Repository of NSF-funded Publications and Related Datasets: "Back of Envelope" Cost Estimate for 15 years

Beth Plale[1,2], Inna Kouper[1,3], Kurt Seiffert[4], and Stacy Konkiel[3]

[1] Data To Insight Center, Indiana University

[2] School of Informatics and Computing, Indiana University

[3] Indiana University Libraries

[4] University IT Services Research Technologies

March 2013

Executive Summary

In this back of envelope study we calculate the 15 year fixed and variable costs of setting up and running a data repository (or database) to store and serve the publications and datasets derived from research funded by the National Science Foundation (NSF). Costs are computed on a yearly basis using a fixed estimate of the number of papers that are published each year that list NSF as their funding agency. We assume each paper has one dataset and estimate the size of that dataset based on experience. By our estimates, the number of papers generated each year is 64,340. The average dataset size over all seven directorates of NSF is 32 gigabytes (GB). A total amount of data added to the repository is two petabytes (PB) per year, or 30 PB over 15 years.

The architecture of the data/paper repository is based on a hierarchical storage model that uses a combination of fast disk for rapid access and tape for high reliability and cost efficient long-term storage. Data are ingested through workflows that are used in university institutional repositories, which add metadata and ensure data integrity. Average fixed costs is approximately $.0.90/GB over 15-year span. Variable costs are estimated at a sliding scale of $150 - $100 per new dataset for up-front curation, or $4.87 – $3.22 per GB. Variable costs reflect a 3% annual decrease in curation costs as efficiency and automated metadata and provenance capture are anticipated to help reduce what are now largely manual curation efforts.

The total projected cost of the data and paper repository is estimated at $167,000,000 over 15 years of operation, curating close to one million of datasets and one million papers. After 15 years and 30 PB of data accumulated and curated, we estimate the cost per gigabyte at $5.56. This $167 million cost is a direct cost in that it does not include federally allowable indirect costs return (ICR).

After 15 years, it is reasonable to assume that some datasets will be compressed and rarely accessed. Others may be deemed no longer valuable, e.g., because they are replaced by more accurate results. Therefore, at some point the data growth in the repository will need to be adjusted by use of strategic preservation.

Introduction

National Science Foundation (NSF) is one of the organizations that are at the forefront of data access and exchange initiatives. In addition to making substantial investments in supporting research and education, it promotes responsible ownership and management of data and materials collected during research. Beginning January 18, 2011, proposals submitted to NSF must include a document that describes types of data and standards of data description (metadata), policies for access, sharing and re-use as well as plans for archiving data, samples, and other research products (National Science Foundation, 2011).

Support of data management and exchange requires significant investments in cyberinfrastructure and making sure that such infrastructure is stable, user-friendly and cost-effective. Most likely, the data cyberinfrastructure will be a networked environment that supports many heterogeneous ways of capturing, storing, processing, analyzing and publishing data. A repository that consolidates publications and research data from all NSF-funded research can become a key component of such a networked environment.

In this study we estimate the cost of setting up and running a repository (or "database") to store and serve the publications and datasets derived from NSF-funded research. We approach the estimation as a task that needs to be conducted within a limited timeframe and propose "back of envelope" longitudinal cost estimate that provides concrete numbers. The study relied on high level of expertise at Indiana University, particularly in its libraries, university research and IT group, and our extensive engagement with NSF office of cyberinfrastructure and computer and information science and engineering directorates. We believe our approach and findings will be of significant interest to the research universities as well as to broader data preservation initiatives worldwide.

Relevant Work

Several national and international projects attempt to model costs of digital preservation and curation. The list compiled by the Open Planets Foundation (2013) contains links to more than seventeen models and approaches that offer guidance on how to calculate the costs of preserving digital information. Some approaches emphasize conceptual modeling and outline categories and dependencies that need to be taken into account during cost calculations. Others, for example, LIFE: Life Cycle Information for E-Literature[1] or CET: Cost Estimation Toolkit[2], offer tools to calculate cost of digital preservation.

Most of the existing models are based on or mapped against the Open Archival Information System (OAIS) Reference Model (Consultative Committee for Space Data Systems, 2002); they also rely predominantly on the activity-based costing (ABC) approach (Palaiologk, Economides, Tjalsma, and Sesink, 2012). Overall, models represent the costs that are incurred at all stages of the lifecycle of the digital materials, such as the costs of ingesting and migrating materials, the costs of hardware and software upgrades, the costs of retrieval, and the costs of disposal (APARSEN, 2013). A series of case studies from Cambridge University, King's College London, Southampton University, and the Archaeology Data Service at York University have demonstrated that a significant share of cost is

[1] http://www.life.ac.uk/
[2] http://opensource.gsfc.nasa.gov/projects/CET/

comprised by acquisition and ingest of materials (42%), archive and preservation (23%) and facilitation of access (35%) (Beagrie, Chruzs, and Lavoi, 2008).

A few models provide numeric estimates of how much it would cost to create and maintain a repository. The Royal Library of Denmark and the Danish National Archives estimate the accumulated "lifetime" cost of digital preservation to be 2,000,000 euros per 200 terabytes of materials over 20 years (Kejser, 2012). Re-calculating this amount into dollars per gigabyte, this estimate is approximately $12.9 per gigabyte (GB). This study from Denmark also proposed a prototype for calculating the costs and acknowledged the difficulties of cost estimation due to lack of empirical data. Goldstein and Ratliff (2010) offered an empirical estimate of the cost of storage as $5.47 per gigabyte. Their study calculated only the cost of storage and was based on the formula that included the initial cost of the physical storage, the rate of storage cost decrease, and the average number of years between infrastructure replacements.

Assumptions

Prior work on repository cost modeling demonstrates that the overall cost of maintaining a repository depends on the goals of the repository and its type, the types of curated objects, as well as on specific activities and related stages of digital object lifecycle that are associated with digital storage and preservation. Most cost modeling studies acknowledge that because of such a variety of factors, it is difficult to provide actual cost numbers.

The goal of this study is to provide decision-makers with an empirical cost estimate that is based on the characteristics of a particular organization or repository, in this case it is the National Science Foundation and its repository, and can guide repository planning and development. To provide empirical estimates, our study is grounded in the following simplifying assumptions.

We assume that the overall cost of a repository depends primarily on two factors: the production rates of digital objects and the actual cost of processing and storing those objects (see Figure 1 below).

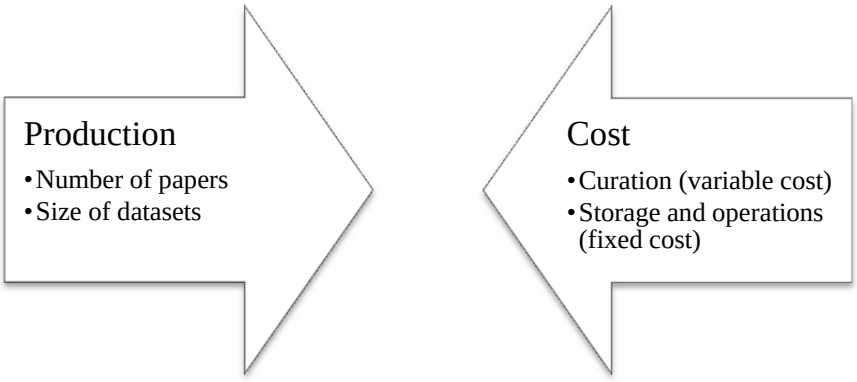

Figure 1. A simplified cost model.

As determined by the goals of creating a repository that stores NSF-funded research output, we limit our digital objects to published papers and datasets that were used in those papers. We take a paper-oriented approach rather than a data-oriented approach to calculation of repository costs. The paper-oriented approach presumes that the primary variable in calculating the number of digital objects to store and process is the published paper. Each paper is assumed to have one data set associated with it.

The advantages of a paper-oriented approach include relatively easy tracking of publications via existing scientific databases and portals and reliable association of datasets and papers. The disadvantages of the paper-oriented approach include difficulties in representing continuous data sets, such as longitudinal time-series data or observational data, and by extension misrepresentation of data that are collected for future rather than for immediate use. Massive observations of large distributed phenomena or over long periods of time that are facilitated by instruments such as astronomical observatories or ecological surveys are intended to serve communities of researchers who can sample active data repositories and generate many papers. Such data are not traceable through a paper-oriented approach in their entirety, unless rather artificial methods are introduced to keep the data available in full.

The size of each dataset depends on the type of research that is carried out. A study that relies on observation satellites or astronomical telescopes will produce datasets that are significantly larger than a study that uses field data collected via interviews and surveys. For simplicity, we assume that a dataset can be small, medium or large. The following average sizes are used for each of these categories:

- 1GB: small data set average
- 10GB : medium dataset average
- 100GB large data set average

Another important aspect of long-term cost estimation is the change rate of certain expenses or asset values over time, for example the amortization rate of capital investments. To simplify the model, we assume amortization and inflation rates to be stable and changing at a fixed percent over time.

Given the conditions and assumptions above, we provide crude estimations of the cost of the development of a repository to store NSF-funded papers and related datasets. In addition to serving as guidance in developing a repository, we hope that these estimates stimulate transparency and further sharing of cost information among data archives and repositories, reduce overall expenses and ultimately increase benefits from data preservation and sharing.

Production Estimates

Number of Published Papers

The annual number of papers published from NSF-supported research is derived from querying the Thomson Reuters *Web of Science* citation database as well as from estimating the number of publications in conference proceedings at conferences sponsored by the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery (ACM), and Springer Publishing House. We included conference publications from those three organizations because *Web of Science* does not capture conference proceedings in their entirety. Peer-reviewed conference proceedings, which are often more selective and rigorous than journal submissions, serve as a primary dissemination venue for research from computer and information sciences and engineering.

The *Web of Science* citation database was searched across its Science and Technology, Social Sciences, and Arts and Humanities databases for papers published in the 2011-2012 year that list National Science Foundation or its spelling variants (e.g., NSF, US NSF, and so on) as its funding agency. Additionally, we estimated the number of conferences that are sponsored by IEEE, ACM and Springer (1334, 150 and 150

conferences per year correspondingly). Each conference produces about 30 papers per year. Using our expert knowledge of these conferences, we estimate that about 20% of the conference papers come from research that is funded by NSF. The number of papers published in 2011-2012 is presented in Table 1 below.

Table 1. Estimated number of papers from NSF-supported research published in journals and conference proceedings in 2011-2012.

Source	Number of Papers
Journals and conference proceedings indexed by *Web of Science*	54,536
IEEE conferences	8,004
ACM conferences	900
Springer conferences	900
Total papers	**64,340**

The paper production rate is held constant over the 15 years of the estimate. Even though science and engineering output is growing at annual rate of 2.6% worldwide and 1% in the US (National Science Board, 2012), this rate could see a decline due to economic recess and budget cuts; therefore, we hold the paper production rate constant. Because the size of published papers is small compared to dataset sizes, for purposes of this study, we consider paper size to be negligible.

Size of Datasets

The NSF supports research and education through seven directorates, each encompassing several disciplines[3]:

- Biological Sciences (BIO)
- Computer and Information Science and Engineering (CISE)
- Engineering (ENG)
- Geosciences (GEO)
- Mathematical and Physical Sciences (MPS)
- Social, Behavioral and Economic Sciences (SBE)
- Education and Human Resources (EHR)

We categorize the size of a dataset based on the NSF directorate that supported the research from which the paper has been published. We make the simplifying assumption that the total number of publications is uniformly distributed across the seven directorates, so each directorate produces 1/7 of the total number of publications or 9,191 publications. We assume that every published paper has one dataset associated with it, and that every dataset is of the average size for its category. The approximate amount of data within each directorate is presented in Table 2 below.

[3] http://www.nsf.gov/staff/orglist.jsp

Table 2. Annual production of data within each NSF directorate.

Data set size / NSF Directorate	Approx. amount of data per year (terabyte, TB)
Social, Behavioral and Economic Sciences Directorate (SBE)	9
Computer & Information Sciences & Engineering (CISE)	9
Education & Human Resources Directorate (EHR)	9
Engineering Directorate (ENG)	90
Biological Sciences Directorate (BIO)	90
Mathematical and Physical Sciences Directorate (MPS)	900
Geosciences Directorate (GEO)	900
Total	**2,007**

The total amount of data added to the repository on the annual basis is two petabytes, or 30 petabytes of data over the 15 years. When calculated over all seven NSF directorates, the average dataset size is approximately 32 gigabytes.

Cost Estimates

Curation Costs (Variable Costs)

We separate the cost of curating a published paper from the cost of curating its dataset. We assume that because NSF encourages open access to research it supports, published papers and their citation information will be readily available. Considering that scientific journals can still impose up to 12 months of post-publication embargo on federally sponsored research (Basken, 2013), we can also assume that when a post-print copy arrives into NSF repository, it will already have a DOI assigned to it, resolved author ID and other types of metadata required for ingest and preservation. Hence, the cost of additional curation needed per paper is considered at zero.

The cost of data curation is estimated at around $150 per dataset. This charge includes up-front curation activities, such as ingesting datasets and supplying necessary metadata, ongoing preservation services, such as file format migrations. Commitment to maintaining access to datasets over 15 years is a fixed cost. This estimate is consistent with existing pricing models in data repositories [see, for example, (Dryad, 2013)]. Additionally, we factor in an annual 3% decrease in curation costs. Actual cost of curating a new dataset will vary and may initially exceed $150 per dataset. However, it can be expected that ongoing research in data science and digital data preservation will result in solutions that reduce these costs through task optimization and automated metadata and provenance capture. The cost of curating two petabytes of data annually as well as the decreasing average cost of data curation per gigabyte over 15 years is presented in the table below.

Table 3. Cost of curating datasets over 15 years.

Year	Curation cost per new dataset	Curation cost of datasets from 64,340 papers	Average curation cost per new GB
2013	$151	$9,732,002	$4.87
2014	$147	$9,448,546	$4.72
2015	$143	$9,173,346	$4.59
2016	$138	$8,906,161	$4.45
2017	$134	$8,646,758	$4.32
2018	$130	$8,394,911	$4.17
2019	$127	$8,150,399	$4.08
2020	$123	$7,913,009	$3.96
2021	$119	$7,682,533	$3.84
2022	$116	$7,458,770	$3.73
2023	$113	$7,241,524	$3.62
2024	$109	$7,030,606	$3.52
2025	$106	$6,825,831	$3.41
2026	$103	$6,627,020	$3.31
2027	$100	$6,434,000	$3.22

As can be seen from Table 3 above, the average cost of curating a dataset decreases from $4.87 per gigabyte in the first year of the repository to $3.22 per gigabyte in the 15th year of its existence.

Costs of Storage and Operations (Fixed Costs)

Our storage and operations cost estimations are based on the actual costs of storage infrastructure at a research university. As a baseline for 2013 we take expenses on capital infrastructure, including the purchase of a high performance storage system based on disks and tapes and offering up to 36 petabyte capacity, its maintenance and outside technical support, software and hardware license agreements, amortization and the cost of staffing. Staff personnel include system administrators, programmers, managers and a director for the project. We specify **$**3.7 million first year cost for initial purchase of storage equipment, facilities, and support. We project $1.9 million equipment refresh in year six and $1.2 million equipment refresh in the eleventh year. Licensing, hardware maintenance, and staff incur ongoing yearly costs of around $1.2 million. The distribution of storage and maintenance cost is shown in the figure below.

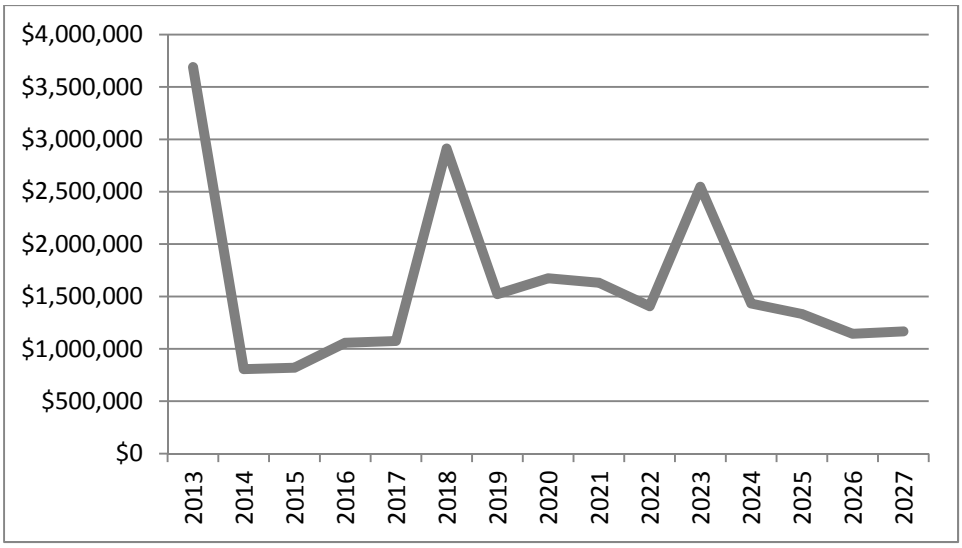

Figure 2. The cost of storage and operations over 15 years.

Overall Cost

A repository that contains published papers and related datasets from NSF-supported research will grow at 64,340 publications and 64,340 datasets per year on the annual basis. The curation and storage requirements for published papers are minimal and can be ignored for the purposes of quick estimates. The datasets will accumulate at the rate of approximately two petabytes per year and will require significant resources to store and curate them. According to our "back of envelope" calculations, the initial first-year cost of curation, equipment purchase, and operations is:

$$\$3,689,767 + \$9,732,002 = \$13,421,769$$

The annual cost of creating and maintaining a repository for NSF-funded published papers and related datasets will be around $1.6 million a year with spikes every six years for refresh and upgrades due to rapid changes in technology.

The overall cost of the repository over 15 years is estimated at $167 million. After 15 years and 30 PB of data accumulated and curated, the cost per gigabyte will be $5.56.

Conclusion

This study provides a quick "back of envelope" estimate of the cost of creating and maintaining a digital repository in the circumstances when speed and urgency surpass accuracy and thoroughness. In such circumstances it is important to generate simple yet meaningful numbers, rather than delve into extensive comparisons and calculations.

Our approach prioritizes the availability of information and internal research and administrative expertise over extensive research and predictive modeling. It provides empirical estimates that can be used as a ground for comparison across case studies that test cost models with real data. The confidential nature of financial information makes it difficult to make cost data public. We believe that a certain amount of sharing of cost data will be valuable in the efforts to improve data sharing and reduce its cost.

In the future we plan to address limitations of "back of envelope" estimate approach, incorporate other digital preservation cost models into our model and address the issues of OAIS compatibility. The future model would also need to address rapid changes in technology and cost-sharing opportunities, i.e., the so called economies of scale (Jones and Beagrie, 2000), as well as the cost of distributed and computational access over data.

Acknowledgements

We thank Craig Stewart, Associate Dean, Research Technologies and Executive Director of Pervasive Technology Institute and Robert H. McDonald, Associate Dean for Libraries and Deputy Director of Data to Insight Center for helpful discussions.

References

Alliance for Permanent Access to the Records of Science Network (APARSEN). (2013). D32.1 Report on Cost Parameters for Digital Repositories. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D32_1-01-1_0.pdf

Basken, P. (February 24, 2013). NSF Anticipates Pushing Boundaries on Open-Access Plan. *The Chronicle of Higher Education*,*59*(26). Retrieved from http://chronicle.com/article/Volume-59-Issue-26-March-8/137673/

Beagrie, N. and Jones, M. (2000). *Preservation Management of Digital Materials Workbook: a pre-publication draft*. Retrieved from http://www.jisc.ac.uk/dner/preservation/workbook/

Consultative Committee for Space Data Systems (CCSDS). (2002). Reference model for an open archival information system (OAIS), CCSDS 650.0-B-1 Blue Book. Retrieved from http://ddp.nist.gov/refs/oais.pdf

Dryad. (2013). Pricing plans and submission fees. Retrieved from http://datadryad.org/pages/pricing

Goldstein, S. J. & Ratliff, M. (2010). DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data. Retrieved from http://dspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf

Kejser, U. B. (2012). Cost Model for Digital Preservation. Retrieved from http://files.d-nb.de/nestor/veranstaltungen/praktikertag2012/kejser.pdf

National Science Board. (2012). *Science and Engineering Indicators 2012*. Chapter 5. Academic Research and Development. Arlington VA: National Science Foundation (NSB 12-01). Retrieved from http://www.nsf.gov/statistics/seind12/c5/c5h.htm

National Science Foundation. (2011). Grant Proposal Guide, Chapter II.C.2.j. Special Information and Supplementary Documentation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

Open Planets Foundation. (2013). Digital Preservation and Data Curation Costing and Cost Modelling. Retrieved from http://wiki.opf-labs.org/display/CDP/Home

Palaiologk,A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. *International Journal on Digital Libraries, 12*(4), 195-214. Retrieved from http://link.springer.com/content/pdf/10.1007%2Fs00799-012-0092-1.pdf