Preservation Challenges

October 2007

John Kunze, jak@ucop.edu

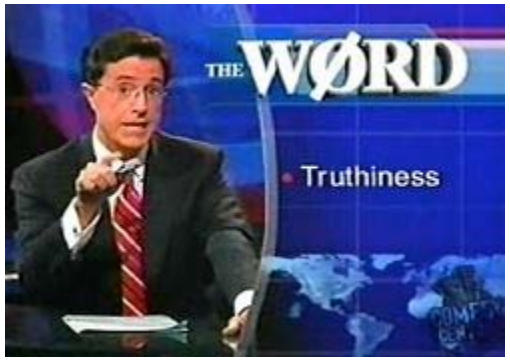
Preservation challenges: case studies

With benefit of hindsight, what's hard?

- Policy
- Making files small
- Fast data transfer
- Cheap, reliable storage
- Lots of annoying files
- What if we run out of money?
- Imagining the non-repository

What's digital preservation?

Storing digital objects while retaining a balance of usability and faithfulness (truthiness) to their creators' original intentions



Policy Challenges

- How faithful
- How long
- How many replicas
- How much manipulation
- Right(s)mare

Fast data transfer challenges

Lots of files, lots of data

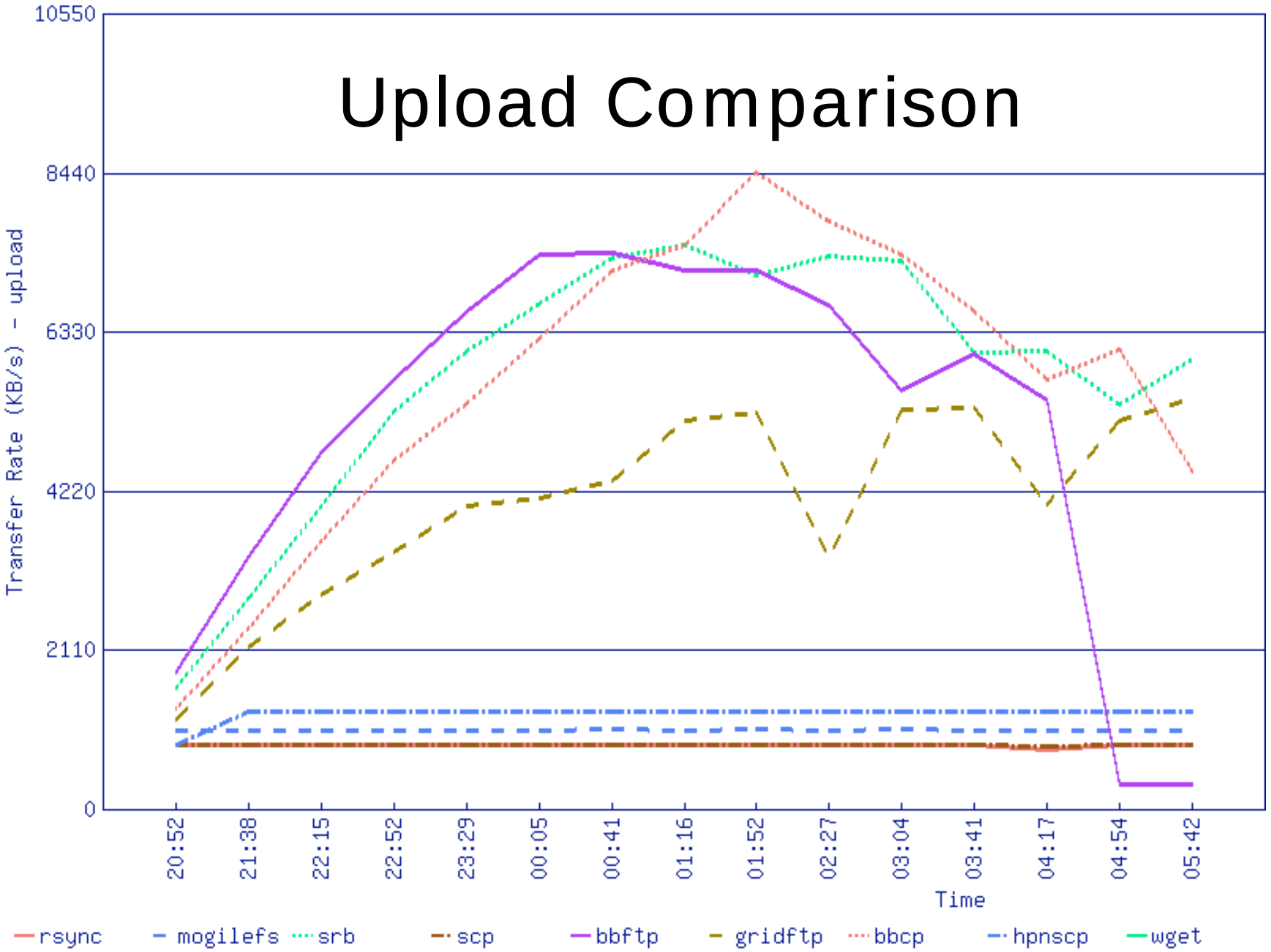
- Could take months to move and replicate

Survey tool performance and usability

Lesson: parallelism works

Transfer tools tested

- SRB (bundled Sget/Sput tools)
  http://www.sdsc.edu/srb/index.php/Main_Page
- GridFTP (hi security, from Grid community)
  http://www.globus.org/grid_software/data/gridftp.php
- MogileFS (simple distribute filesystem, Perl scripts)
  http://www.danga.com/mogilefs/
- High Performance SSH (no system gaming)
  http://www.psc.edu/networking/projects/hpn-ssh/
- BBFTP (easy installation and use)
  http://doc.in2p3.fr/bbftp/
- BBCP (easy installation and use)
  http://www.slac.stanford.edu/~abh/bbcp/
- RSYNC, SCP, SFTP, FTP (ubiquitous)


Upload Comparison
10550
8440
6330
4220
2110
0
Transfer Rate (KB/s) - upload
20:52
21:38
22:15
22:52
23:29
00:05
00:41
01:16
01:52
02:27
03:04
03:41
04:17
04:54
05:42
Time
rsync
mogilefs
srb
scp
bbftp
gridftp
bbcp
hpnscp
wget

Making files small

Case: page image compression for mass book digitization

Conclusion: Full-color JP2K, lossy compression 60:1 works!

Cheap, reliable storage

- RAID (Redundant Arrays of Inexpensive Disk) 1980s
- JOBD (Just a Bunch of Disks) 1990s
- MAID (Massive Arrays of Idle Disks)

Lots of annoying files

Case: web archiving

- AIHT
- W/ARC

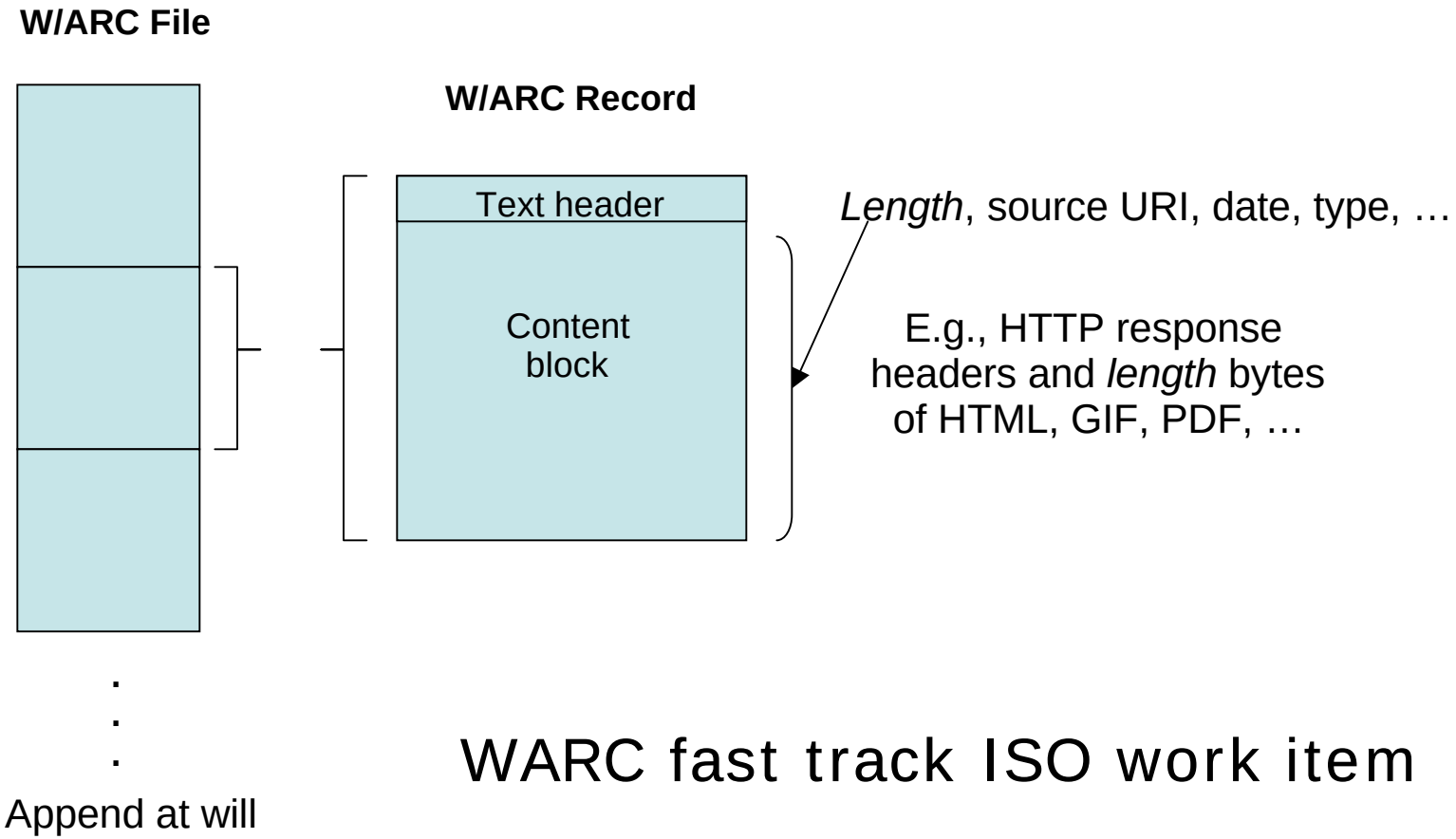
Generalization: aggregate file format

Records are sort of peers of files

  - Many “files” in one file for speed and ease

W/ARC File Anatomy

WARC = Web ARChive file format



WARC fast track ISO work item

What if we run out of money?

One Narrow Case:

... Data Desiccation, creating no-frills, sometimes feature-poor derivatives that retain most of the original scholarly value but are likely to be less perishable than original format

Save desiccated derivatives along *with* original, just in case there are

- No funds to touch files
- No expertise to convert them properly

Example

Photo of Mission San Luis de Tolosa [2]About the City
[3]Visiting SLO
[4]What’s New
[5]City Government
[6]Employment Opportunities
[7]Bids & Proposals
[8]Economic Development
[9]FAQs
[10]How are we doing?
City of San Luis Obispo
About the City

[Choose a Destination....]
[11]Search [12]Contact Us [13]City Home
A Brief History

Who we are and how we got started. The City of San Luis Obispo serves as the commercial, governmental and cultural hub of California’s Central Coast. One of California’s oldest communities, it began with the founding of Mission San Luis Obispo de Tolosa in 1772 by Father Junípero Serra as the fifth mission in the California chain of 21 missions. The mission was named after Saint Louis, a 13th Century Bishop of Toulouse, France. (San Luis Obispo is Spanish for "St. Louis, the Bishop".) It was first incorporated in 1856 as a General Law City, and became a Charter City in 1876.

Where we’re located. With a population of 44,000, the City is located eight miles from the Pacific Ocean and is midway between San Francisco and Los Angeles at the junction of Highway 101 and scenic Highway 1. San Luis Obispo is the County Seat, and a number of federal and state regional offices and facilities are located here, including Cal Poly State University, Cuesta Community College, Regional Water Quality Board and Caltrans District offices. The City’s ideal weather and natural beauty provide numerous opportunities for outdoor recreation at nearby City and State parks, lakes, beaches and wilderness areas.

Great place to live and visit. While San Luis Obispo grew relatively…

Endnotes

…
[18]About the City | [19]Visiting SLO | [20]What’s New | [21]City Government
| [22]Employment
[23]Bids & Proposals | [24]Economic Development | [25]FAQs | [26]How are we doing?
[27]©2006, City of San Luis Obispo

References

1. http://www.ci.san-luis-obispo.ca.us/briefhistory.asp#content
2. http://www.ci.san-luis-obispo.ca.us/about.asp
3. http://www.ci.san-luis-obispo.ca.us/visit.asp
4. http://www.ci.san-luis-obispo.ca.us/whatsnew.asp
5. http://www.ci.san-luis-obispo.ca.us/government.asp
6. http://www.ci.san-luis-obispo.ca.us/humanresources/index.asp
7. http://www.ci.san-luis-obispo.ca.us/finance/bids.asp
8. http://www.ci.san-luis-obispo.ca.us/economicdevelopment/index.asp
9. http://www.ci.san-luis-obispo.ca.us/faq.asp
10. http://www.ci.san-luis-obispo.ca.us/how.asp
11. http://www.ci.san-luis-obispo.ca.us/search2.asp
12. http://www.ci.san-luis-obispo.ca.us/contact.asp
13. http://www.ci.san-luis-obispo.ca.us/index.asp
14. http://www.ci.san-luis-obispo.ca.us/visit.asp

…

Imagining the non-repository

The repository’s deadly embrace

Filesystems vs databases