Apr 6th, 9:30 AM - 10:30 AM

Keynote Address: "DataSpace: A Model for Long-Term Preservation and Dissemination of Research Data"

Serge Goldstein
Princeton University

DataSpace

A model for long-term preservation and dissemination of research data

Storing Research Data “Forever”

Special Thanks To

- MacKenzie Smith, MIT Libraries
- “Managing Research Data 101”
- https://libshare.library.gatech.edu/clearspace/docs/DOC-3634.pdf;jsessionid=DF96E09B9D6BE9E5EC62A27717DC5868

What is “Forever”?

- Longer than a typical project?
- Longer than a typical career?
- Longer than a typical institution?
- 5 years, 10 years, 25 years, 100 years?
- How long to you keep stuff?
- Do you publish preservation policy?

What is “forever”?

Suggestion: treat data same way library treats books

- Intent is to preserve indefinitely
- As long as practical, feasible
- Cannot be precisely defined

Is “forever” reasonable?

- Cliff Lynch, CNI: “Granting agencies only expect data to be preserved for a few years, maybe 5-10. No-one is expecting data to be stored forever.”.
- Is that true of publication?
- Why should research data be treated differently from research publication?

What is “Data”?

- Numbers?
  - Recorded? Collected? Generated?
- Images? Video? Audio?
  - Shoah Foundation
  - In what format?
- Programming Code?
- Publications/Text?
  - In what format?
- Transcription service
- Is pure “raw” data useful
  - May require extensive meta-data to be useful

Working definitions:

Data:

Any storage entity produced or obtained as part of a research effort.

Forever:

Indefinitely. As long as practically “feasible”. Until we think no-one will complain if deleted.

Current Storage Models

- Let someone else do it
  - Government agency/lab/bureau
    - NOAA National Geophysical Data Center
    - GenBank (DNA data)
    - fMRIDC (fMRI publications and data)
    - NCSA Astronomy Digital Image Library

Why Save Data "Forever"

- Because we want to:
  - Available to ourselves and our students and colleagues
    - Where are the data sitting today? On a departmental server? On a computer under your desk? On a CD or DVD somewhere?
    - Where is your dissertation data?
  - Available to future scholars, including ourselves

Why Save Data "Forever"

- Because we need to:
  - Encourage honesty?
    - Gregor Mendel probably cheated
  - Like open-source: help uncover mistakes, bugs?
  - Open Data Movement
    - Mostly library/catalog data, map data, WordNet
  - Open Access Movement
    - Mostly publications
- Because it's not "our" data

Why Save Data "Forever"

- Because we have to:
  - Funding agencies want data "sharing" plans
  - NIH Data Sharing Policy (2003):

  http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

  "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."

NIH Data Sharing Policy

- “Applicants may request funds for data sharing and archiving. The financial issues should be addressed in the budget section of the application.”
- Specifics depend on grant, published in RFP, RFA or PA

NSF Former Data Archiving Policy

- Division of Social and Economic Scienes
- http://www.nsf.gov/sbe/ses/common/archive.jsp
- **“Grantees from all fields will develop and submit specific plans to share materials collected with NSF support, except where this is inappropriate or impossible.”**

NSF New Data Sharing Policy

Beginning January 18, 2011, proposals submitted to NSF must include a supplementary document of no more than two pages labeled “Data Management Plan”. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See Grant Proposal Guide (GPG) Chapter II.C.2.j for full policy implementation.

Other agency Policies

- See Gary King's Page on "Data Sharing and Replication"
- http://gking.harvard.edu/pages/data-sharing-and-replication
- See National Academy of Sciences "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age", July, 2009
- http://www.nap.edu/catalog/12615.html

Current Storage Models

- Let someone else do it
  - Government agency/lab/bureau
    - NOAA National Geophysical Data Center
    - GenBank (DNA data)
    - fMRIDC (fMRI publications and data)
    - NCSA Astronomy Digital Image Library
  - Professional society/Journals
    - Global Ocean Observing System: coordinates distributed data
    - Dryad: ecology/evolutionary biology

Current Storage Models

- Nice folks at another University
  - ICPSR, University of Michigan (political/social)
  - Dryad: ecology/evolutionary biology
  - Protein Data Bank (PDB): 3-D protein data
  - NCSA Astronomical Image Library
  - Sloan Digital Sky Survey
- The “Cloud”
  - Amazon S3
  - Google
  - DuraSpace

DuraSpace

- Fedora/Dpsace merged organization
- “DuraSpace is the independent 501(c)(3) not-for-profit born from a vision to help save our shared scholarly, scientific and cultural record. We are dedicated to sustaining and improving Fedora and DSpace, two of the most dominant open source repository solutions.”

DuraCloud

- Uses existing “cloud” storage providers (Amazon, RackSpace).
- Overlays “access and preservation tools”
- “Elastic capacity" coupled with a "pay as you go" approach.
- Released open-source code in July, 2010.
- Pilot program (NDIIPP).
- Public service in second-quarter 2011.

Current Funding Models

- Institution/department pays
- Grants pay monthly/yearly
- Haphazard
  - Some grant money
  - Some departmental money
  - Use whatever is available
  - Don’t worry, someone will pay

Current Funding Models

- Most require some form of on-going payment
- Advantages
  - Capitalist approach to data storage
  - If someone wants to pay, data gets saved
  - “Natural” expiration process
- Disadvantages
  - Capitalist approach to data storage
  - Who pays to save rarely used data?

Cloud Storage Costs

- Per Month charges
- Amazon S3
  - About 10/cents per gig for storage
  - Plus 10/cents per gig data transfer
  - Plus 1/cnet per 1000 requests
- Google
  - About 17cents per gig
  - Plus 10/cents per gig data transfer
  - Plus 1/center for 1000 requests

Cloud Storage Costs

- About $1.20-$1.50/year for storage
- ?? For data transfer ??? If 1 gig downloaded 20 times, another $2/year
- For a terabyte, that’s between $1K and $3K+ year
- You can buy a terabyte drive for $100

Different Approach

PAY ONCE, STORE ENDLESSLY (POSE)

Why Pay Once?

- Grants expire often and quickly
  - Monthly payments are problematic
  - Unfixed payments VERY problematic
- Researchers expire/leave pretty often

How Store Forever?

- Administrators/Librarians expire slowly
- Institutions expire rarely

Computing cost of storage

- Initial cost of disk drives, electronics
- Drives must be replaced every few years
- People must be paid every month (or so)
- Data must be managed and migrated
- How can you pay these on-going costs with a one-time payment?
  - Key: cost of storage decreases over time
  - Key: magic of math

The Business Model (1)

- I = Initial cost of storage
- D = rate at which storage costs decrease yearly, expressed as a fraction (e.g., 20% would be 0.2)
- R = How often, in years, storage is replaced
- T= Cost to store the data “forever”

$$T = I + (1-d)^{r} * I + (1-d)^{2r} * I + \ldots.$$

If d=20%, r = 4:

$$T = I + (.8^{4}) * I + (.8^{8}) * I + \ldots.$$

The Business Model (2)

If d >0,

$T = I + (1-d)^r * I + (1-d)^{2r} * I + \ldots$
$= I/(1-d)^r$

The series CONVERGES!

For d=20%, r = 4: T=I * 2

Charge 2x initial storage cost, save half, store forever!

Simplified Equation

Total Cost (T) =

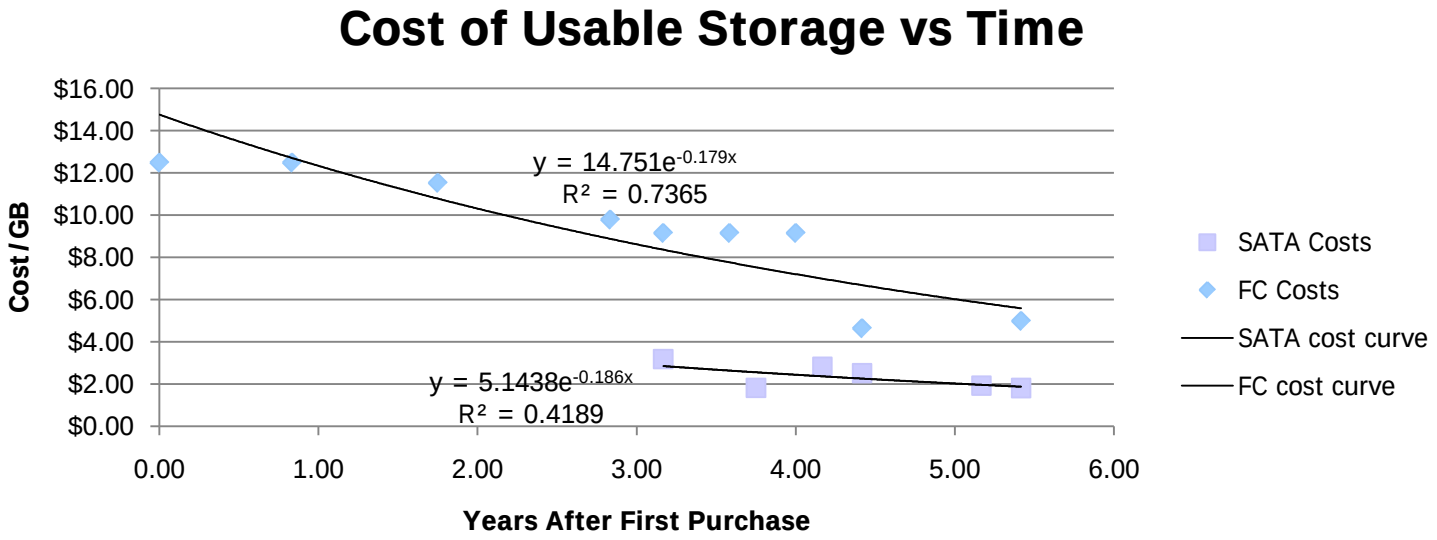
Initial Cost (I) * Storage Factor (S)

T = I * S

S is computed based on:

- Replacement cycle
- Rate of decrease in storage costs

An Example: DataSpace at Princeton



- FC costs decrease by about 16% per year
- SATA costs decrease by about 17% per year
- Additional savings every few years from new storage

The “Storage Factor ” for DataSpace at Princeton

- SATA cost = $1.81/gb
- Replace every four years
- Costs decrease by 20% year

S = 1.81/(1-.8 **4) = $3/gb

Adding tape backup jumps this to $5/gb

$5K one-time to store a terabyte forever

Objections to the Model

1. Not legal!
2. Breaks down if storage costs don't decrease rapidly or replacement cycle is short
3. Doesn't consider all costs
    1. People
    2. Electronics
    3. Meta-data (curation)
4. Too expensive!

3. Not legal

- Can’t squirrel-away NSF money to pay future costs
- NOT doing that
  - Researcher pays me for long-term storage
  - I can do whatever I want with the money
  - If you buy a freezer, vendor probably puts (some of the) money in the bank
- University operating as a vendor
- Real issue: must get account that rolls over

1. Cost/Replacement issues

- Model very robust to varying cost-decrease rates and replacement cycles:

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

1. Cost/Replacement issues

- People constantly predict that storage cost decreases will tail off.
  - In 1981, a Morrow Designs 10 megabyte drive cost $3,000, or $300,000 per gigabyte, or $300 million per terabyte
  - In 2000, an IBM 20 gigabyte drive sold for approximately $280, or $14,000 per terabyte
  - Can get a terabyte drive today for $100. A good one for $300.
  - No real indication that cost decreases will not continue

2. Doesn't consider all costs

- Studies: 5% of total data preservation costs for disk drives
- Bulk of cost is for "people" (staff)
- True only if people costs are added to every storage request.
- "Marginal" people costs decrease as quickly as disk drives.
- Staff 20 years ago->1 gig; today-> 1 petabyte

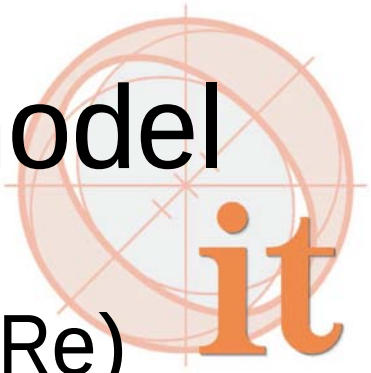
Mitigating model risks

- Must minimize ancillary costs
- Keep these to a minimum
  - Minimal
    boutique" services"
  - Customer pays for data curating or delivery if not web based
  - No re-use of disk space.
  - ALL data public; no special handling.

Mitigating model risks

- Handle fluctuation in media costs
  - Recompute S every year
  - If off, adjust S
  - Rob Peter to Pay Paul
  - How a bank operates
    - That 3% 40-year mortgage you gave out 30 years ago
- NOT a Ponzi scheme
- NOT an endowment model

DataSpace operational model



- Write Once, Read Endlessly (WoRe)
  - No storage re-use
  - Permanent URL; no deleting
  - Publicly accessible (quarantine)
  - Bit storage + meta-data
  - Fate of “owner” irrelevant
    - Repo has permanent right to store and distribute
- Critical to keeping costs manageable

4. Too expensive

- Most serious objection
- Can buy terabyte drive for $100
- Researchers have sysadmins who whisper in their ears: “we can do it for less”
- Don’t need to store data “forever”: why pay for that

Dealing with “too expensive”

- First, “forever” is not a factor
  - 90%+ of cost in for first 3 replacement cycles
  - Minimal difference between storing for 10 years and storing “forever”
- $5K/terabyte “forever” is too much if sysadmin can buy drive for $100
  - Does not factor in tape backup and replacement
  - Researcher probably doesn’t care
  - Ok if reviewers say it it ok

Why so expensive?

- University IT shops not missioned to provide cheap storage
  - Must provide high-speed storage for local applications
  - Must cost-recover other parts of IT organization
  - Cannot benefit from scaling
  - Cannot let people go (overstaffed)
  - People “have to” use their service

A business opportunity?

- Enterprising group of grad students
- Set up data center
- Save through scaling
- Offer POSE service
- Better than cloud?
  - Amazon S3: 1 Terabyte for 20 years =
  - **$30-$50K**
- **$5K one-time starting to look good**

Want to know more?

Excellent critique of model by David Rosenthal:

http://blog.dshr.org/2011/02/paying-for-long-term-storage.html

- **http://arks.princeton.edu/ark:/88435/dsp01w6634361k**
- **dataspace.princeton.edu**
- **serge@princeton.edu**
- **Are knives sharpened? Question?**