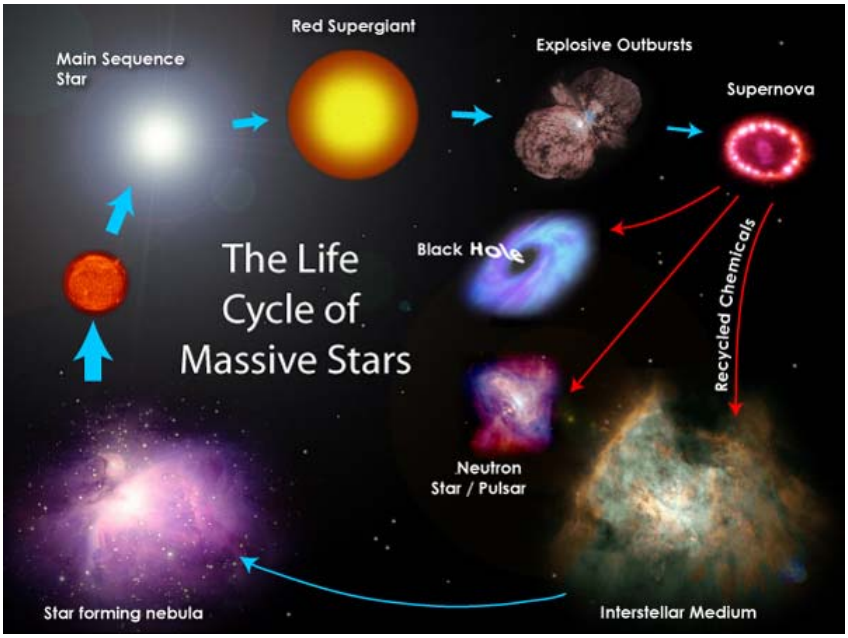

Fig. 1. The Life Cycle of Massive Stars (published on www.star.ucl.ac.uk/groups/hotstar/research.html).

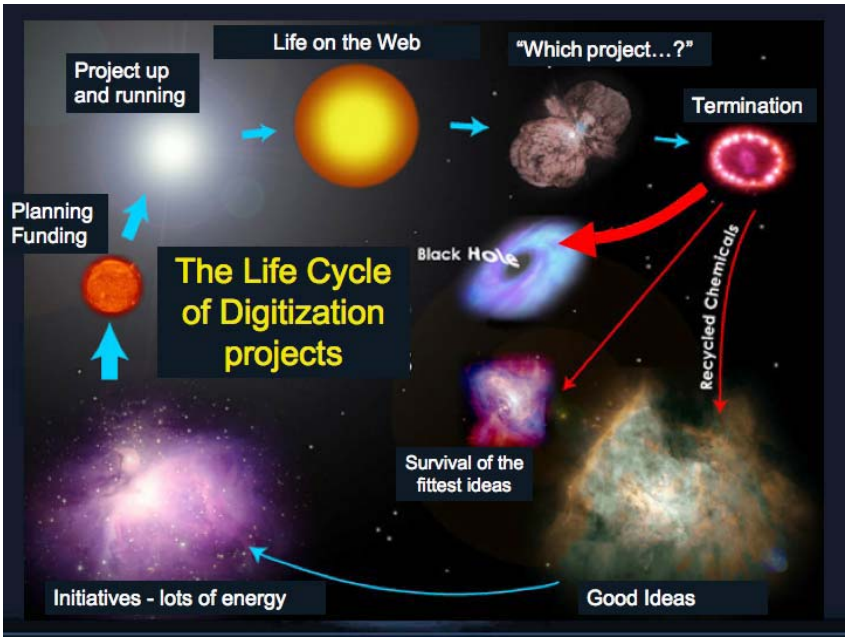

Fig. 2. The Life Cycle of Digitization Projects (modified from Fig. 1 by the author).

良いアイデアは星間媒体のように、常に周りに存在する。アイデアの初期段階は膨大なエネルギーが集まって、プランニングとプロジェクト資金の確保という、次の段階に受け継がれる。次にプロジェクトが動き出して、計画は形になってきて、デジタル化が始まる。全ての情報がデジタル化され、ウェブサイトに 掲載されたときに、その姿は世界に見えるようになり、光り輝く強力なスーパージャイアントへと成長する。

しかし、別の新しいプロジェクトの開発が始まると、その他の対象は邪魔になる。やがて我々のプロジェクトは放置され、崩壊を始め、組織はその中止を決定する。これが優れたプロジェクトの末路である。しかし、重要な情報の全てが死ぬわけではない。ダーウィンの進化論に従えば、もっとも重要な情報は生き残るだろう。そして、その他の古き良きアイデアは、新しいプロジェクトを形成する、新しく良いアイデアに取り入れられるだろう。

電子化は期待寿命が比較的短い電子記録のサイクルを考えると、短期間だけ重要性のある小規模なプロジェクトには最適かもしれない。地平線の彼方までファイル寿命があるような大規模なプロジェクトにはコストがかかり過ぎるかもしれない。そのような案件で、プロジェクトを開始するか、しないかの選択は将来を見通して計画する意思の強さにかかってくる。あるプロジェクトを開始するか、しないかの局面で、開始を決めたとしても、その決断は長期の財政的責任を伴う。

この報告書はスウェーデン国立公文書館（Riksarkivet, RA） における、デジタル化と長期保存のための実践コストを分析したものです。給与等のコスト計算の前提条件が国や組織によって異なるため、この報告書のコストは一つの例として読んでください。そうは言っても、コスト評価のモデルは全般的に適切で、他の状況で同様の計算を行う場合にも使えると思います。

長期保存のコスト

ストックホルムにある国立公文書館（以下、公文書館）では1970年代からデジタル形態での記録の受け入れが増えてきた。2005年には25テラバイトを受け入れた。公文書館はこのデータを将来に渡って使えるようにするため、２年前に大規模なデータストレージシステム（HSMシステム：階層型記憶装置管理システム）を導入した。このシステムはサーバと、コンピュータに接続されたストレージロボット（この場合は、カセットテープシステム）を中心に構成されている。システムは①保管されたデジタル情報のデータエラーを検知して修正する、②データを次世代の大容量記憶システムにマイグレーションする、機能を持っている。使用されるデータは最初にテープからサーバにコピーされるため、ストレージロボット内の情報が実際に使われることはない。システムのコストは、全体の5%から10%に該当するストレージ媒体にではなく、ハードウェア、ソフトウェア、サポート、メンテナンス、管理／オペレーションにかかる。

公文書館は基本的に2種類のデジタル情報を受け入れている。一つは最初からデジタルで作成（デジタルボーン）された情報で、もう一つは紙文書をデジタル化したものである。デジタルボーンの情報は政府機関の記録で、紙文書のデジタルコピーは、閲覧者のアクセス性を改善するために、公文書館の収納文書をデジタル化したものである。デジタルボーンのファイルはデータベースにあるため、容量は小さい。しかし、デジタル化された記録のほとんどはイメージファイルであるため、情報量は多くなり、結果としてハンドリングコストが高くなる。デジタル化の活動は、政府機関に情報を1日24時間提供できるようにするという国の政策によって推進された。

デジタル情報を長期保存するコストと課題についての議論が3年前に始まった。議論のテーマは、資料をデジタル化した後、デジタルファイルとして維持していく方が安いのか、あるいは、COM（コンピュータ　アウトプット　マイクロフィルム）を使って、デジタル

ファイルからマイクロフィルムを作成して、イメージとして長期保管するのが安いのかを検討することだった。どちらの場合でも、原本は保存される。

議論の入り口は、デジタル化プロジェクトの中で、大量のファイルが作成され、様々な目的に使われたが、それらのファイルを長期に渡って維持し続けていくべきか、否か、が明確になっていないことについてであった。二つの論文がこの議論の引き金になった。一つはデジタルイメージングプロジェクトのコスト(注1)に関する、米国立公文書館のスティーブン・プーリアの論文で、一つはレポジトリストレージのコスト(注2)に関するハーバード大学図書館、ワイスマンプリザベーションセンターのステファン・チャップマンの論文である。二つの論文は、多くの人が考えるほどデジタルのコストは単純ではないということを明確に示している。デジタルファイルの保存コストは高いということである。我々はデジタル情報の保存に使っている公文書館の階層型記憶装置管理システムのコストをベースにして計算を行った。計算結果をチャップマンの結果と比較したところ、図3の通り、一致した。両方のケース共、同量の情報を保存するコストを以下の条件で比較した。平均的な332頁の本で、（１）原本を空調設備のある書庫で保管、（２）マイクロフィルムをコントロールされた環境の書庫で保管、（３）600dpiの白黒イメージで保管、（４）300dpiのグレースケール（8ビット）イメージで保管。グレースケールイメージはストレージスペースが非常に安くなっていて、主要なコスト要因では無くなっているとはいえ、より多くのスペースをとるため、保管コストも高くなる。ストレージコストの中には、データの完全性チェック、バックアップ手順、情報復旧のためのチェック、新しいテープへの自動書き換え等をカバーする、データを管理し保存するのに必要なシステムが含まれる。

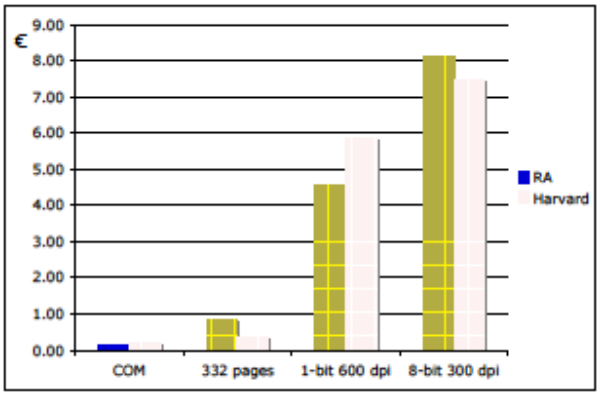

図3. 公文書館とハーバード大学図書館による同一情報を、フィルム、紙、二種類のデジタルフォーマットでストレージした場合のコスト比較。

デジタルストレージのコストは一般的に信じられているよりも高くなる。理由は一般の人が考えているよりも、多くの要因が関係するためである。これらの課題を討議する中で、業界からは ストレージが増えると、経済的な負荷も、より早く増加するという話を聞かされた。

(注1) Steve Puglia, 'The Costs of Digital Imaging Project,' in RLG News, Oct. 1999.
(注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.

毎年、ストレージメディアの容量が倍増しているため、ストレージのコストは急速に下がっているという誤解を生んでいる。これは短期的（一般的に5年以内）保存の場合は、ファイルをアクセス可能状態にしておくために実行すべきことが少ないので事実である。しかし、長期に渡って管理するためのコストは上がり続けるだろう。

マイクロソフトのベイエリアリサーチセンターの、ジム・グレイ所長は次のように語っている。

> 「…しかし、ストレージの本当のコストは管理費です。ウオール街の人達は、ストレージの管理費としてテラバイト当り、年間30万ドル（約3,450万円）を費やしていると私に話しています。彼らは、テラバイト当り、1人以上のデータ管理者を抱えています。別の情報源は10テラバイト当り１人と言っています。Googleとインターネットアーカイブは100テラバイト当り、１人で管理しているようです。バックアップ／復旧、アーカイブ、再編成、拡張、容量管理のコストは強固なコストを過少に見せているようです。これはソフトウェア関係者にとっての課題になっています。事業が普通に続くとすれば、ペタバイトの保管にはストレージ管理者が千人必要なことになります。」（注3）

一般的にハードウェアコストは下がり続けている。今や、ストレージメディアは非常に安くなり、コストについての全体的な討議の中での重要度は低くなっているかもしれないが、メディアとコンピュータのコスト（図4と図5）の間には違いがある。容量の点から見ると、コンピュータの価格は目に見えて下がってきた。同時に、コンピュータが扱うデータ量とそれに伴う、ファイル処理に必要な容量は非常に大きくなっている。これは必ずしも、より多くの情報を扱っているという問題ではなく、より多くのオプションを扱っていることを意味している。これは1台の2テラバイトハードディスク（450ユーロ）を、その10倍（4,500ユーロ以上）はする一般的な2テラバイトバックアップ ハードウェアシステムと比較すれば明らかである。HSMシステムの主要なコストはストレージメディアではなく、それを取り巻くハードウェアとソフトウェアである。

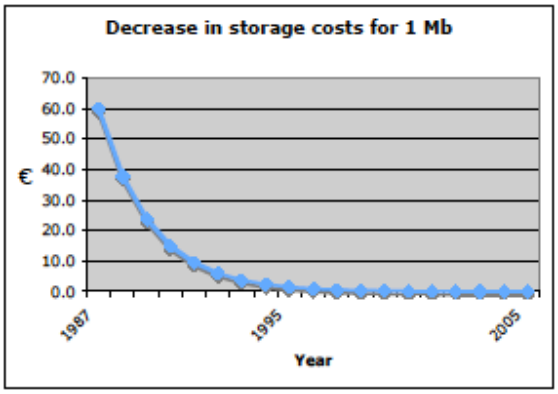

図４．１メガバイトの情報を保存する場合の磁気ストレージメディアのコスト傾向。

（注3）Interview in ACM Queue Vol. 1(4), June 200

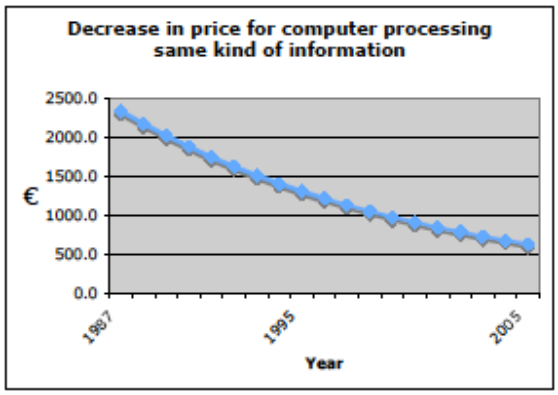

図５．同じ情報を処理する場合のコンピュータ価格の減少

長期保存用の大規模システムは違う価格展開パターンになっているように思える。一般的に、これらのシステムの製品寿命は5年程度であると考えられている。システムが新しい時が、価格がピークの時である。価格は次世代の製品が発売されるまで減少していく。それから、次世代製品で新しいピークが起こり、価格は再び減少する。しかし、価格サイクルの最初の時と同じ水準ではない。我々の計算では、各世代間の価格は25%程度減少すると推測している（図6）。これは、デジタルストレージの5年から10年先の将来を予測するための単なる仮定であるが、それにしても、このような予測は将来の経済的な条件についてのアイデアを得るために役立つであろう。

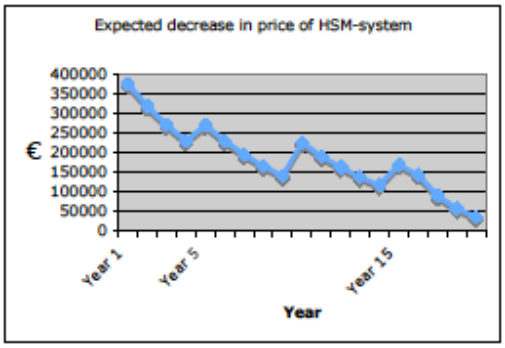

図６．大規模ストレージシステムの長期的な価格の減少予測。

デジタル記録の長期保存用の、カセットテープロボットを搭載した公文書館のHSMシステムは200テラバイトの容量を持っており、年間40テラバイトの拡張ができるように設定されている（前に述べたように、実際は年25テラバイト増加している。）。システムは2003年に導入され、約18ヶ月稼動している。システムとオペレーションのコストを図7に示す。

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	27600	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**

図7．公文書館のHSMストレージシステムのコスト。（単位はユーロ）

図8と9は給与と施設のコストが上昇する中で、機器とストレージメディアのコストがどう減少するのかを示している。通常、サポートとアップデートのコストは上昇曲線を示すが、公文書館と業者の契約に従えば、コストは5年間の契約で平均化される。

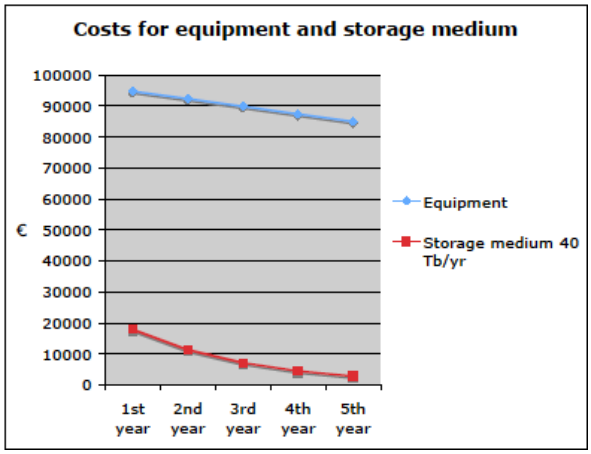

図8．公文書館のHSMストレージシステムのハードウェアコスト。

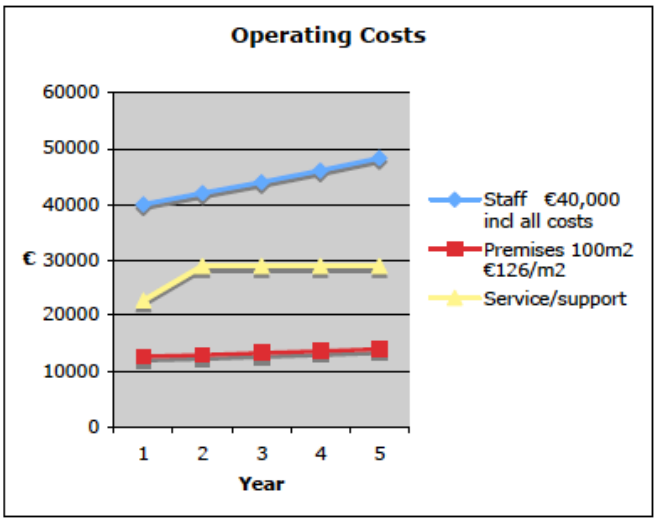

図９．公文書館のHSMストレージシステムのオペレーティングコスト。

コストを技術、職員、施設に分けて見ると（図10）、人件費が全体の39%になる。人件費は給与の上昇と、システムの拡大に伴う管理要員の増加によって、年々増加すると予測できる。全ての要員が高い資格を有している必要は無いが、スウェーデンの給与は他の国とは違っており、資格の有無が計算に大きな影響を与えることは無い。

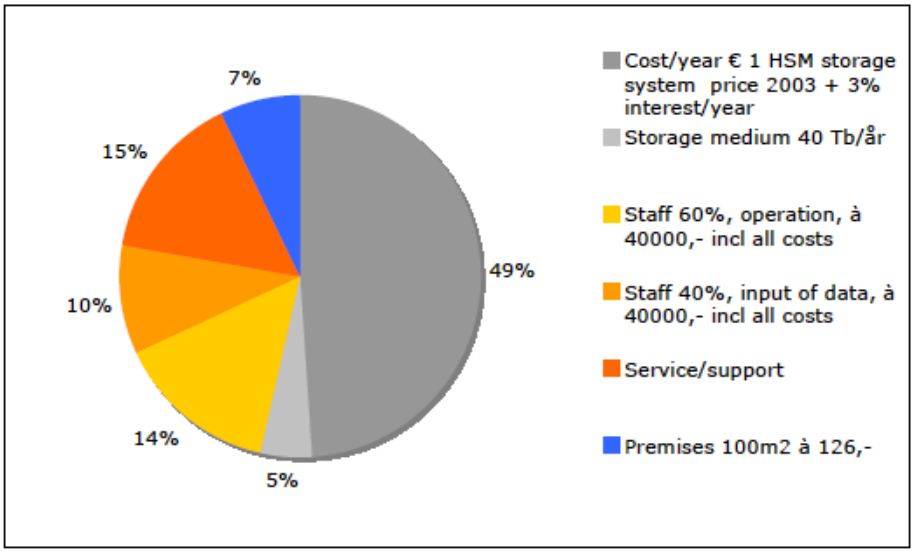

図10．公文書館のＨＳＭストレージシステムのコスト割合

不可能で、恐らく、馬鹿げていると言えるかもしれなが、10年先を予想してみると、唯一確かなことは、全般的なコスト指標と同様に給与が上昇するということである。
公文書館が行った計算は、公文書館が保存しているデジタル情報の大部分は活用頻度の少ない情報であるという仮定に基づいている。人件費はアクセス活動に関連しており、公文書館の状況は、企業、銀行、Google（前述のマイクロソフトのジム・グレイが例を述べて

いる）とは違って、システムの維持に必要な職員数は限られている。そうだとしても、公文書館の人件費と施設費は上昇し続けるだろう。図11は、職員、サポート、施設を合算したコストが30年間で機器コストの12倍以上になることを示している。ストレージメディアのコストはこのグラフでようやく見える程度で、それも最初の10年だけである。

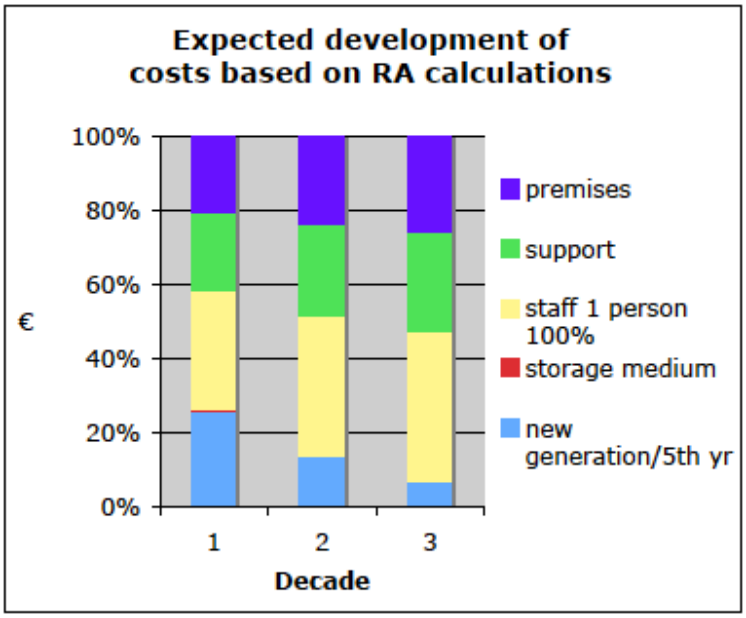

図１１．公文書館の10年毎のコスト予想

長期保存のコストは活動の度合いによって変わってくる。保存情報の利用度が上がれば、管理コストも上がる。情報の利用度が上がると、情報がアクセスされる外部サーバも必要になる。保存情報の利用度の増加と、控えめな経済規模を計算に入れると、システム管理とオペレーティングの人件費は機器コストの何倍にも増加すると予測できる。職員の増加に伴って、建物と土地のコストも増加する。サポートコストの先行きを予測するのは難しいが、恐らくシステムの規模と共に増加するだろう（図12）。

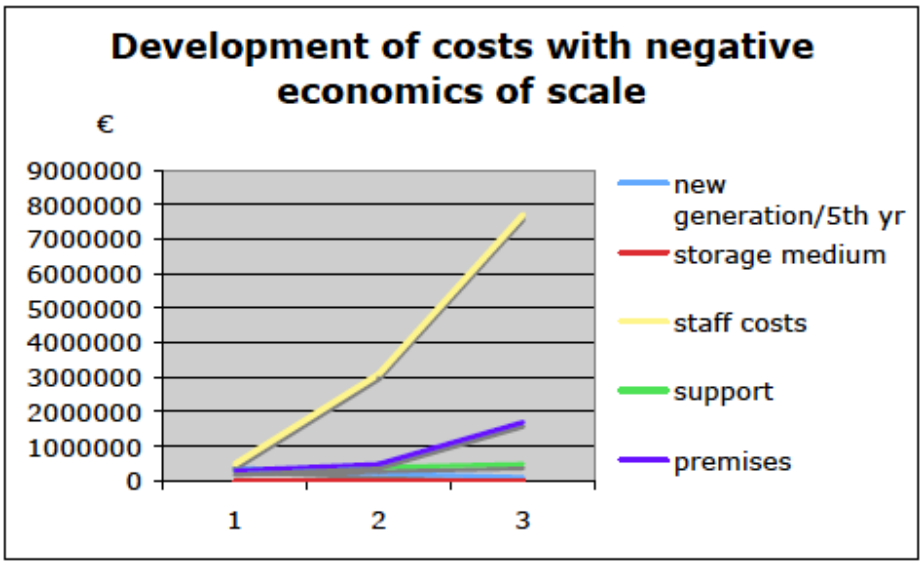

図１２．控えめな経済規模の公文書館コストへの影響度

デジタル化

ストレージのコストシナリオはコストを考えるための材料になる。コストシナリオは、常に更新とマイグレーション処理をすることでファイルを生かしておくのは当然のこととして、デジタルファイルの保存には、どの程度の長期的財政負担が必要なのかを示すので、デジタル化の初期の投資は慎重に考えなければならないことを強調している。デジタル化のコストは高いので、資料がデジタルブラックホールに飲み込まれて終わってしまうリスクは、より深刻である。デジタル化には、選別、記述とメタデータの作成、プロジェクト管理、実際の変換（デジタル　カメラでのスキャニング或いは取込み）といった異なった活動が含まれている。今では、スキャニングのコストは広く知られている。

スキャニングの品質は機器、処理仕様、スキャンする資料とその取り扱いについての知識に基づいている。機器の選択はスキャンする資料に関係する。スキャニングの仕様はオリジナル情報の特性と品質に関係する。品質管理にとっては、デジタル化する資料についての知識は必須条件である。取り扱いとセットアップはワークフローの責任である。イメージと音声のデジタル化においては、オリジナルが持っている情報を最適に取り込むため、コンテントとキャリアについての専門的な知識が絶対的に必要である。

2005年に、公文書館で、紙資料のデジタル化のコスト計算が実施された。公文書館は約80人の職員を有する独自のスキャニング施設、MKC（メディア変換センター）を保有している。スキャニングの対象物は、製本されたもの、１枚ものの記録、大型の図表と図面である。以下の図はMKCの情報を元にしている。

MKCでは毎年、A4サイズ白黒600dpiのファイルで5百万イメージが電子化されている。スキャンされたイメージ毎のコストは約0.10ユーロである。記録は自動給紙装置の付いたスキャナでスキャンされる。1個のデジタルファイルを作成するコストの内訳を図13に示す。コストの3分の1はスキャニングにかかり、準備、品質管理、抽出、管理の4項目が主なコストで同じ割合になっている。

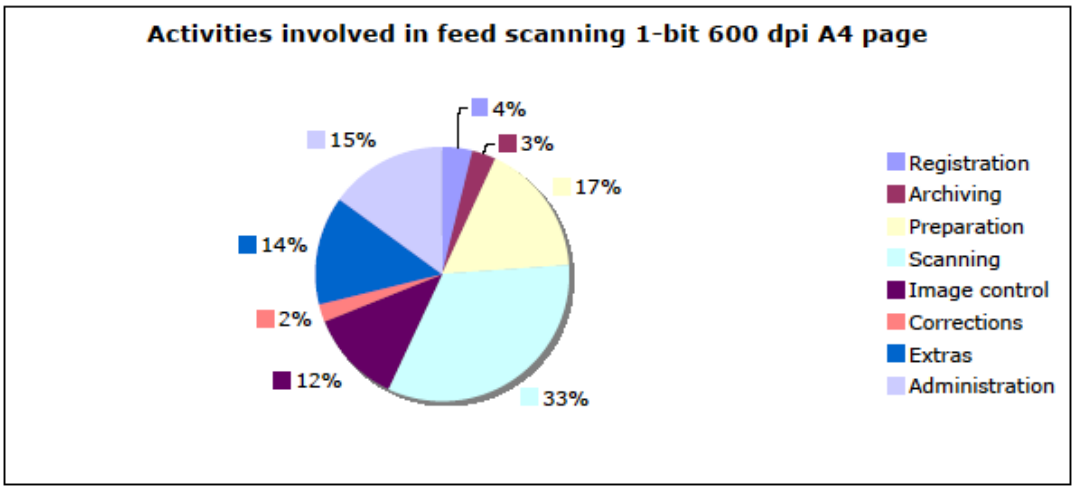

図13. 公文書館のスキャニング施設、MKCのA4サイズ紙のコスト割合

大型の図表と図面のスキャニングは297dpiで8ビット、グレースケールでマニュアル給紙のスキャナで行っている。1ファイルの作成コストは約0.61ユーロで、毎年、132万1千イメージファイルが作成されている。スキャンされたファイルのコスト内訳を図14に示す。ここでのスキャニングのコストはA4サイズ紙のスキャニングコストの約2倍、65%になっており、管理が2番目のコスト要因になっている。残りの項目は、ほぼ同様のコスト割合になっている。

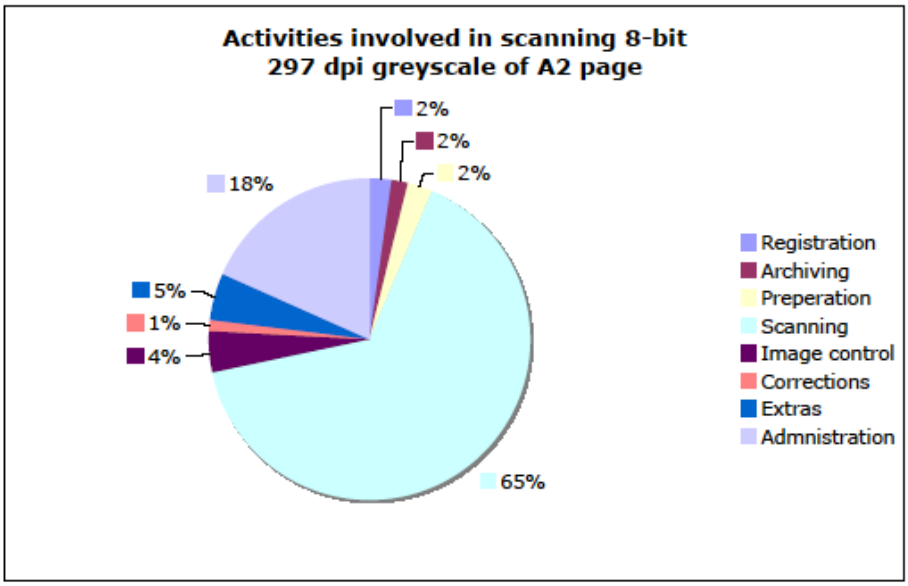

図１４．公文書館のスキャニング施設、MKCのA2サイズ紙のコスト割合

これがオーディオビジュアル情報のデジタル化になると、話は全く別になる。オーディオビジュアル情報のデジタル化は非常に時間がかかると共に、巨大なデジタル情報が作成

される。これはデジタル化以外には、将来のために資料を保存する手立てが無い場合に限って行われる事例である。言い換えれば、オーディオビジュアル資料をデジタル化しなければならない場合、その結果は巨大な量のデジタルデータが生まれ、それを長期間保存しなければならないことになる。

2004年に、スウェーデン文化省は国立オーディオビジュアルアーカイブス（Statens Ljud och Bild Arkiv, SLBA）のコレクションを保存する戦略を検討した報告書、「Preserving sounds and Images」(注4)を出版した。コレクションは4千5百万時間のオーディオテープ（30%）とビデオテープ（70%）で構成されている。これが、CD（16ビット、44,100khzサンプリングレート）やDVD(MPEG2)の様な、多くの人が最小限の品質と考えている、若干の圧縮或いは特定のフォーマットでデジタル化された場合、情報量は合計で８テラバイト（8百万ギガバイト）になる。これをもし、現在最高の“本当の品質”で取り込もうとすれば、データ量はもっと多くなるだろう。そしてこの問題についての技術の改善が早いため、“本当の品質”の定義は控えめに言っても流動的である。コレクションには様々なフォーマットや記録方式が含まれているので、デジタル化のプロセスも多様になってくる。例えば、1/4インチの講演テープはオリジナルの記録速度の2倍でデジタル化するよう提案されている。この種の大量のテープにとっては時間の節約が重要なので、これが十分な品質であろう。しかし、最高の効率で行っても、事業全体は9千万ユーロのコストで10年かかると予測された。

この報告書は「変換は、条件及び技術的な環境を考慮すると10年以内に行われることが望ましい。」と述べている。陳腐化している機器と同様に、オリジナルのメディアは劣化し続けており、正常に作動するように維持するのが困難なため、この種の資料はできるだけ早い時期に保存しなければならない。

報告書の中で、コスト割合は明確にされていないが、数人の職員が、たくさんの機械をセットアップして並行的に作業していることから、コストの殆どは変換作業にかかるだろう。準備と抽出が恐らく2番目のコスト要因であろう。

オーディオビジュアル資料を変換する場合は、アナログ機器の保守コストと、最適な信号を抽出するための、調整のコストを計算に入れなければならない。これは専門家の作業で、時間がかかる。他の資料と比較した、AV用のギガバイト当りの作成コストを図15に示す。

(注4) Bevara ljud och rörling bild (SOU 2004:53), Swedish Ministry of Culture, 2004

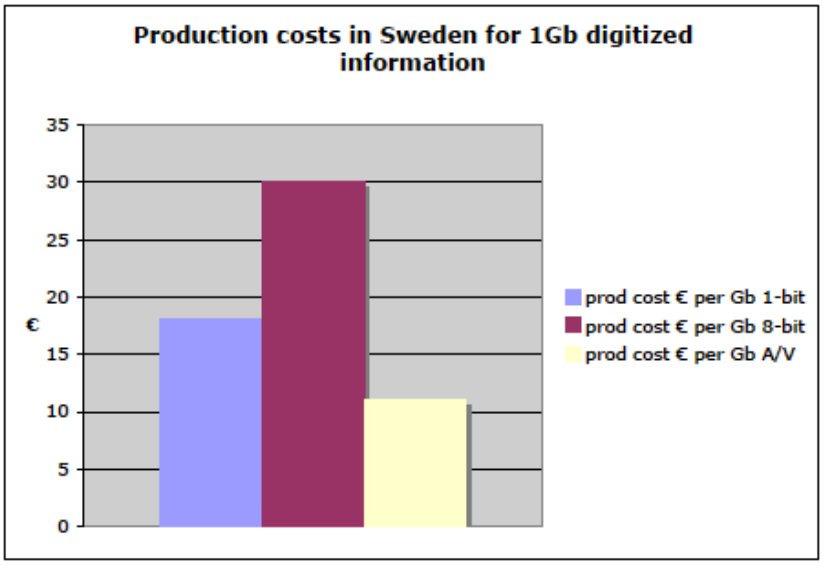

図15．ギガバイト当りの、１ビットと８ビットのイメージファイル及びオーディオビジュアルファイルの作成コスト。

全ての資料を変換すると、年間80万ギガバイトの情報が作成されることになる。図16はオーディオビジュアルファイルの年間作成予測を、MKCで毎年作成されるイメージファイルの量と比較した表である。このデジタル化に関する巨大な投資は、長期に渡って行われてきた仕事を保存するための適切な準備と釣り合わなければならない。

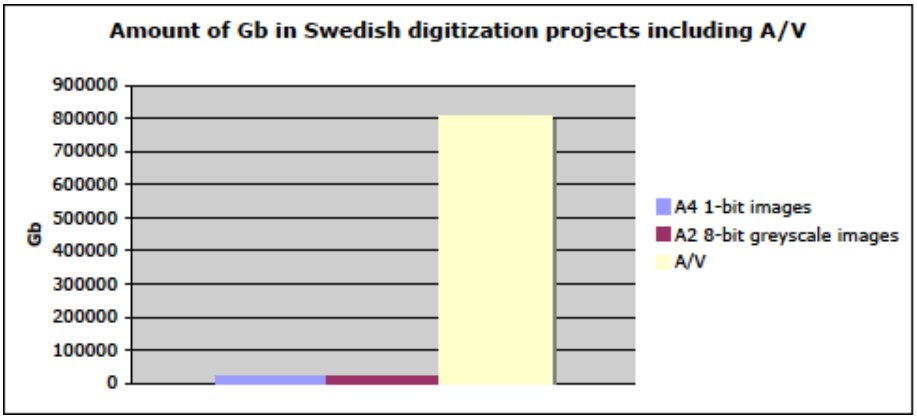

図１６．今後のSLBA向けのMKCでのデジタル化プロジェクトで1年間に作成されるイメージとオーディオビジュアルファイルの量の比較。

しかし、ストレージのコストだけを考えると、将来のことを期待するのが困難な規模の組織的な資金投資が必要になるのは明白である。 オーディオビジュアル資料は、大規模なプロジェクトへの資金投資と継続的なメンテナンスが保証されるのかが疑わしい状況の中で、変換と長期保存は何とかして実行しなければならない。そのためにはデジタルフォーマットに変換してデジタルファイルで保存する以外に選択肢が無いため、コスト的に苦

境に陥ることになる。
　紙資料の場合は、アクセス目的でデジタル化し、保存はオリジナル又はマイクロフィルムで対応することが可能である。
　「まず、デジタルイメージを作成し、そのデジタルイメージからCOM（これが保存用フォーマットになる）を作成する。そして、デジタルファイルを将来的に保有することは約束しない」という戦略は財政的な観点からは賢明かもしれない。将来デジタルが必要になった時は、COMから安価にデジタルに再スキャンできるので、デジタルコレクションは限定された期間だけ利用し、永久に保有する必要は無いかもしれない。
　現在、公文書館は、劣化した資料情報の保全に長い歴史を持っているマイクロフィルム化の戦略を改善する動きの中で、COMの利用が実行可能かどうかを調査している。公文書館ではイメージファイルを単にマイクロフィルムに変換するだけでは、検索のためのデータが無いことになるので、検索用メタデータも一緒にマイクロフィルムに変換することを検討している。
　デジタルイメージは直接使われるが、マイクロフィルムは情報を保存する容器なので、デジタルと同じ検索データは必要としない。
　フィルムは将来必要になった時、非常に早く（再）スキャンできる。そして、検索可能な状態でイメージをデジタルの世界に提供できる。

　どのような戦略を選択したとしても、大規模なデジタル化を開始する前に、必ず検討しなければならないことは、長期的に必要な資金が実際に確保できるのかということと、その予測に基づいて長期保存戦略を立案することである。
　かつて他のプロジェクトがそうであったように、プロジェクトが新たなデジタルブラックホールに飲み込まれて破綻する危険を回避するためには、あらゆる可能性をカバーしたコスト予測を計画プロセスの一部として組み込むことである。

（訳者後記）
　この報告書は、TAPE（Training for Audiovisual Preservation in Europe）のテクノロジーワーキンググループのサイトに掲載された、スウェーデン国立公文書館における、デジタル情報の長期保存に関する最良の実践とは何かをコスト面から考察した報告書です。
　Electronic Records Archives（ERM）の技術開発が進む中で、デジタル化保存に長い実績を有するスウェーデン国立公文書館がデジタル情報をマイクロフィルムに変換して長期保存する選択肢を検討することになった理由は日本でも参考になると考え、著者のJonas Palm氏の許可を得て翻訳した。オリジナルは下記のサイトで入手できます。
http://www.tape-online.net/technology.html - palm

デジタルをマイクロフィルムに変換する技術については下記を参照ください。
http://wwwjp.kodak.com/JP/ja/business/products/digitalpreservation/guide.shtml

2006年8月24日
楢林幸一