Recovering Data of Forensic Interest

A report on challenges confronted by the National Archives and Records Administration

Prepared for the Computer Forensics and Cultural Heritage Meeting

University of Maryland,

May 14, 2010

Heritage Preservation "For the Life of the Republic…"

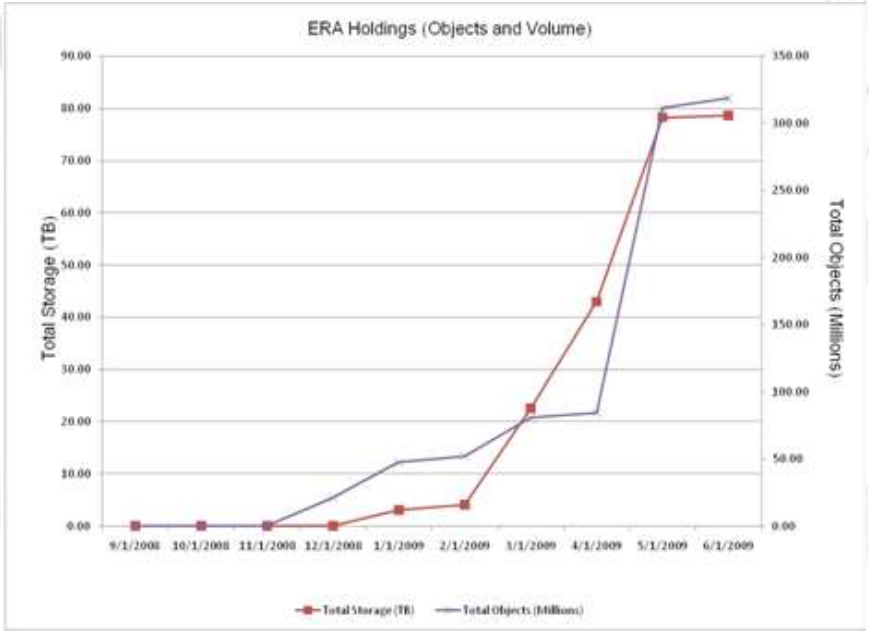
"The National Archives and Records Administration serves American democracy by safeguarding and preserving the records of our Government, ensuring that the people can discover, use, and learn from this documentary heritage. We ensure continuing access to the essential documentation of the rights of the American citizens and the actions of their government. We support democracy, promote civic education, and facilitate historical understanding of our national experience…"

Preserving the digital heritage

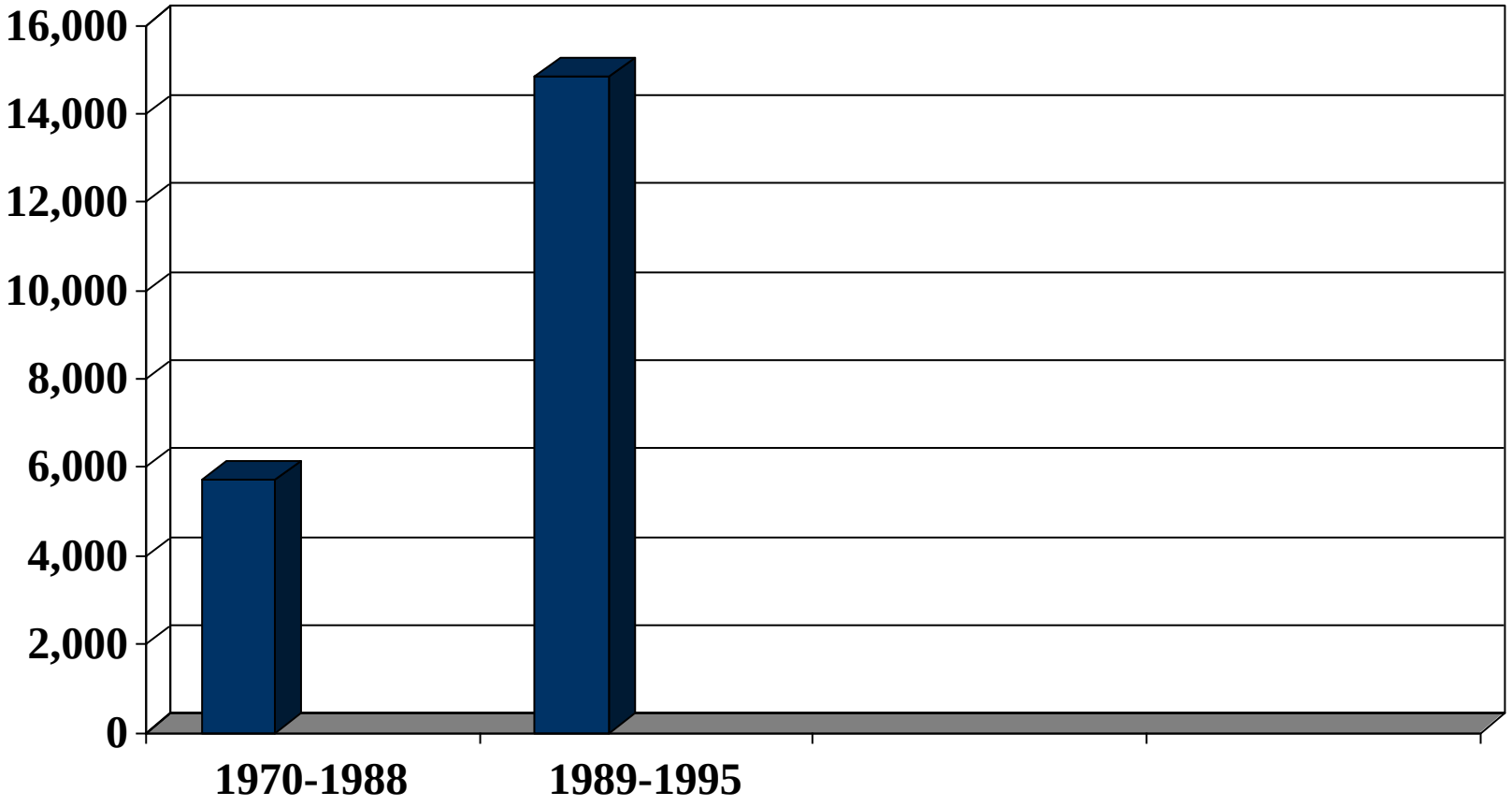
The Electronic Records Archives (ERA)
is a comprehensive, systematic, and dynamic means for preserving electronic records that will be free from dependence on any specific hardware or software and will improve preservation of , and access to, electronic records into the future.

“Digital Tsunami” of data with forensic interest

Between FY 2007 – FY 2022, the annual average yearly transfer volume to ERA is expected to grow to 54.95 Petabytes (PB) (equivalent to 300 trillion pages of printed text). During the same time period, the volume of records accumulated and stored within ERA will grow to 227 PB (56.75 trillion pages). ERA will be loaded with electronic records from Federal agencies, Congress, and Presidential administrations. Currently there are approximately 80 terabytes of data, stored in ERA.

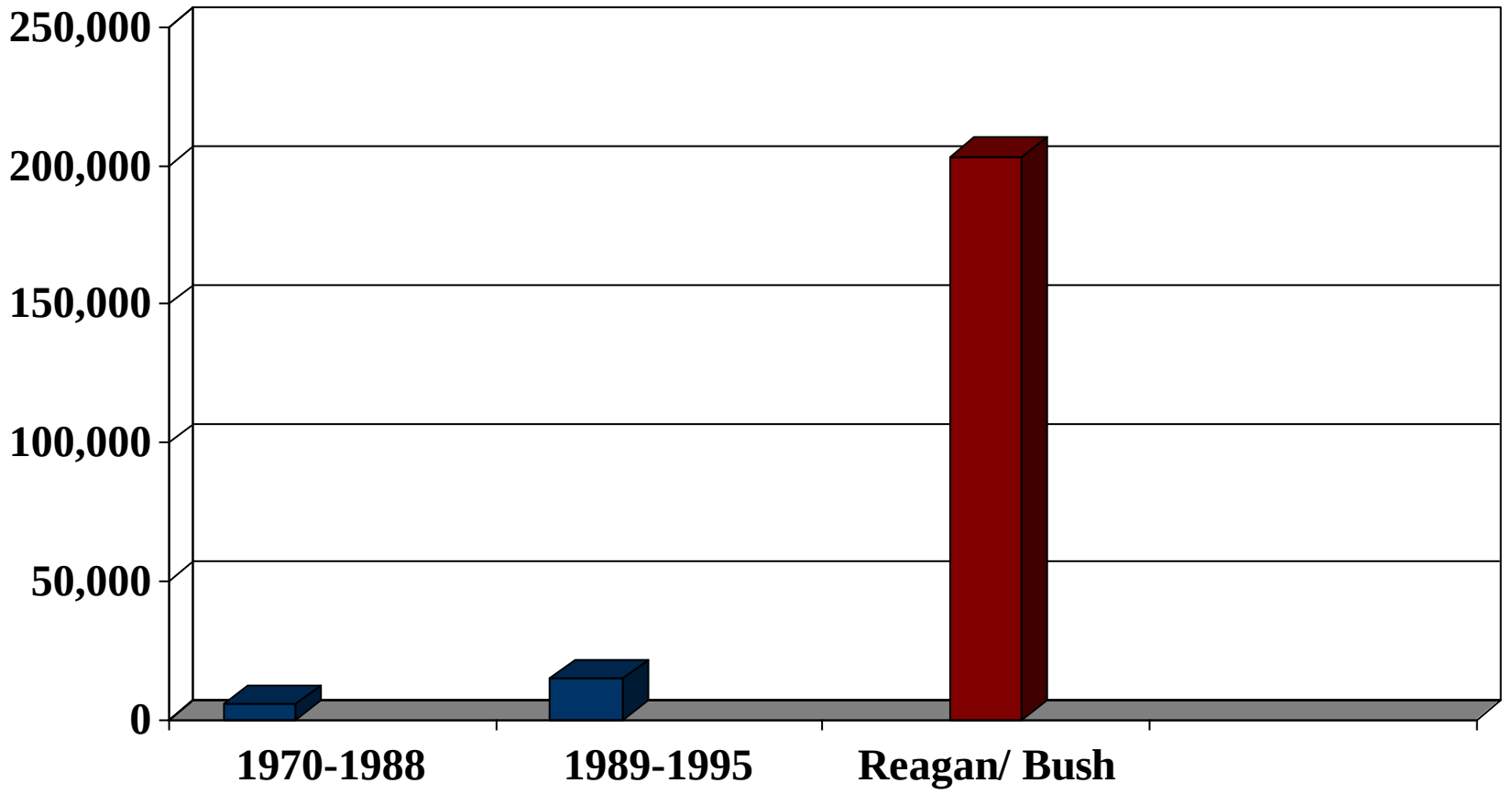


Transfers of Digital Files to NARA



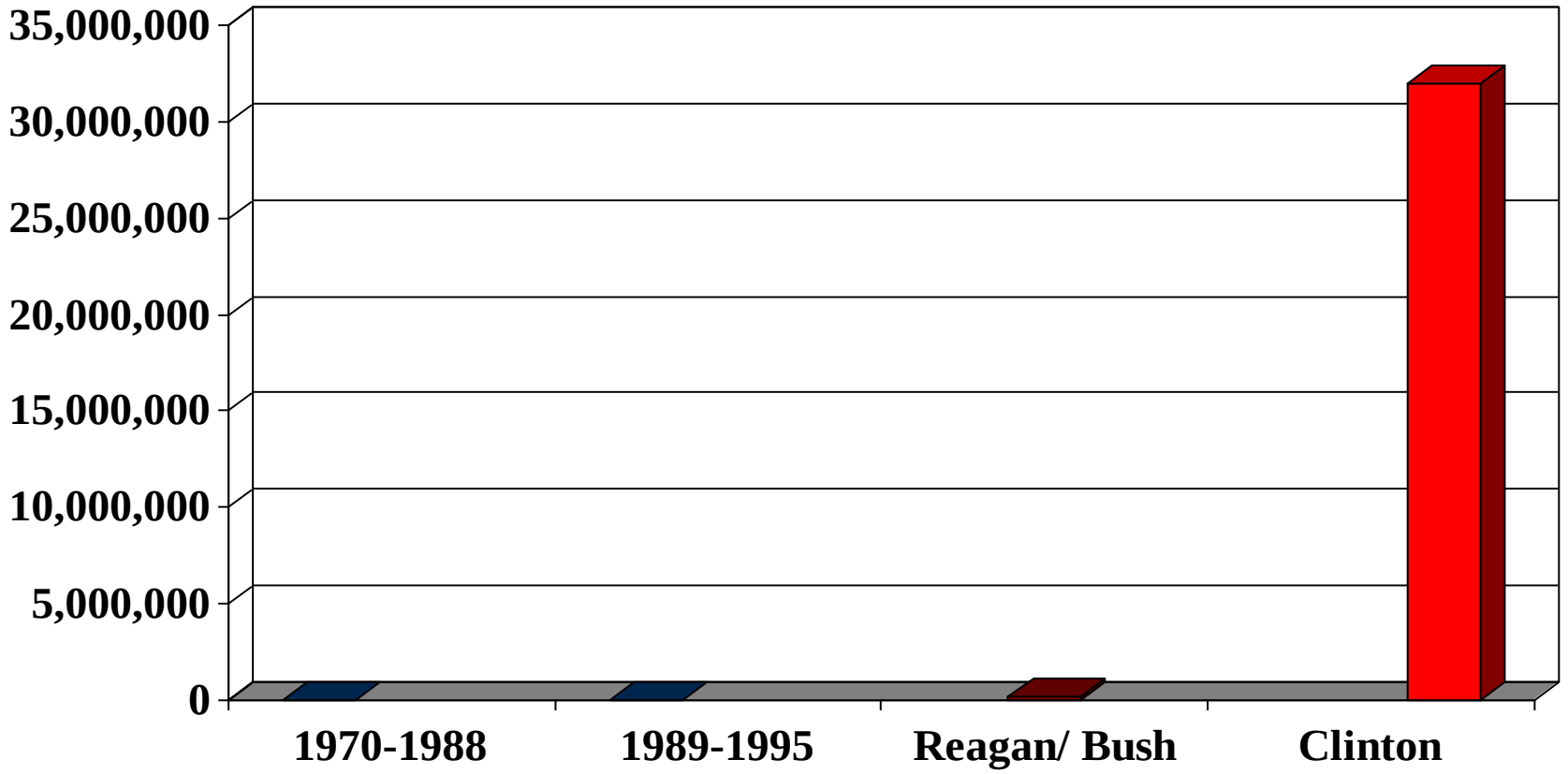
Total files, in terabytes

Transfers of Digital Files to NARA, in terabytes



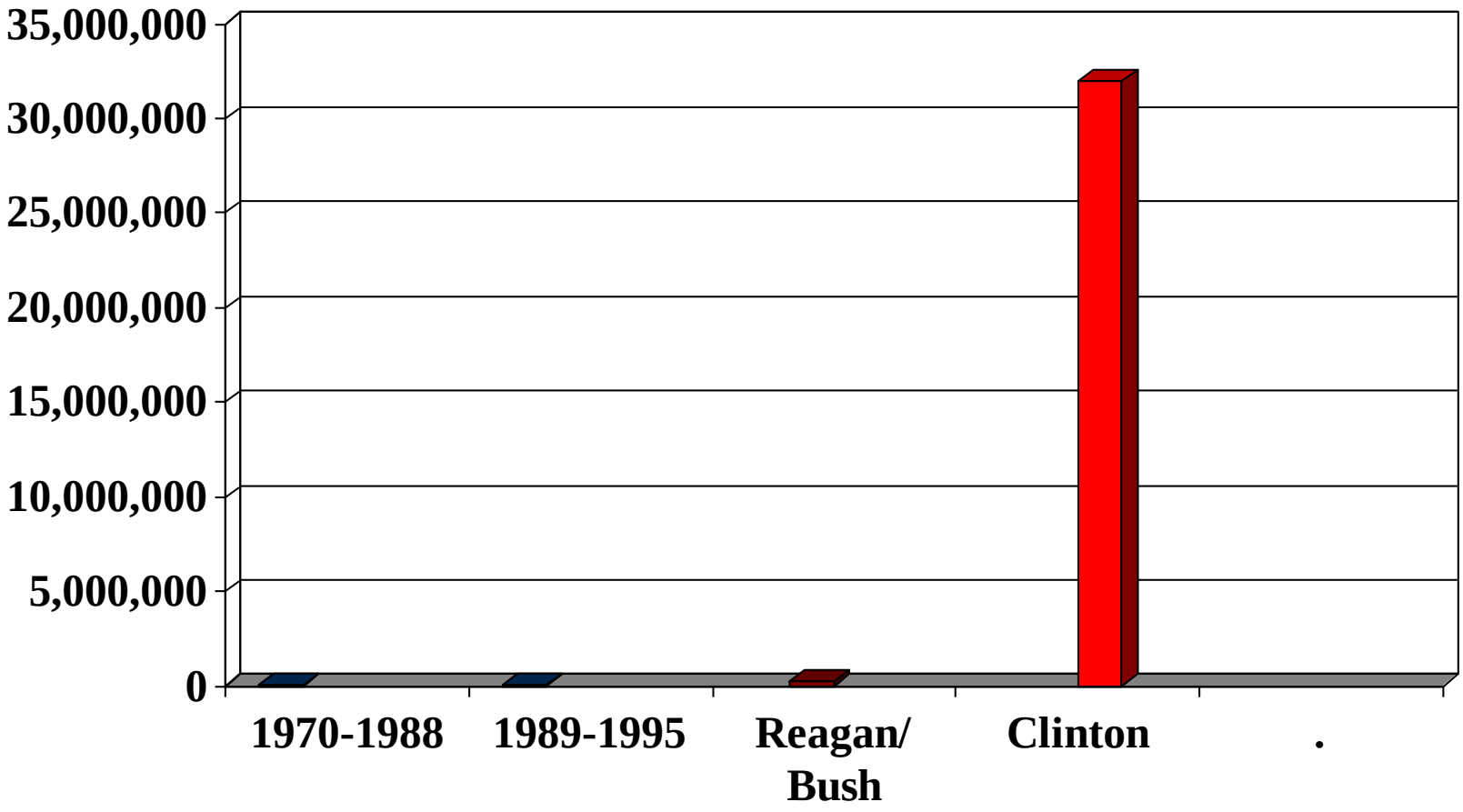
White House begins producing significant amount of digital records. Presidential records are heritage data of significant interest, but could not be presented on line prior to ERA

Transfers of Digital Files to NARA



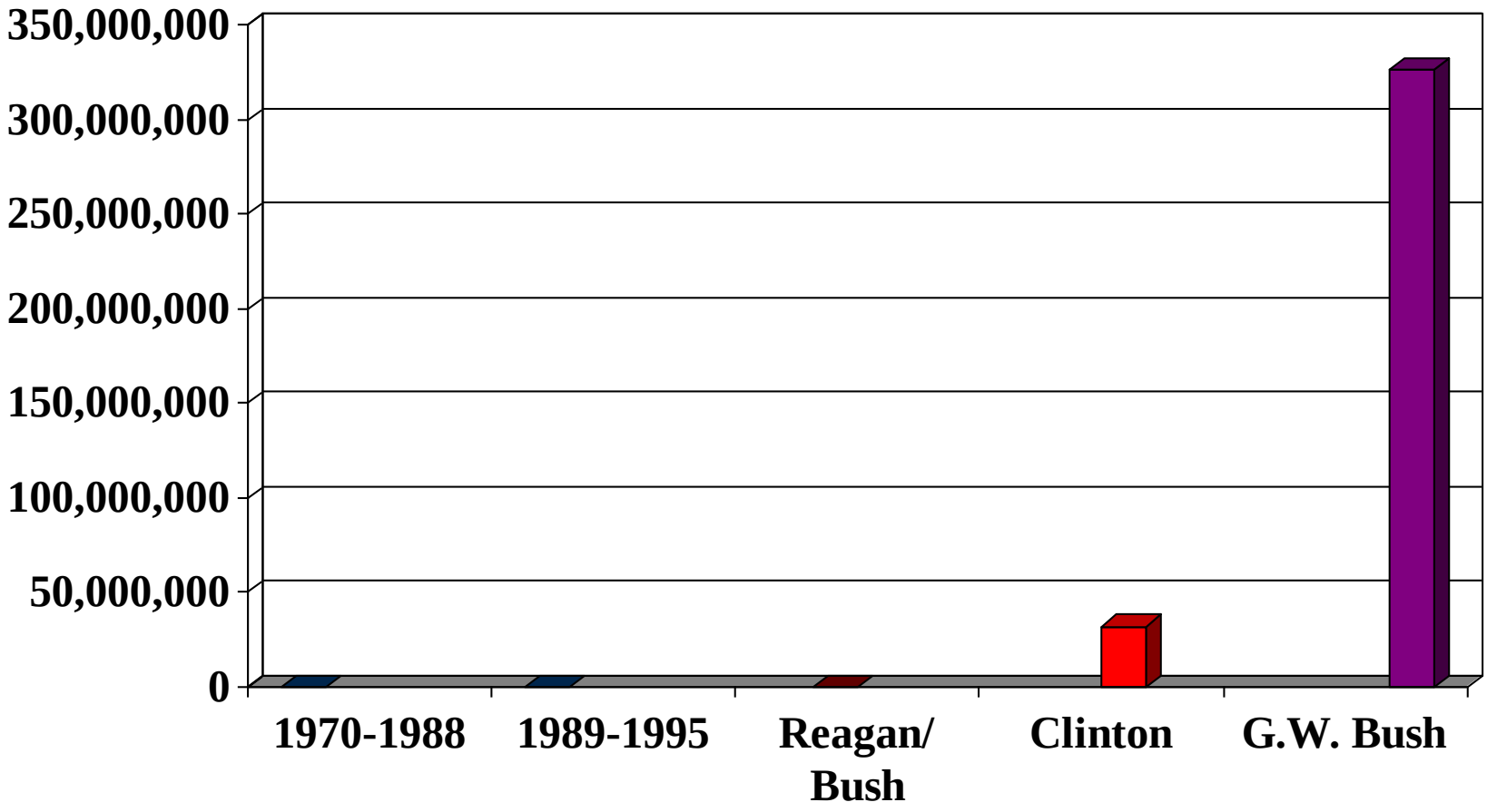
In terabytes

Transfers of Digital Files to NARA



In terabytes

Transfers of Digital Files to NARA



In terabytes

Case Study

- 2 TB portable drive containing office automation files from mid to late 1990s was discovered to be missing
- The drive contained an extract of EOP data that was of forensic interest and included large amounts of sensitive and PII data which needed to be identified
- The 2 TB disk not only presented a challenge in terms of size, but also content due to the breadth of
  - Categories of files (plain text, word processing, spreadsheets, databases, graphics, email, etc)
  - Source applications (e.g. MS Office products, DB III – V, Paradox, Revelation, Word Perfect, Lotus, Quattro Pro, etc)
  - Data formats.
  - Had to include considerations for 80 different date formats.

Count of Files and Types

- Files: 8,792,132
- Bytes: 1,933,270,158,678
- Unique file extensions: 13,247

- File exclusion and de-duplication left:
  - 2,165,994 files in 217 different file types

File categories and numbers of each

File Category	Number of Files
BITMAP/VECTOR	147
COMPRESSED	5,711
DATABASE	13,889
EMAIL	57,742
GENERIC	380,708
GRAPHIC	234,940
MULTIMEDIA	3,854
PRESENTATION	3,658
SPECIAL	67,379
SPREADSHEET	132,180
UNKNOWN	502,959
WORD PROCESSOR	762,006

"Unknown" includes Lotus Notes email and Revelation databases which use multiple files to produce a table

Structured Data by type and number

Database Type	Number of Files
DBASE	7,043
PARADOX	5,416
MS ACCESS	1,119
Q&A DATABASE	267
Microsoft Project 98	41
Others	2

Spreadsheet Type	Number of Files
EXCEL	19,590
Lotus 123	110,309
Quattro Pro	1,775
Others	31

Challenges to identification of PII in structured data

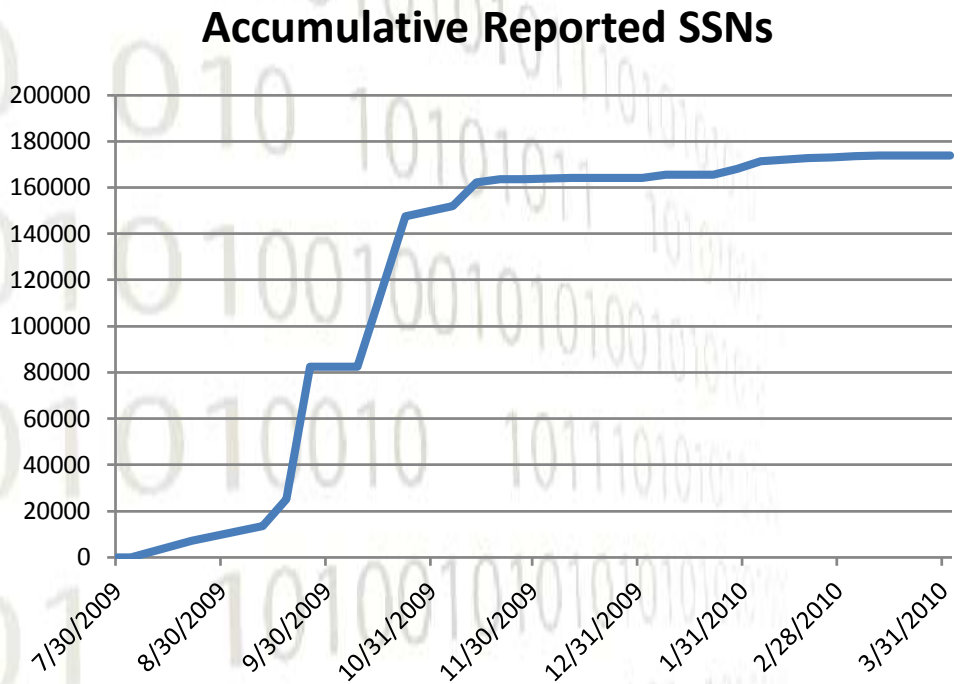
- Structured data posed analytical challenges to PII extraction:
  - Functionally similar files were organized with different schemas
  - data inconsistencies existed within columns (ex. DOB in the SSN column)
  - typographical errors
  - format inconsistencies in all PII data elements (name, DOB, SSN, etc.,)
- Disambiguation could not be done in an automated fashion and required human evaluation

Final file counts after analysis and extraction

	Totals
Number of Files copied from Master 2	**8,792,132**
Number of Processed Files with PII	**41,606**
Number of Files Processed with No PII (Moved to Excluded)	**7,138,514**
Number of Files with HIPAA or Passport numbers	6
Total Processed Files	**7,180,126**

- 1.6 million files contained data in unstructured sources.
- Majority of files with no PII (excluded) were duplicates
- 341,595 plain text files were searched using various pattern recognition techniques to extract structured content.
- Quality of extractions degrade rapidly compared to structured sources

Search results over time



- Flat line trend illustrates diminishing rate of return as efforts move from structured to unstructured data sources
- 198,097 SSNs have been extracted from 41,000 files

Summary and observations

- The amount of structured and unstructured data of forensic interest created by the USG alone is staggering and the curve is getting steeper
- Systems such as ERA are offering promise higher up the curve
- The NARA effort was directed at a small subset of heritage data less than 20 years old
- Data recovery from a 2 terabyte drive has cost over $600,000 (or $300k per t.) to date

Contact Information

Leo Scanlon
Chief Information Security Officer
National Archives and Records Administration
301 837 0752
leo.scanlon@nara.gov