Web Archives

A Storage & Content Management Use Case

Laura Graham

Library of Congress Web Archives Team

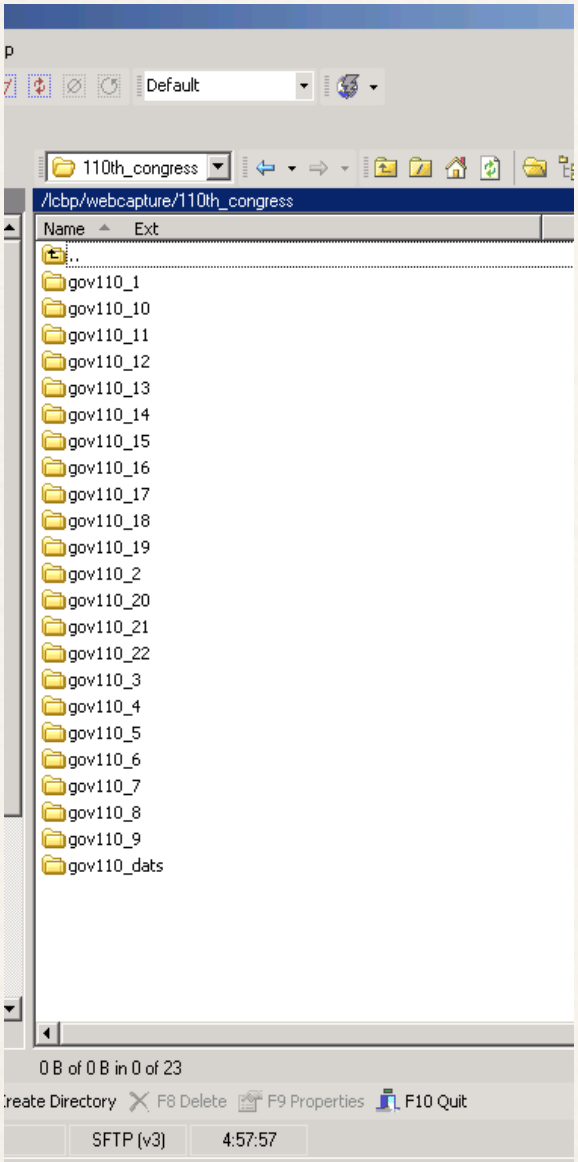
The Tools & Formats

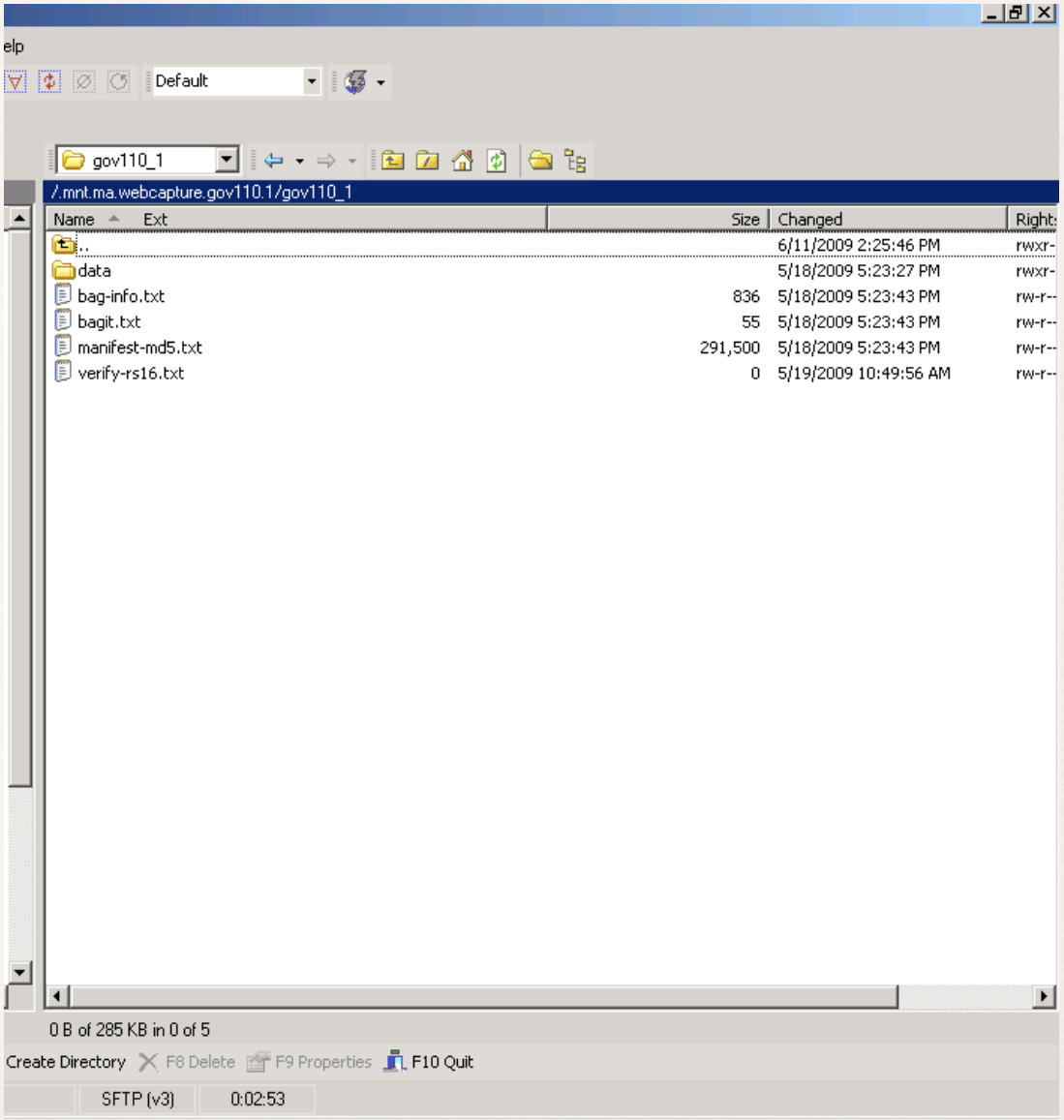
- **Heritrix: Produces the Content**:
  - Developed by Internet Archive
  - Open Source Java App
  - **Output File Format**:
    - ARCs
    - Warcs: ISO 28500:2009
- **Wayback “Viewer” Provides the Access**:
  - Developed by Internet Archive
  - Open Source Java App

Content: A “Plain Vanilla” Use Case

- **File & Content is fixed**
  - What Heritrix writes = what we transfer = what we store = what we make accessible to the public
    - No master or derivatives
    - WARC files themselves may remain compressed
- **Organization**
  - No items … just files in buckets
    - Collection
    - Crawl Frequency
- **Bagit File Packaging Format**
  - Bags persist from transfer to storage to access

Bags and more bags … !

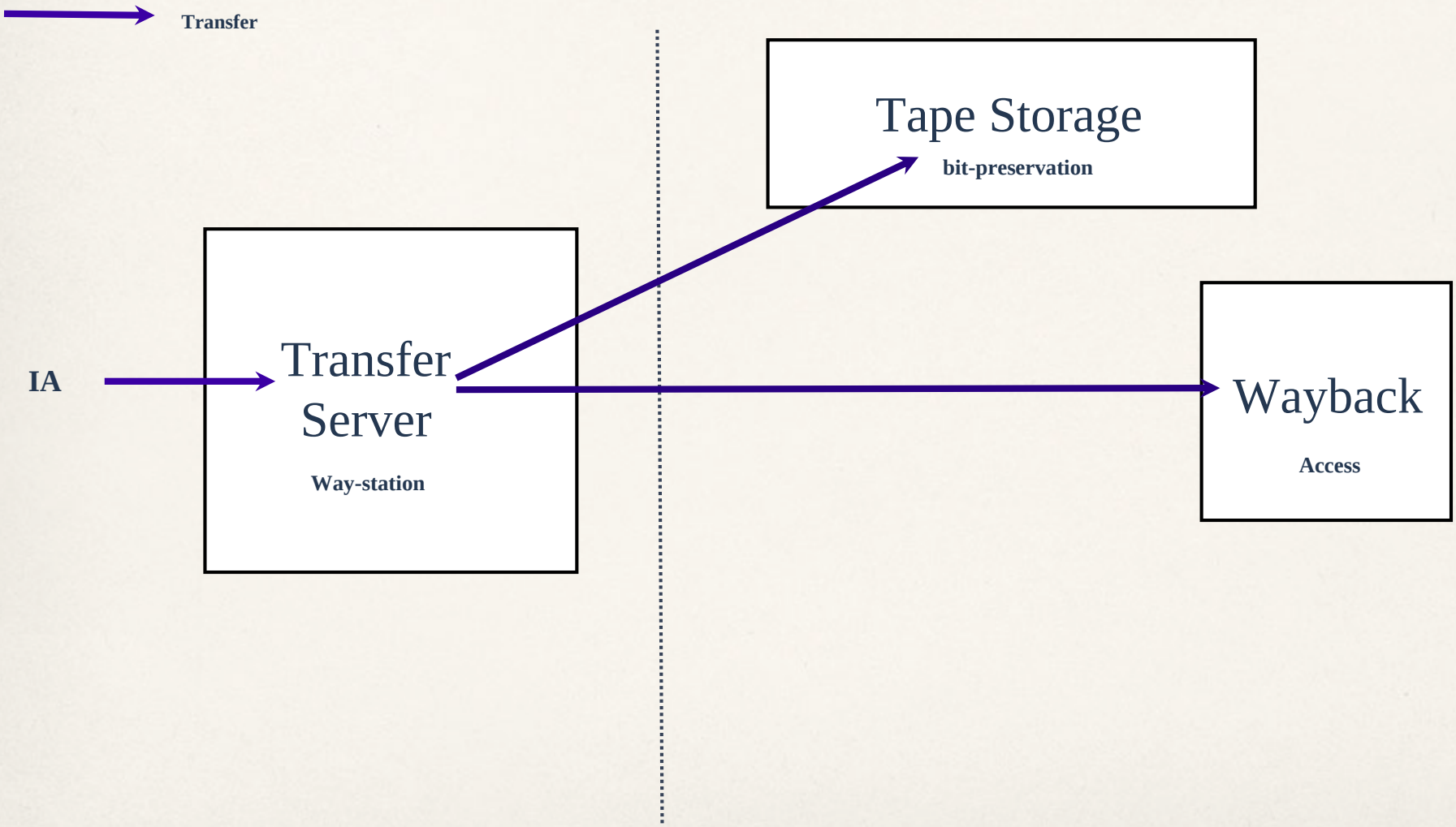




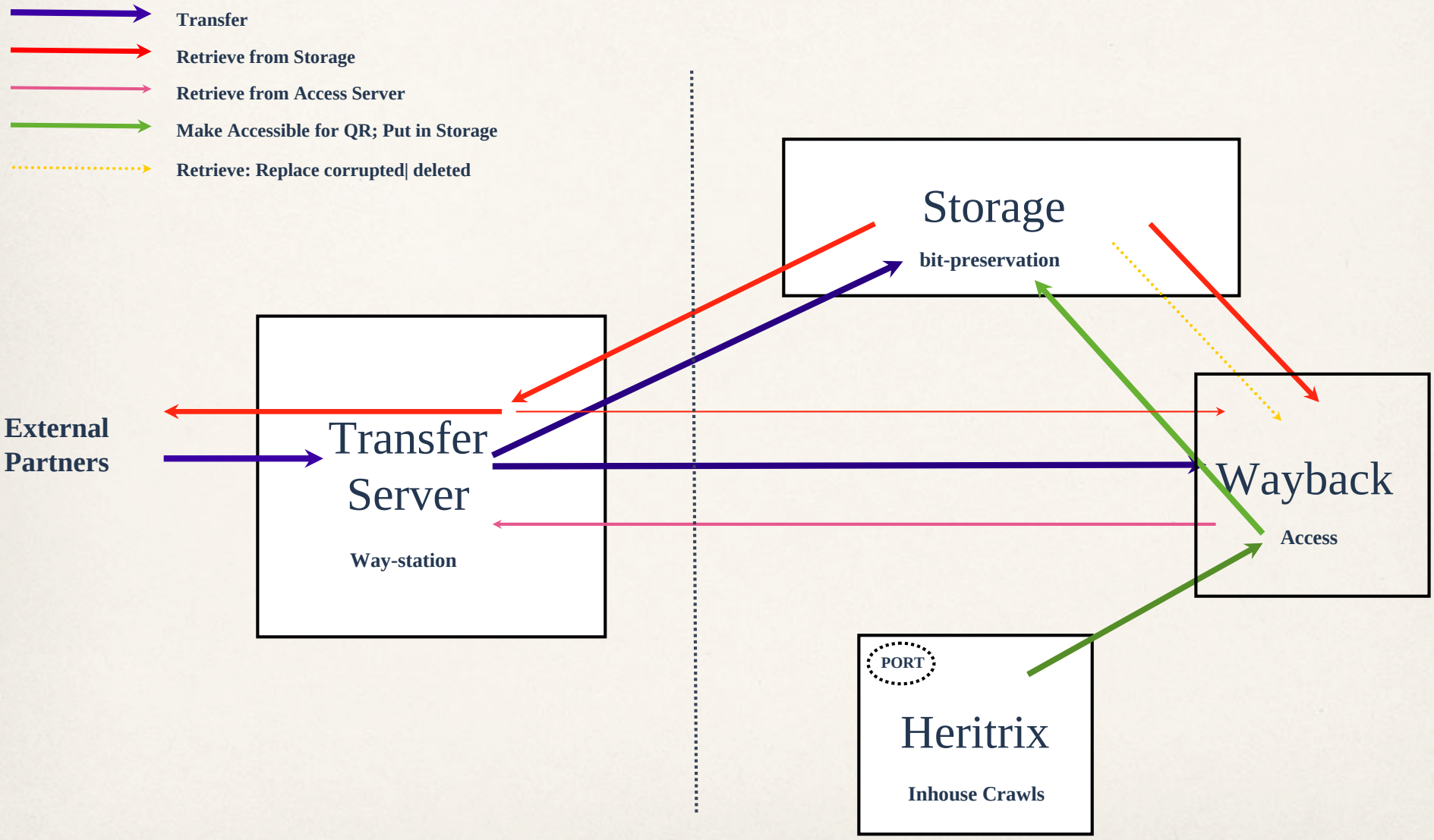
Scale of Content (Internet Archive)

Year	GBs	Change % - | +	Transferred
2001	136		0
2002	10221	7415	0
2003	762	-92	0
2004	7771	919	0
2005	16364	110	0
2006	21557	31	0
2007	19003	-11	9391
2008	12863	-32	16569
2009	32875	155	38746
	121552		**64706**

Starting out in 2008…



…and here we are in 2009



Requirements (aka wish list) …

- Easy & efficient ‘interoperability’ between machines
  - automation / elimination of “manual labor” steps
- Universal Tracking
  - workflow status & data attributes of everything everywhere … not just in storage

Example: Copying & Verifying

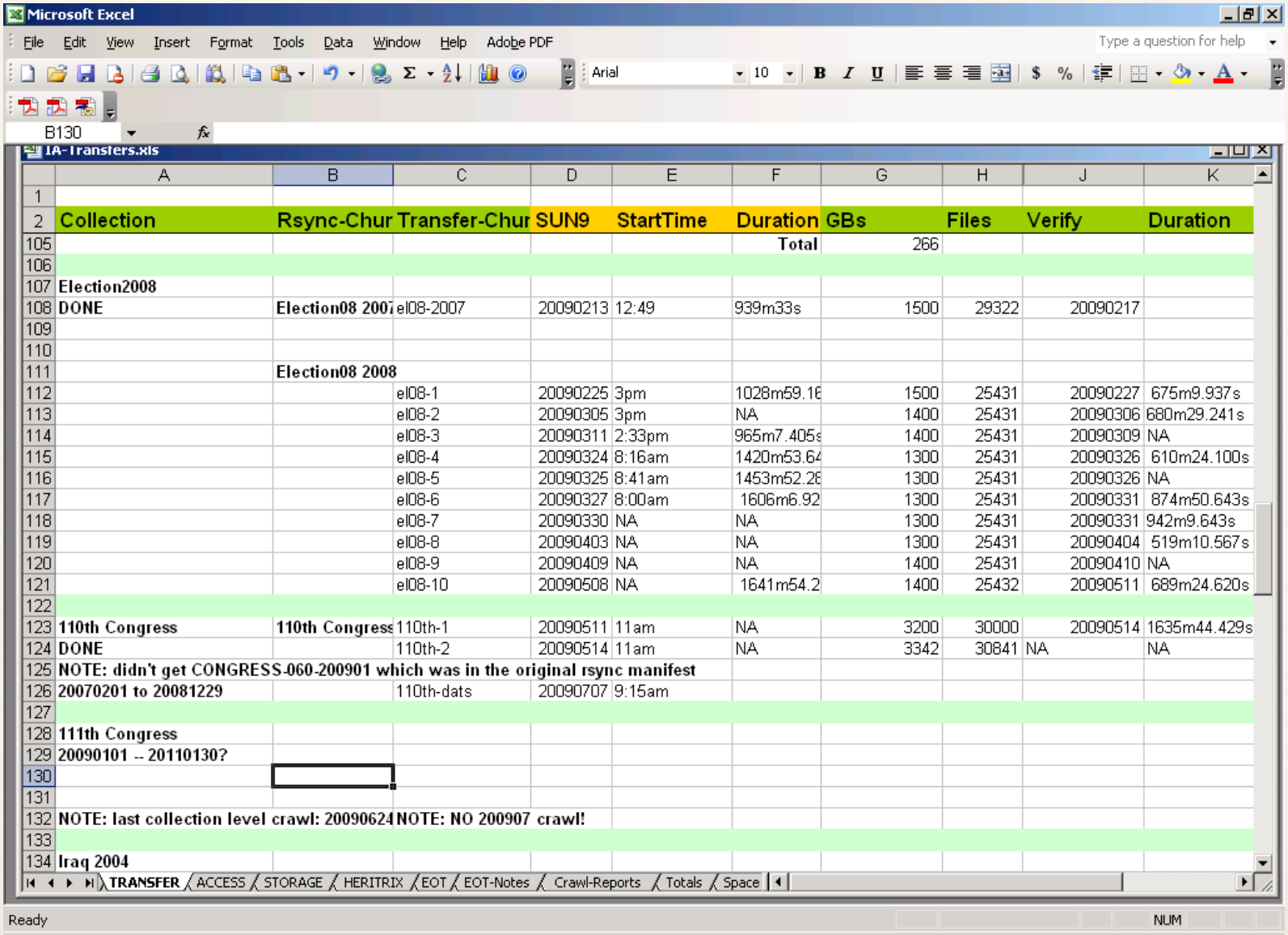
- Copying Bags:
  - Where possible, we'd like *not* to have to move stuff so much…

- Verifying Bags:
  - Straightforward file fixity
  - We want to know our content is safe
    - …but we'd like to verify-on-the fly just about everywhere we go

Example: Repackaging

- Automated ‘Repackaging’ of Bags “en route” to target server
  - Workflows/Machines have different Requirements
  - Size:
    - Transfer: 1+ TB is efficient
    - Storage: >=650-800 GBs good for management
    - Access Server: 300 GBs upper limit
  - Source bag → Destination bag**s**
    - 1 to many
    - Many to 1
    - and more…

Tracking (multiple machines/processes)



Microsoft Excel

File Edit View Insert Format Tools Data Window Help Adobe PDF

Type a question for help

Arial 10

B130

IA-Transfers.xls

	A	B	C	D	E	F	G	H	J	K
1										
2	Collection	Rsync-Chur	Transfer-Chur	SUN9	StartTime	Duration	GBs	Files	Verify	Duration
105						Total	266			
106										
107	Election2008									
108	DONE	Election08 200	el08-2007	20090213	12:49	939m33s	1500	29322	20090217	
109										
110										
111		Election08 2008								
112			el08-1	20090225	3pm	1028m59.18	1500	25431	20090227	675m9.937s
113			el08-2	20090305	3pm	NA	1400	25431	20090306	680m29.241s
114			el08-3	20090311	2:33pm	965m7.405s	1400	25431	20090309	NA
115			el08-4	20090324	8:16am	1420m53.64	1300	25431	20090326	610m24.100s
116			el08-5	20090325	8:41am	1453m52.28	1300	25431	20090326	NA
117			el08-6	20090327	8:00am	1606m6.92	1300	25431	20090331	874m50.643s
118			el08-7	20090330	NA	NA	1300	25431	20090331	942m9.643s
119			el08-8	20090403	NA	NA	1300	25431	20090404	519m10.567s
120			el08-9	20090409	NA	NA	1400	25431	20090410	NA
121			el08-10	20090508	NA	1641m54.2	1400	25432	20090511	689m24.620s
122										
123	110th Congress	110th Congress	110th-1	20090511	11am	NA	3200	30000	20090514	1635m44.429s
124	DONE		110th-2	20090514	11am	NA	3342	30841	NA	NA
125	NOTE: didn't get CONGRESS-060-200901 which was in the original rsync manifest									
126	20070201 to 20081229		110th-dats	20090707	9:15am					
127										
128	111th Congress									
129	20090101 -- 20110130?									
130										
131										
132	NOTE: last collection level crawl: 20090624		NOTE: NO 200907 crawl!							
133										
134	Iraq 2004									

TRANSFER ACCESS STORAGE HERITRIX EOT EOT-Notes Crawl-Reports Totals Space

Ready NUM

In Sum: Use Case = Interaction

- 2008: Web Archives began with a simple, one-directional workflow, and a focus on the content in the “storage box”
- 2009: Storage box now part of a complex of interactions, workflows, and machines
  - The User “lives” in that “complex”
  - Tools & utilities to manage the relationship between storage & its larger environment