August 27, 2010

DataSpace:
A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data

Serge J. Goldstein
Mark Ratliff

Office of Information Technology, Princeton University

At the May 5th, meeting of the National Research Board, the NSF announced that, in October 2010, it would require that all grant proposals include a data management plan. This announcement represents the next step in what has been a growing trend on the part of government agencies to require researchers to plan for the preservation and sharing of the data produced by publicly funded research.

The NIH data sharing policy, published in 2003, specifies that "all investigator-initiated applications with direct costs greater than $500,000 in any single year will be expected to address data sharing in their application (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html). The data archiving policy of the NSF's division of social and economic sciences states that "grantees from all fields will develop and submit specific plans to share materials collected with NSF support" (http://www.nsf.gov/sbe/ses/common/archive.jsp). In July of 2009, the National Academy of Sciences published a white paper entitled "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age" (http://gking.harvard.edu/replication.shtml) which strongly argues for the long-term preservation of research data (http://brtf.sdsc,edu/biblio/BRTF_Final_Report/pdf).

It is now abundantly clear that researchers must consider the preservation and sharing of their data as a key component of any research effort. The problem that arises, for the researcher and the granting agency, is how to fund and manage such preservation in a sustainable way. Grant funding typically is for projects of limited duration. How can we fund and sustain long-term, indefinite preservation of research data if our grant models involve short-term, limited resourcing? This article proposes a model for doing exactly that. The model can be summed-up with the phrase:

Pay Once, Store Forever (POSF)

We propose that long-term data storage be funded by one-time payments that cover the current costs of storage, and leave enough excess funds to cover on-going replacement and management of that storage. This is made possible by the steady decline in the cost of physical data storage over time, as well as the steady increase in the amount of storage that can be managed by a given number of staff.

The Funding Model

We begin by considering the funding of physical storage devices (disk drives, tapes, etc.). We define the following terms:

Let:

C = the initial cost of the physical storage required to preserve a file.

D = the rate (as a fraction) at which the cost of storage decreases, on a yearly basis.

R = the average number of years that elapse before the storage device must be replaced.

T = the total cost of the storage, assuming we are storing the data "forever".

We can now compute T as follows:

$$T = C + [(1-D)^r \times C] + [(1-D)^{2r} \times C] + [(1-D)^{3r} \times C] + \ldots$$

The total cost, being the sum of the initial cost and all recurring replacement costs, is known as a "power series." This power series approaches a finite value (called a limit) when D is greater than 0. That value is:

$$T = C \times \frac{1}{1-(1-D)^r}$$

For example, if we take D to be 0.20 (storage costs decrease by 20%) per year, and R to be 3 (we replace the storage every 3 years), then:

$$T = C \times \frac{1}{1-0.512} \approx C \times 2$$

This equation shows that, if storage costs decrease over time by about 20% a year, and you replace the storage every 3 years, then, if you initially charge twice what the storage costs, you will have enough money to fund the replacement of that storage "forever".

To simplify the following discussion, we will adopt the convention of referring to the initial cost multiplier as "S" (the "storage" factor). That is:

$$S = \frac{1}{1-(1-D)^r}$$

and

$T = C \times S$ (Total Cost = Initial Cost * Storage Factor)

The table below shows the value of S for a range of D and R values (storage cost depreciation and replacement cycles). Notice that, for a wide range of values, S remains relatively small. Also note that the terms of the equation above get very small very quickly; that is, after about twenty years, the marginal cost of storing a given amount of data has become much smaller than today's cost.

Computation of S (Storage Factor) for various values of D and R (rounded)

D R	0.1	0.2	0.3	0.4	0.5
3	3.7	2.0	1.5	1.3	1.1
4	2.9	1.7	1.3	1.1	1.1
5	2.44	1.5	1.2	1.1	1.0

Is this a reasonable model? The notion that storage has to be replaced every so-many years seems reasonable and in keeping with most institutional practices. The key to the model then is the assumption that storage costs will decline steadily over time. We believe that this assumption is supported by the data on storage costs over the recent, and even remote, past.

In 1981, a Morrow Designs 10 megabyte drive cost $3,000, or $300 per megabyte (advertisement in *Creative Computing* magazine, December 1981, page 5). Today one can purchase a 500 gigabyte drive for $600. That's 500,000 times more storage for twice the cost, or a 250,000 fold decrease in cost over 30 years, which averages to about a 35% cost decrease per year. More recently, an IBM 20 gigabyte drive sold for approximately $280 in 200 (Advertisement on page 64 of *The Computer Paper*), or about $15 per gigabyte. Compared to today's 500 gigabytes for $600, that's a twelve-fold decrease in 10 years, which averages to about 23% per year. Given these numbers, we feel that a 20%/year average decrease in cost is reasonable, but the model produces comparable storage factors even with lower average yearly decreases.

Thus far we have addressed the costs associated with physical storage devices. What about the logistical costs like staffing and facilities needed to support these devices? Unlike the storage itself, staff and facilities costs do not decline steadily over time … in fact, the opposite is the case; these costs steadily increase over time. However, this is only true if one considers these costs on an unrated (flat) basis.

If we look at the costs of staff on a PER GIGABYTE (pro-rated) basis, then these costs do indeed decline steadily over time. Although one may be paying a storage administrator twice what one was paying twenty-five years ago, the amount of storage that administrator is managing has gone up by a factor of 100 or more. Thus the staff costs pro-rated across storage can also be modeled using the above equation. The same is true for facilities costs. The point here is that, in deciding

what to charge a researcher to store any given file, we need to look at the **marginal** cost of that storage to the institution providing the storage service. If we include logistical costs in this computation, then those costs (people, buildings, software) need to be calculated on a per-unit-of-storage basis and then incorporated into the model. We believe that, on a per-unit-of-storage basis, all of the costs associated with storing data decrease over time, and thus can be modeled as above.

We do not think, however, that it makes practical sense to include all of these costs directly in our funding model, because we do not believe that it makes sense to charge these costs directly to the researchers. Funding for staff and facilities represents a fixed, relatively stable charge that can be cost-recovered from grant overhead, or handled through central funds, rather than being recovered directly from grant payments. However, if desired, these costs could be built into the model through an appropriate adjustment of the S factor.

The Operational Model

The POSF funding model makes sense only if storage costs decline steadily over time, and if ancillary costs associated with storing data are kept to a minimum. To that end, we propose that the *POSF* funding modeled be married to an operational model which minimizes ancillary costs and which meets the emerging "sharing" requirements of the NSF and other granting agencies. The management and sharing model we propose can be summed-up with the phrase:

Write Once, Read Forever (WORF)

The essential notion behind this model is that the *POSF* funding scheme only works if management costs are kept to a minimum, and if a uniform approach can be taken to the sharing of the research data. Unlike *POSF*, which is described using an equation, *WORF* can be described using a set of management "principles" which have a twofold aim: first, to insure that research data is, and continues to be, publicly accessible, and, second, to minimize the costs associated with storing and disseminating such data.

WORF Principles:

1) The storage provided and paid for may not be "re-used". The researcher is paying for the permanent storage of a file, not for a given amount of storage.

2) At the time a file is written to the store, a permanent URL is assigned, and the data itself becomes read-only. It may not be changed (although it may be made unavailable). The meta-data associated with the file may be changed by the user or by authorized staff members, but the data itself is fixed and cannot be changed. If the data are in "error", the user may choose to store a corrected version, but he/she may not change the original data.

3) All data in the repository are to be made publicly accessible. Complex access management functionality will be avoided. The researcher may "embargo" the data for a short initial period (a few years), but ultimately all of the data must be publicly accessible. The sole

exception is in cases where the repository is legally obligated to make the data unavailable (e.g., copyright violation, material that violates institutional rules, material that violates state or national statutes).

4) The repository only provides storage for the bits associated with the data, and a variable set of meta-data (all files have a common set of meta-data, and may have additional meta-data, as warranted). No data conversion/migration services are required to be provided.

5) The fate of the researcher is irrelevant to the fate of the data. Once paid for, the repository assumes all responsibility for the storage and management of the data. The researcher retains "copyright", and may pass such copyright on to others, but at the time of the original submission, the repository is granted a permanent an unalienable right to store, publish and disseminate the data.

For completeness, we can add to the above the principles that encapsulate the *POSF* funding model, as follows:

6) At the time a file is stored, a charge will be assessed based on the amount of storage occupied by the file (e.g., gigabytes). Once paid, no further charges will accrue. Pay Once, Store Forever.

7) The repository may choose to offer ancillary services, such as data conversion and/or specialized data delivery. Such services will be separately priced, and will typically be paid by the person requesting the service.

We propose to call any repository that abides by the above set of principles a *Dataspace Repository*.

Dataspace at Princeton University

The Office of Information Technology (OIT) and the Library at Princeton University have partnered to create a Dataspace Repository at Princeton. We have selected DSpace (http://www.dspace.org/) as our repository software because it affords us the file storage, sharing and management (meta-data) capabilities that are needed to implement the Dataspace architecture. In addition, we have tied our DSpace system to system that generates permanent URLs following the Archival Resource Key (ARK) identifier scheme (http://tools.ietf.org/html/draft-kunze-ark-15). We have recently (spring, 2010) made this repository available to our faculty for storing research data. The system may be accessed at:

http://dataspace.princeton.edu

A key challenge in implementing this system was to come up with our "storage factor". The following graph [Figure 1] shows purchases OIT has made of Fibre Channel (FC) and SATA disk

drives over the past 6 years. The first purchase of FC storage was made in October of 2003 with the latest purchases shown having occurred in March of 2009.

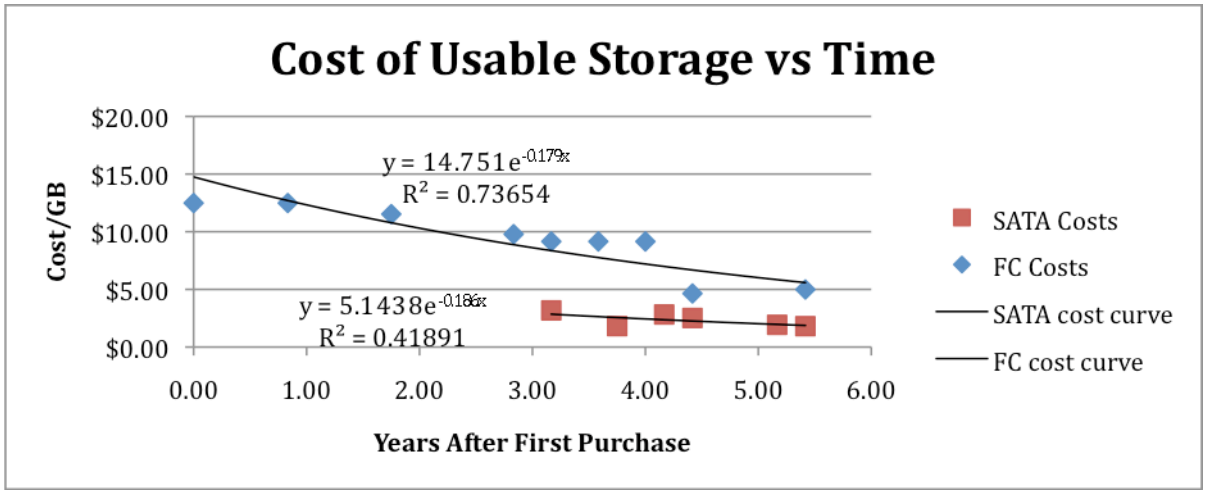

Figure 1

The graph also shows least squares fits of exponential curves to the FC and SATA purchase data. The fit shows FC storage costs have dropped by 16% from year-to-year (e.g. $e^{-0.179} = 84\%$) and SATA storage costs have dropped by 17% from year-to-year (e.g. $e^{-0.186} = 83\%$). These two curves show how the cost of a given technology (FC and SATA in this case) independently decreases over time. But we should also realize that the ability to migrate to new storage technologies as they become available will allow us to benefit from an even steeper decrease. For instance, when SATA became available in December of 2006, we could have moved data from the previously existing, and more expensive, FC storage to less expensive SATA storage and realized an immediate, one-time 65% decrease in cost.

For the purposes of our calculations we then made the conservative assumption that storage costs will decrease by 20% each year. The total unit cost of SATA storage purchased today and refreshed every four years would be:

$$T = \frac{\$1.81/GB}{1-(0.80)^4} = \$3.07/GB$$

Costs of periodic tape backups should also be included. Backups of this storage are written directly to tape for disaster recovery purposes. At this time, there is no internal charge for this service, however OIT is considering making this service available more broadly and introducing a charge. Although this service is not yet fully defined, we can make some estimates on the proposed costs based on tentative pricing for this service. We calculate the cost of the backup service at \$0.24/GB/yr and the cost decrease to be 10% each year. The total unit cost of backup will be:

$$T = \frac{\$0.24/GB}{1-(0.90)} = \$2.40/GB$$

Combining these two results we find that the total one-time cost for storage and backups will be $5.47 per Gigabyte. Rounding up, we are now offering a POSF charge of $6/Gigabyte. That is the current cost to have data stored, and shared, forever.

FAQ

In discussing this model with colleagues and potential customers, a number of questions come up frequently, and these are listed, along with our responses, below:

Question: What happens if you are wrong in your computation of S (the storage factor)?
Answer: We think that 20% is a conservative estimate for the year-to-year decrease in storage costs, but, even if turns out to be wrong, we can adjust our S value to compensate over time. Note that S is potentially computable every time a customer uses the system to store a file. It is not a fixed value, but rather one that can be adjusted as needed to track changes in storage cost. Note also that, even with relative low percentage decreases in storage costs, after about 20 years, storage costs are nearly nill (e.g., how much does a kilobyte of storage cost today?)

Question: Your model only accounts for the cost of disk drives; what about all of the other costs?
Answer: Actually, the model can account for any costs which decrease over time when computed on a per-unit-of-storage basis. That applies to equipment and people as well as to actual drives, so long as those costs per unit-of-storage decreases over time. What the model does not account for are ancillary services, such as special requests for data delivery, data migration/transformation, and other end-user services. We believe that these costs should be borne by the person requesting the service, not by the researcher. The model does provide for storage for an indefinite period as well as basic internet-based sharing, but it would not, for example, fund the duplication and transfer of data into another repository, or the conversion of data from one format into another.

Question: What happens if the researcher leaves Princeton, or dies?
Answer: Nothing. The data continues to be stored and shared. The repository has an unalienable license to store and share data indefinitely.

Question: Who manages the data, the meta-data, and access to the data?
Answer: The owning researcher, or whoever the repository designates as the owner. If/when that person (or persons) leaves Princeton and/or ceases to be active, management shifts to whoever is in charge of the DSpace "community" in which the data is stored. Ultimately, it shifts to whoever runs the DataSpace service.

Question: How long is "forever"?
Answer: We are using "forever" in this document in the sense of "indefinitely". There are no built-in expiration dates. Further, the institution warrants that it will make a best-case effort to continue to support the repository. At a minimum, the institution warrants that data will be migrated into other repositories if the current repository can no longer be supported, and that this migration will not affect the "permanent" URL associated with the date.