Libraries and Large Data

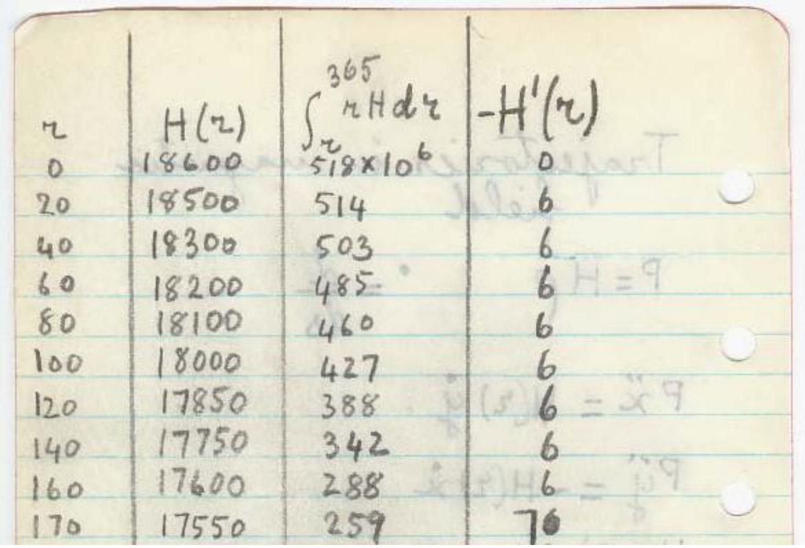
Super Computing 2012

Elisabeth Long
University of Chicago Library

What is the Library's Interest in Big Data?

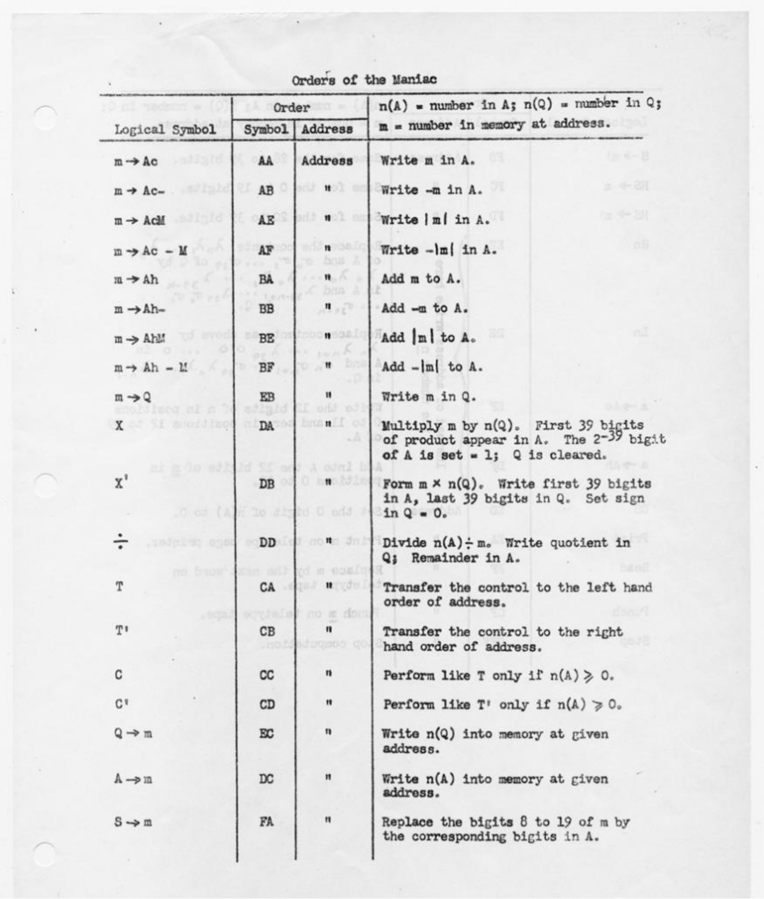
We’ve Always Collected Data

- Intellectual output of our faculty
  - Ongoing usefulness of data
  - Historical record of activity
- Preserving = keeping indefinitely



r	$H(r)$	$\int_r^{365} rH\,dr$	$-H'(r)$
0	18600	518×10^6	0
20	18500	514	6
40	18300	503	6
60	18200	485	6
80	18100	460	6
100	18000	427	6
120	17850	388	6
140	17750	342	6
160	17600	288	6
170	17550	259	76

Trajectories in Magnetic Field. Notebook E3, Enrico Fermi Collection



Orders of the Maniac

Logical Symbol	Order		n(A) = number in A; n(Q) = number in Q; m = number in memory at address.
	Symbol	Address	
m → Ac	AA	Address	Write m in A.
m → Ac-	AB	"	Write -m in A.
m → AcM	AE	"	Write |m| in A.
m → Ac - M	AF	"	Write -|m| in A.
m → Ah	BA	"	Add m to A.
m → Ah-	BB	"	Add -m to A.
m → AhM	BE	"	Add |m| to A.
m → Ah - M	BF	"	Add -|m| to A.
m → Q	EB	"	Write m in Q.
X	DA	"	Multiply m by n(Q). First 39 bigits of product appear in A. The 2^{-39} bigit of A is set = 1; Q is cleared.
X'	DB	"	Form m × n(Q). Write first 39 bigits in A, last 39 bigits in Q. Set sign in Q = 0.
÷	DD	"	Divide n(A) ÷ m. Write quotient in Q; Remainder in A.
T	CA	"	Transfer the control to the left hand order of address.
T'	CB	"	Transfer the control to the right hand order of address.
C	CC	"	Perform like T only if n(A) ⩾ 0.
C'	CD	"	Perform like T' only if n(A) ⩾ 0.
Q → m	EC	"	Write n(Q) into memory at given address.
A → m	DC	"	Write n(A) into memory at given address.
S → m	FA	"	Replace the bigits 8 to 19 of m by the corresponding bigits in A.

Computer program for calculating cyclotron orbits, Enrico Fermi 1951

New Drivers

- New data management requirements from funding agencies
- Need for infrastructure to manage data preservation after end of research project
  - Relationship between data preservation and data sharing

Best Practices for Data Management

- Data collection & organization
  - Unique identifiers, standardized formats, file naming schemes, etc.
- Quality control and assurance
  - Data collection workflows, automated data checks
- Metadata
  - Describing: digital context, stakeholders, scientific context, parameters, data
  - Use of structured metadata, identifying existing standards
- Workflows
  - Documenting the path from raw data to final product
- Data stewardship and reuse
  - Trustworthy digital repository checklist, data citation practices

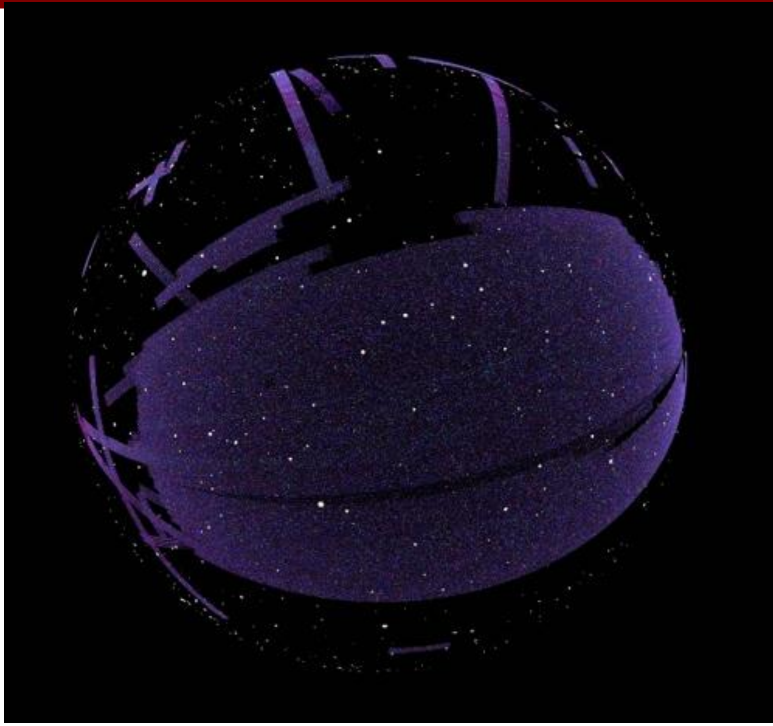
Data Preservation: Case Study

Preserving the Sloan Digital Sky Survey

An SDSS – Library Collaboration

SDSS Archiving Project Context

- SDSS-I & II was an 8 year, inter-institutional astrophysics project to image ¼ of the sky
  - Images of more than 350 million objects
  - Over 3.4 billion rows in the SQL database
- 2 years before end of project funding, approached UChicago library to talk about archiving the project. (Had project money remaining to pay for archiving)



SDSS Data Capture

The SDSS Telescope at Apache Point, NM

CCD Camera: to **capture images** of thin strips of the sky (later tiled)

Plug plate: Fiber optics to **capture spectra** of “interesting” objects

SDSS Data Processing

“The task of data managers for the SDSS is to take digitized data - the pixels electronically encoded on the mountaintop in New Mexico - and turn them into real information about real things.”

Data Pipelines process the raw telescope data to

- extract information such as position, size, shape, redshift, wavelength, etc.
- correct for atmospheric and other anomolies

SDSS Data Sharing

- Annual Data Releases
  - 1 year embargo to give priority access to SDSS members
- Multiple Data Interfaces
  - Public Oriented:
    - **SkyServer** (includes featured images, lesson plans, projects)
  - Scientist Oriented:
    - **Catalog Archive Server (CAS)**
      - SQL Database with web-based query interface
      - CASJobs (login required) for saved datasets and running long jobs
    - **Data Archive Server (DAS)**
      - Access to flat files of all the data for download (wget, curl, rsync)
- Help Desk

SDSS Data Sharing: A Success Story

- How/Why did it work?
  - Embargo Period: balances collaborator interests with public interests
  - Clear rules for authorship and citation
  - Extensive documentation on how to understand and use the data
- Results
  - Publications based on SDSS Data include more than…..
    - 70 PhD theses
    - 2000 articles in refereed journals
    - 70,000 citations
  - Galaxy Zoo: citizen scientist website

What to Archive?

- Raw data
- Processed data
  - Note: data was often reprocessed to fix something discovered down the road – how many copies to keep?
- Data pipeline algorithms
- Interfaces
  - **SkyServer:** Web Interface will age, links rot
  - **CAS**: SQL Database/Web Interface will age and need migrating
  - **DAS**: Flat files more straightforward
- Administrative archive
  - Documents the history of the project and many decisions
- SDSS email listservs
  - Includes a lot of explanation of what was done to the data – especially when things were re-processed

Elements of the Archive Project

- **5 year MOU** with Astrophysical Research Consortium (ARC) which oversees SDSS
  - Collaboration between UChicago, Fermi Lab, and Johns Hopkins
- SDSS's **administrative records** went to UChicago Archives. (Mix of paper and electronic). Important for documenting what the data is and how it was created
  - The "Project Book", Meeting minutes, Policies, Manuals, Email listservs, Project website, etc
  - Copy of email listerv at Johns Hopkins
- **DAS** running at UChicago and FermiLab
  - Live server running at Fermi with manual switchover to UChicago as needed
  - UChicago committed to archiving of DAS in perpetuity
- **CAS** running at Uchicago, FermiLab and Johns Hopkins
  - Database and hardware nearing end of life. Migrate? or time to cut CAS loose?
- **Virtual Help Desk** run at UChicago Library, with group of astrophysicists "on-call" behind the scenes to answer complex questions

SDSS Takeaways

- “Archiving” may really mean “keep project going”
  - How to plan for after the funding runs out
  - Desire to continue to serve complicated interfaces to data
- Importance of ARC
  - Infrastructure that persists after SDSS project closed
- Context, Context, Context
  - Data archiving includes A LOT more than just the data
  - SDSS-II was extremely well-run project = critical to the success of the archiving project
    - Data well described
    - Structured data release schedule with documentation of what had been done to data

SDSS Takeaways

- The best laid plans of mice and men….
  - Even with a well-organized project we have just discovered we are missing one set of data from the DAS
- Need to negotiate different levels of support for different types of data
  - Long-term **archiving** for the DAS
  - Near-term **serving** for the CAS
  - 5-year MOU with chance to review and revisit serving decisions
  - Use statistics to aid decision-making

Next Steps

- It ain’t over ‘till it’s over
  - SDSS-III was funded and is reprocessing the SDSS-II data
  - Archiving discussions have begun with SDSS-III (even though UChicago not a participant in SDSS-III)
- Researchers move on…
  - UChicago researchers now participating in the Dark Energy Survey (DES)
- …And the data keeps growing
  - DES images are 1GB each. 400 to be taken each night...