



The Cornell Experience: arXiv.org

Anne R. Kenney

DCC/DPC Workshop on Cost Models for
Preserving Digital Assets 26 July 2005

CORNELL



How do we identify the
requisite resources to
maintain a digital
repository?



"If we had some horses, we'd have a cavalry – if we had some men."

Victor Mature in *Demetrius and The Gladiators*, 1954



Step 1. Identify Cost Categories

Cost categories

- Startup costs
- Ongoing costs
- Contingency costs



Step 2. Identify Cost Centers

- Capital costs
 - Technical/procedural infrastructure
- Direct operating costs
 - By OAIS functions, e.g., ingest, data management, storage, preservation planning
- Overhead
 - Indirect operating expenses, e.g., facilities, general/administrative/support services



Step 3. Calculate Costs

- System(s)
- Staff
- Ancillary staff expenses
- Services and fees
- Supplies and materials
- Other



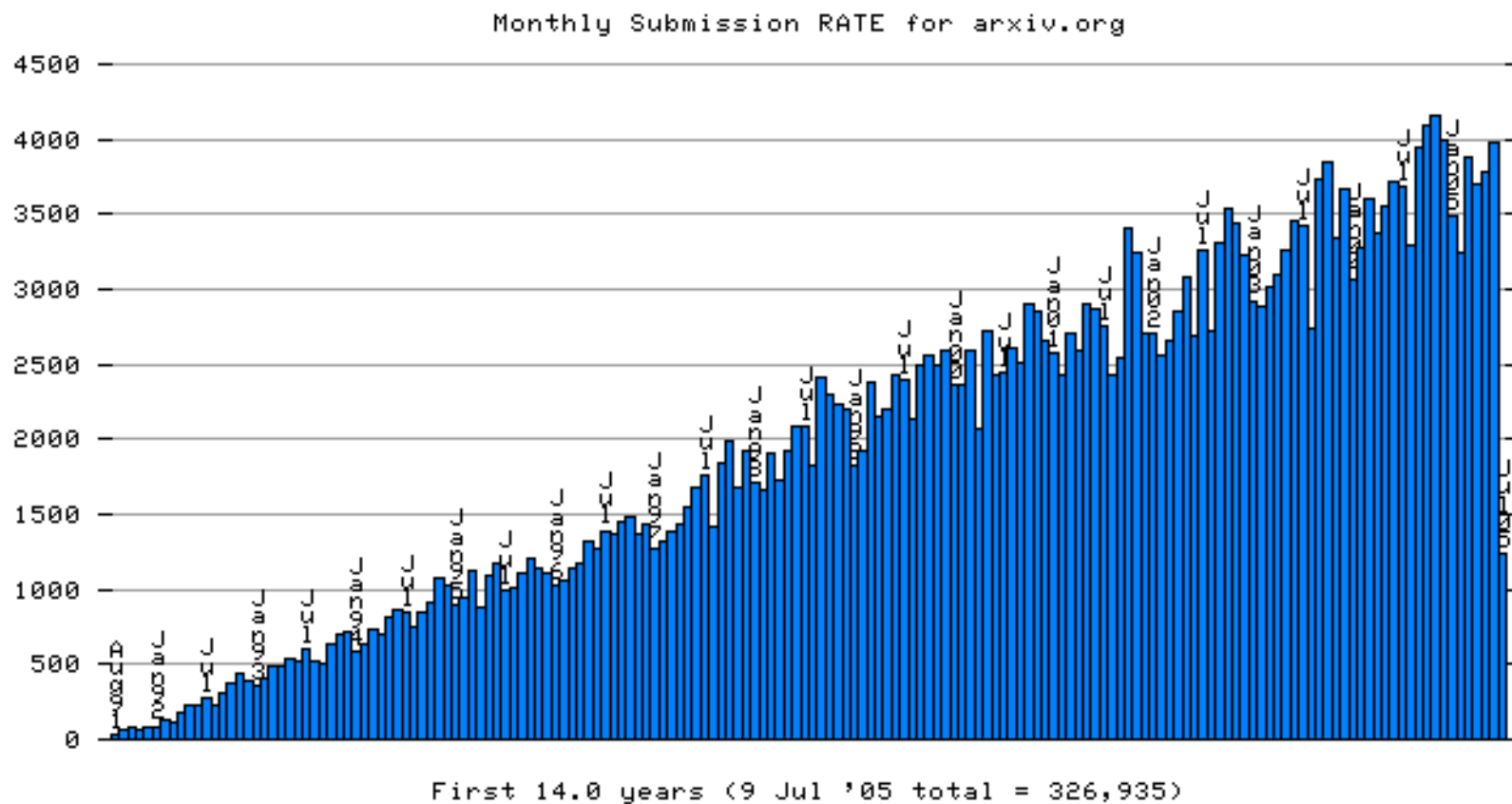
Case Study: arXiv.org

- automated electronic archive and distribution server for research papers in physics, cs, math
- Begun in 1991; transferred to CU Library in fall 2001
- 326,000 items by FY2004/5; 45,000 submissions this year
- Cost/submission: \$1-\$5

arXiv.org Main Page



Monthly Submission Rate



Hourly Use Data of arXiv.org

arXiv web server usage for 21 Jul '05 (arxiv.org site only)

Hour_Total__Connections

00	_11861_		11861
01	_21946_		10085
02	_33240_		11294
03	_46439_		13199
04	_60527_		14088
05	_73932_		13405
06	_87277_		13345
07	_99237_		11960
08	_112155_		12918
09	_126827_		14672
10	_141809_		14982
11	_157772_		15963
12	_170933_		13161
13	_183045_		12112
14	_193269_		10224

Total number of connections = 193269 (+72 local & administrative connections)

Current local time is *Thu, 21 Jul 05 14:56:06 EDT*



Capital Equipment

Annual equipment cost =

- purchase price of hw/sw
- amortization rate (3-5 year range)
- annual maintenance, licenses, and development fees, estimate at 20%-50% purchase price



arXiv Server Costs

- Equipment cost: \$10,600
- Amortization: 5 years
- Annual fees: \$2,403
- Annual equipment cost =
 $\$4,523 (\$10,600/5 + \$2,403)$



Equipment costs/ preserved digital object

- Calculated by dividing the annual equipment costs by the current database and estimated growth rate of digital objects.
- The greater the number of similar digital objects, the lower the per/object equipment cost.
- The more streamlined/automated the processes, the cheaper the cost.



ArXiv Annual Equipment Costs (FY2004/05)

- 326,000 submissions
- 45,000 in FY 2004/5
- $\$4,523/45,000 =$
\$ 0.10/submission
- $\$4,523/326,000 =$
\$.01/item/year



Calculate Personnel Costs

- Staff performing specific tasks, management overhead, ancillary expenses
- Calculate staff costs on “weighted hourly rate” not salary, which is over twice the hourly wage



Weighted Hourly Rate

- 1) Avg. number of workdays/
year
- 2) Number of hours/day
- 3) Number of productive hours
- 4) Weighted hourly rate =
$$\frac{\text{salary} + \text{fringe}}{\text{productive hours}}$$



Calculating The “Real” Costs at Cornell

Staff:

- 222 day work year (excluding vacation, sick, personal and holidays)
- 7.3 hour work day (excluding breaks)
- 1,621 hours/year at work
- Assume 75% “production time”: 5.5 hours/day and 1,216 hours/year
- “Weighted hourly rate” =

salary + fringe

1,216



ArXiv: Programmer/ Analyst Band F

- Annual hours worked: 1,216
- Salary: \$57,000
- Benefit rate: 32%
- Hourly wage: \$28
- "Weighted hourly rate":

$$\frac{\$57,000 \times 1.32}{1,216} = \$62$$

Weighted rate is 2.2 times the hourly wage



Management and Ancillary Costs

- % of management overhead assigned/staff member
- Annual ancillary expenses (min: \$4,000/year)
 - training, travel, workstation support, supplies, phone/data, reference materials
- Overhead



Example: Full Freighted Programmer/Analyst

\$62.00 (WHR)

+ \$ 6.85 (8% mgr overhead)

+ \$ 2.32 (ancillary expenses)

\$71.70

x \$ 1.58 (overhead)

\$113.29



Calculate Services Costs

- Annual or contractual fees paid for out-sourced functions, including backup, space in server farm, mirroring
- Related services & expenses, e.g., technology monitoring, subscription to software repository, consortial fees, royalties, memberships



Calculate Contingency Costs

- Unanticipated expenses associated with trigger events, catastrophes, technical sea changes
- Varies depending on size of effort, complexity, experience



Calculate Overhead Costs

- Negotiated rate vs. itemized expenses
- Recovered or absorbed
- How much comes back to the digital repository



ArXiv Annual Costs for Maintenance/Preservation

- Equipment: \$4,523
- Staff (2 FTE): \$150,480
- Mgt overhead: \$15,000 (8%)
- Ancillary staff costs: \$8,640
- Contingency: \$17,568
- Overhead (58%): \$112,084
- Total: \$305,333/year



Total Costs

- 326,000 submissions; 45,000 in FY 2004/05
- $\$305,333 / 45,000 = \$6.79/\text{submission}$
- $\$305,333 / 326,000 = \$0.94 / \text{item/year}$
- $\$305,333 / 163 \text{ GB} = \$1,873/\text{GB}$



Maintaining/Preserving ArXiv

- Staff costs represent the greatest expense
- Present day costs do not equal future costs
- Total costs will decrease only with automation of human effort
- Per submission and per item costs will continue to decrease as the collection grows



"Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate."

Cliff Lynch, 2003



"They say you can't do it, but remember, that doesn't always work."

Casey Stengel