Updating a Demand Analysis Forecast and Creating a Format Profile for NARA Electronic Records Holdings

Leslie Johnston
Director of Digital Preservation
National Archives and Records Administration

Updating our Understanding of our Holdings

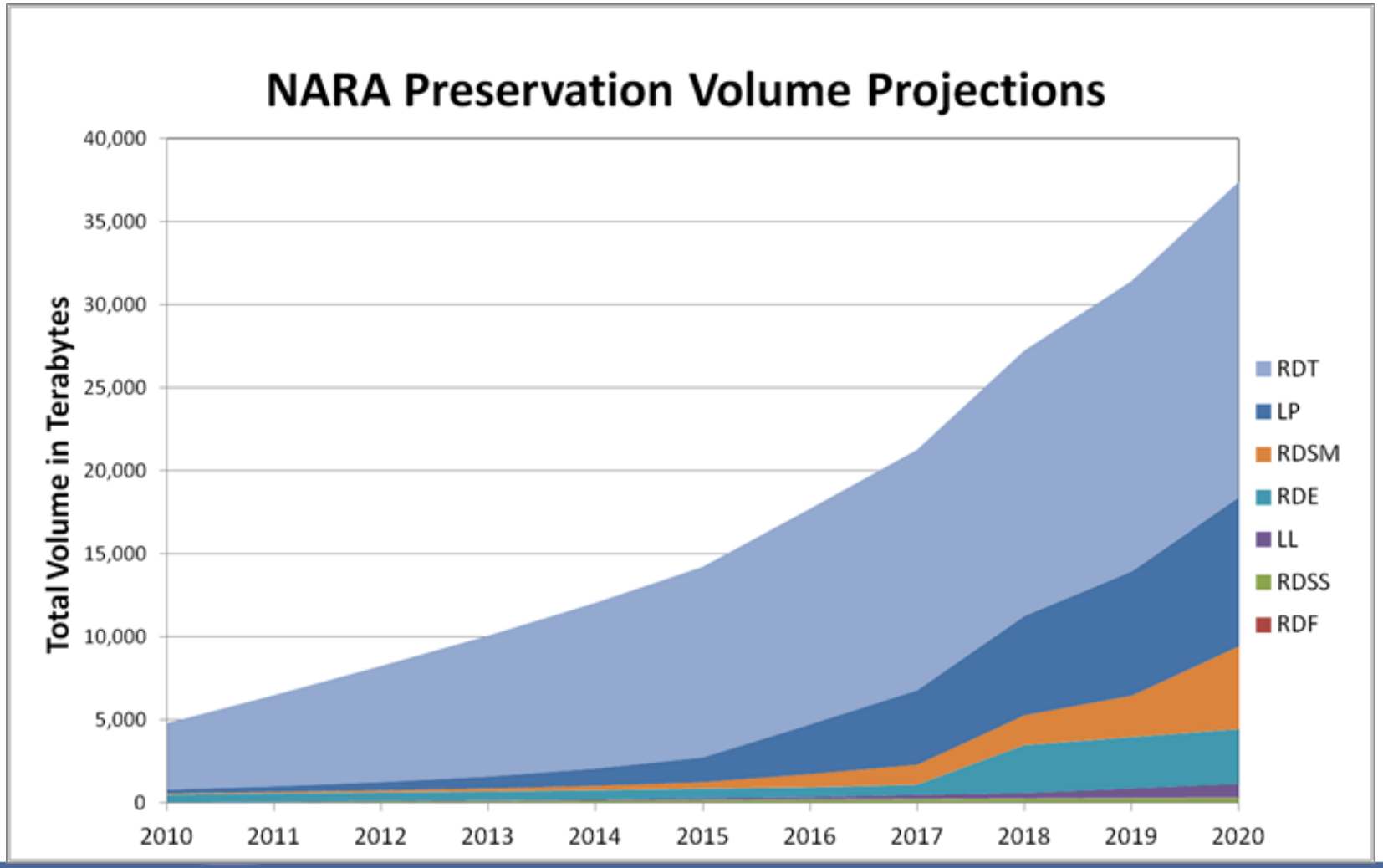
Demand Analysis

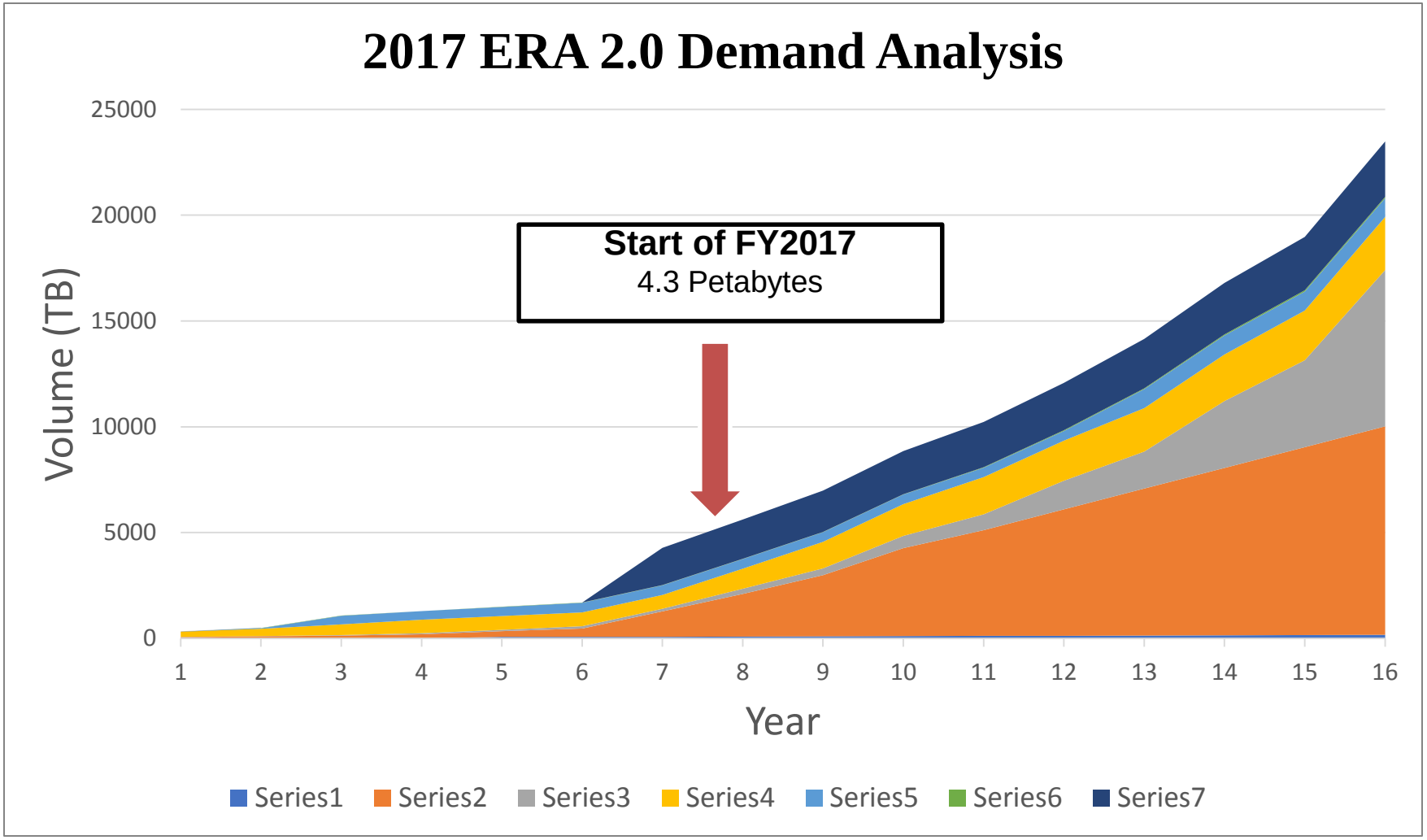
- A study done in collaboration with stakeholders from each unit of NARA expecting to receive or create digital records on the amount of incoming material.
  - Several stakeholder units: LL (legislative records), LP (Presidential libraries), RDE (federal born digital textual records), RDF (FOIA release), RDSS (federal born digital and digitized still pictures), RDSM (federal born digital and digitized motion pictures), RDT (federal textual digitization), and VI (digitization partnerships).
- Was first surveyed in early 2015
- Survey was scoped to include all backlogs and known scheduled transfers over the next five years
- The 2015 results overestimated the capacity to increase the rate of holdings digitization.

Lessons Learned

- NARA backlogs are substantial, cannot be included in current volume forecasts
- NARA has trouble predicting the size and count of incoming records
- NARA record disposition activities greatly affect permanent storage needs

2015 Demand Analysis




2017 ERA 2.0 Demand Analysis
Start of FY2017
4.3 Petabytes
Volume (TB)
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Year
Series1
Series2
Series3
Series4
Series5
Series6
Series7

Updating our Understanding of our Holdings

Electronic Records Format Profile

- Several Systems:
  - ERABase (federal records)
  - CRI (legislative records)
  - Title 13 2000 and 2010 (Census)
  - EOP 43 and 44 (Presidential)
  - PERL 40, 41, 42 (Presidential)
- Each a different infrastructure with different tooling (or none) for identifying and characterizing file formats.
- Collated all current ingest reports, some with format characterization and file extensions, some with only file extensions which were mapped to a matrix of formats/applications. The characterizations were also mixed granularity, such as Adobe PDF vs. Adobe PDF 1.6.

Lessons Learned

- The mix of systems and tooling leads to sufficient characterization to provide a high-level understanding of our holdings, but not sufficient detail (yet) to identify granular format risks so we can audit the holdings in the current environment and take preservation actions.
- These needs were already ideintifed for ERA 2.0 but have increased in priority.

Format Counts out of 1,466,258,253 Total Files
Across All ERA System Instances

Format Total	Format	
776,199,663	Electronic Mail Message file	CRI, EOP43, EOP44,
364,767,209	JPEG bitmap graphics file	CRI, EOP43, EOP44, PERL40, PERL41, PERL42, Title13 2010,
126,445,935	Tagged Image File Format	CRI, EOP43, EOP44, ERABase, PERL42, Title13 2000,
66,615,143	HyperText Markup Language document	CRI, EOP43, EOP44, PERL40, PERL41, PERL42,
52,697,013	ASCII 8-bit Text	CRI, EOP43, EOP44, ERABase, PERL42, Title13 2000,
29,787,055	Extensible Markup Language file	CRI, EOP43, EOP44,
8,181,392	Adobe Acrobat PDF file	CRI, EOP43, EOP44, PERL42,
6,166,056	Document text file	CRI, EOP43, EOP44, PERL42,
5,546,971	Canon Raw Image file	CRI, EOP43, EOP44,
4,381,935	None Reported	CRI, EOP43, EOP44,
3,886,434	Microsoft Word Open XML Document text file	CRI, EOP43, EOP44,
2,649,585	Microsoft Excel spreadsheet; Microsoft Works spreadsheet; DATAIR Data I..	CRI, EOP43, EOP44, PERL42,
1,728,080	CompuServe Graphics Interchange Format bitmap	CRI, EOP43, EOP44, PERL42,
1,054,580	Message	CRI, EOP43, EOP44,
1,007,502	Adobe Acrobat PDF 1.3 - Portable Document Format 1.3	ERABase,
887,126	Microsoft Excel XML file	CRI, EOP43, EOP44,
792,830	Encapsulated PostScript file format; Printer font	CRI, EOP43, EOP44, PERL42,
705,678	JPEG File Interchange Format 1.02	ERABase,
684,394	Plain Text File	ERABase,
612,724	WordPerfect 6.0 document ; PFS:WindowWorks document	CRI, EOP43, EOP44,
595,676	Raw JPEG Stream	ERABase,
540,543	Portable Network graphics file	CRI, EOP43, EOP44,
451,057	FoxPro report	CRI, EOP43, EOP44,
420,552	Uniform Resource Locator	CRI, EOP43, EOP44,
419,029	Adobe Acrobat PDF 1.4 - Portable Document Format 1.4	ERABase,
407,182	ASCII Comma Separated Values text file format; CompuShow Adjusted EG..	CRI, EOP43, EOP44,
382,956	Microsoft Windows Shortcut file; .RTLink Linker response file	CRI, EOP43, EOP44, PERL42,
344,412	Bitmap graphic	CRI, EOP43, EOP44, PERL42,
325,667	Digital Moving Picture Exchange Bitmap	EOP44, ERABase,
229,666	Microsoft PowerPoint Presentation	CRI, EOP43, EOP44,
209,287	Adobe Acrobat PDF 1.6 - Portable Document Format 1.6	ERABase,
205,490	Apple QuickTime Video Clip; Apple QuickTime Audio; AutoCAD AutoFlix Mo..	CRI, EOP43, EOP44, PERL42,
205,246	Microsoft PowerPoint OpenXML presentation	CRI, EOP43, EOP44,
203,108	Standard Generalized Markup Language	CRI, EOP43, ERABase,
169,241	RealPlay SMIL file; Self Mounting Image file	CRI, EOP43, EOP44,
161,910	Adobe Acrobat PDF 1.2 - Portable Document Format 1.2	ERABase,
158,501	Microsoft Windows Dynamic Link Library; OS/2 Dynamic Link Library; Corel..	CRI, EOP43, EOP44,

Discussion?

Leslie Johnston
Leslie.Johnston@nara.gov