

Delivering Unique Numeric Data on the Web

(Projects, Platforms, and Preservation)

Ronald C. Jantz
Government & Social Sciences Data Librarian
Scholarly Communication Center
Rutgers University Libraries

Delivering Unique Numeric Data on the Web

- Introduction – Perspectives and Issues
- Digital Projects and the Scholarly Communication Center
- Projects, Platforms and Preservation

Perspectives and Issues

- Re-usable platforms (either technology or process) can dramatically reduce development time and improve quality.
How do we establish and sustain re-usable platforms in an academic environment?
- Digital preservation
A scenario: A truck loaded with hazardous waste is headed toward a dump site. Will our descendants know where we have buried the waste?
- Unique projects: Those that have specific relevance to Rutgers University and New Jersey

The Scholarly Communication Center

(Rutgers University Libraries)

Goals

- Allow scholars to apply state-of-the-art technology
- Teach and demonstrate the latest electronic tools
- Share the resources of the library

The SCC has given us an opportunity to experiment and innovate.

Environment in the SCC

- A Windows2000/NT Network
- A Social Sciences Data Center (with 12 workstations)
- A digital preservation laboratory (under construction)
 - Large network mass storage (terabytes)
 - Scanners, including large format scanner (40 inch wide) & digital camera
 - Large format printer
 - Image compression software (e.g. djvu from AT&T & LizardTech)
- Staff
 - 2 resident librarians, 3 staff and a staff manager
 - On average, 10 part-time students
- Work areas for special projects

SCC Project Goals

- Develop platforms that can be quickly learned by students and part-time employees.
- Encourage re-use to improve quality, reduce development time, and facilitate training.
- Establish project classes for reusable platforms: *directories of people, reference databases, image archives, numeric data, online surveys.*
- Define end-to-end processes for access and preservation

A Sampling of SCC Projects

Databases/Archives on the Web

- The Alcohol Studies Database (with the Center for Alcohol Studies)
A reference database at: http://www.scc.rutgers.edu/alcohol_studies
- The New Jersey Environmental Digital Library (with NJ DEP)
An image archive at: <http://njenv.rutgers.edu/njdlib>
- Medieval Early Modern Data Bank (with History Department)
Numeric data at: <http://www.scc.rutgers.edu/memdb>
- Public Opinion Data (with Eagleton Institute)
Numeric data at: <http://www.scc.rutgers.edu/eagleton>
- For more – see “Digital Projects” at <http://www.scc.rutgers.edu>

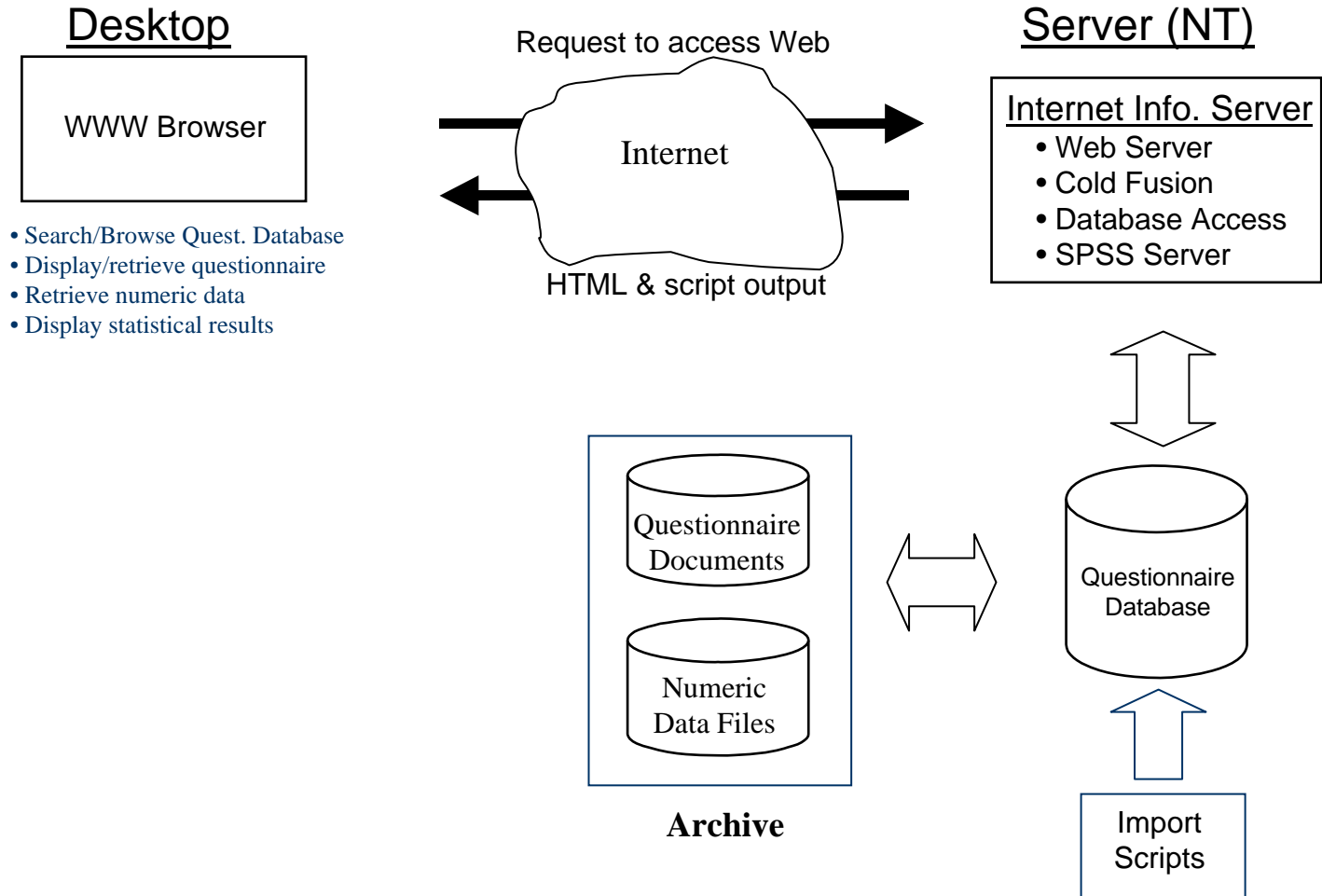
Eagleton Public Opinion Polls

(Delivering numeric data on the web)

Characteristics:

- Prototype at: http://www.scc.rutgers.edu/eagleton_tst
- Content: New Jersey public opinion (1971 -)
- Frequency: four polls per year
- Access: public domain
- Compiler: Eagleton Institute
- Owner: Eagleton/Star Ledger
- Archiver: RUL/Scholarly Communication Center
- Type: database on the Web
- Format: html, pdf, portable spss files, MS-Access, ColdFusion/SQL

Eagleton Project Architecture



Eagleton Poll Archive - SEARCH RESULTS



[Eagleton Institute](#) [SCC](#) [Help](#) [About](#) [Feedback](#)

[Home](#)

[Search](#)

[Browse](#)

[Fund of NJ](#)

POLL NO.	QUEST. NO.	POLL DATE	QUEST. TEXT
121	qc1a1	1999	I'd like to get some idea of how you feel ...
124	qed3	1999	Rutgers is the State University of New Jersey. Compared to ...
124	qed2	1999	From what your experience and what you have heard, how ...
124	qed1	1999	Do you happen to know whether Rutgers is a private ...

Home

Search

Browse

Results

Frequency

X-Table

Questionnaire

Request File

Fund of NJ

Title: State Issues and Presidential Candidates

Poll Number: 121

Question Number: qc1a1

Question Text: I'd like to get some idea of how you feel about different organizations or institutions that exist to serve you and others. For each of the institutions or groups I mention such as... (STARTING POINT) tell me whether you have a lot of confidence IN THEM, SOME CONFIDENCE IN THEM, OR NOT VERY MUCH CONFIDENCE. By confidence I mean that you feel they are doing what they ought to be doing. (RANDOMIZE ITEMS ON EACH FORM) Rutgers University?

Topic Keywords: presidential scandal; NJ government; presidential candidates; gambling; confidence in institutions; sports/nets/newark

Date: 01/07-01/13 1999

Description: Eagleton Poll #121 - January 1999. File is available in .por format - see location. File size is 389KB.

Sponsor: Star Ledger Eagleton Poll.

Respondent Type: General Population of New Jersey, adults 18 and over

Total N: 400

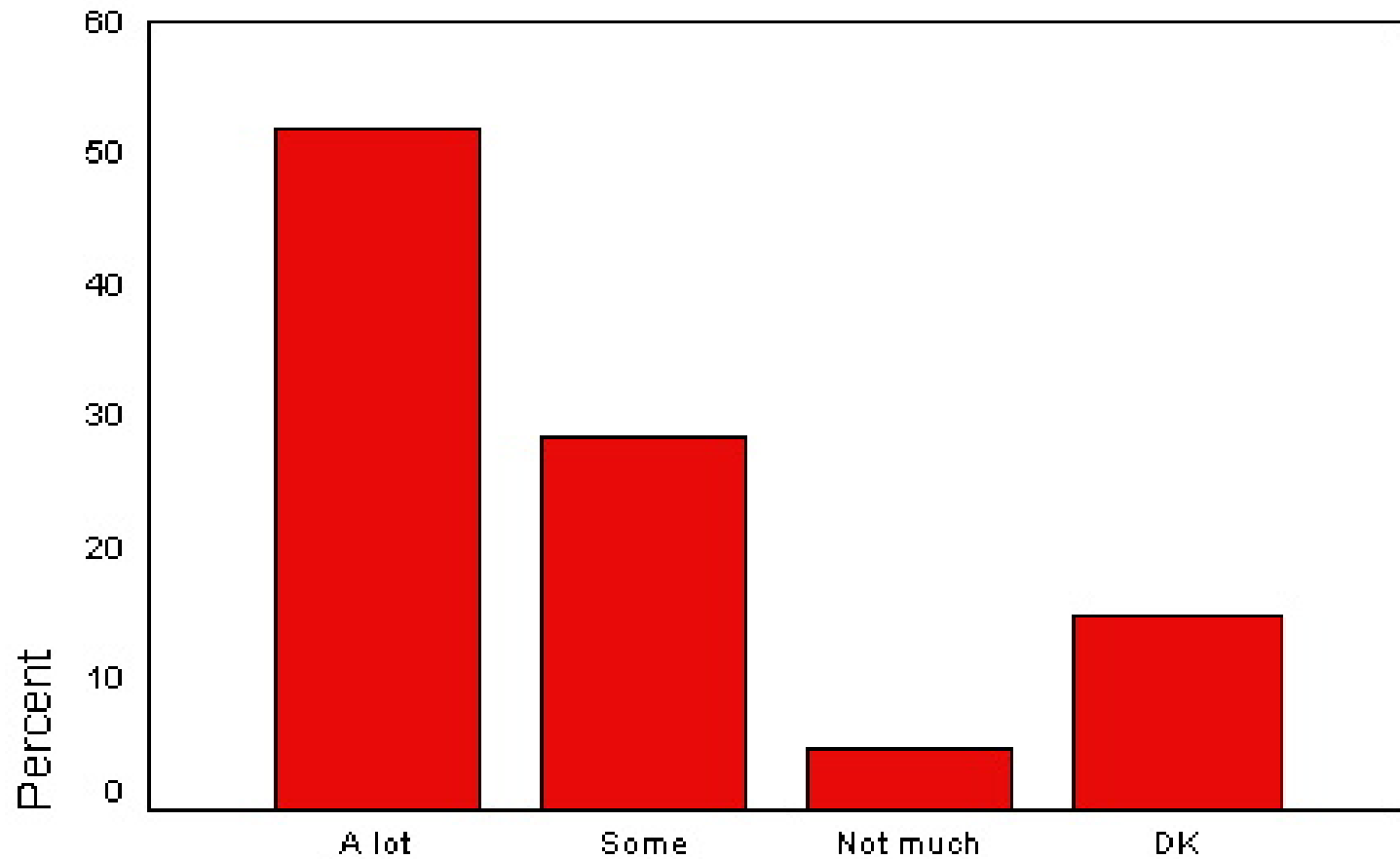
Frequencies
RUTGERS
UNIVERSITY

N	Valid	397
	Missing	397

RUTGERS UNIVERSITY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not much	19	2.3	4.7	4.7
	DK	60	7.5	15.0	19.7
	Some	112	14.2	28.3	48.0
	A lot	207	26.0	52.0	100.0
	Total	397	50.0	100.0	
Missing	System	397	50.0		
Total		795	100.0		

RUTGERS UNIVERSITY



RUTGERS UNIVERSITY

Cases weighted by SLCOSNWT

The Challenges of Digital Preservation

- Lack of standards (*or too many standards*)
- Lack of documentation on production and use
- Cost and rapid obsolescence of technology
- Impermanence of the medium
- Content easily changed – legal issues
- Version control
- Need to guarantee integrity of digital information
- Migration of information (*driven by external factors*)

Archiving Eagleton Poll Data

- In addition to daily and offsite backups,
- We are archiving essential data in the least device and software dependent format.
- Objective: to be able to regenerate the website in another hardware and operating system environment (perhaps in another technological epoch).

Eagleton Polls – What is to be Archived?

Archived Unit	Presentation Format	Preservation Format
1. Website	HTML, coldfusion, sql	Ascii text
2. Questionnaires	Adobe PDF	Ascii text
3. Ref. Database	MS-Access	Ascii text
4. Numeric Data	Spss export	Ascii text (data & syntax)
5. Processes	Readme (ascii text)	Ascii text
6. Metadata	HTML	Ascii text

Preservation Metadata for Digital Collections*

Collection – Eagleton Public Opinion Polls - Questionnaires

1. Persistent identifier:
2. Date of creation:
3. Structural type: ascii text
4. Technical infrastructure: 130 files in ascii text format, one file for each poll
5. File description
6. System requirements:
7. Installation requirements:
8. Storage information:
9. Access inhibitors:
10. Access facilitators:
11. Preservation action permission:
12. Validation: (information about validation mechanism)
13. Relationships (to other objects):
14. Quirks: (any characteristic that may cause loss in functionality)
15. Archiving decision (work):
16. Decision reason (work):
17. Institution responsible for archiving decision:
18. Archiving decision (manifestation):
19. Decision reason (manifestation):

* (from National Library of Australia: <http://www.nla.gov.au/preserve/pmeta.html>)

Summary: What are we learning?

- To take full advantage of platform technology, we need to formalize re-use – processes, platform components and training:

reference databases, numeric data, online surveys, digital archives, and directories.

- For numeric data, we should be able to quickly extend usage beyond the researcher to those who don't normally have access to data.
- End-to-end process definition is critical, especially for successful long term preservation.