DIGITAL ARCHIVES UPDATE

Dongie Agnir, Systems Developer
Micah Takabayashi, Acquisition Specialist
Adam Jansen, Project Consultant

10 October 2014
Electronic Records Day

AGENDA

- Why Archives Are Important
- Importance of Authenticity
- How To Send Records To The Archives
- Preserving Digital Records

WHY ARCHIVES ARE IMPORTANT

Role of the Archives

Above all,

Protect the Record

Involves not only the object itself, but also protecting:

- Naturalness (produced as a by-product of activity)
- Interrelatedness (gain meaning from each other)
- Impartiality (free of external influence)
- Authenticity (uncorrupted)
- Uniqueness (derived from context)

NEW PARADIGM

- Shift from Paper to Electronic more than just format
  - Different process of creation
  - Storage more fluid
  - More easily corrupted
  - Shorter useful 'lifecycle'
- Greater potential
  - Search
  - Extract metadata
  - Validation of integrity

AUTHENTICITY IN THE DIGITAL AGE

Maintaining integrity does not necessary mean authenticity, while it is an important part there is something more…

Authenticity requires understanding the identity:

- Determine one version from another
- Sender from receiver, draft from final

For a record to be "what it purports to be" requires understanding of intent and context of creation, collection and handling (chain and process of custody)

IDENTITY AND INTEGRITY

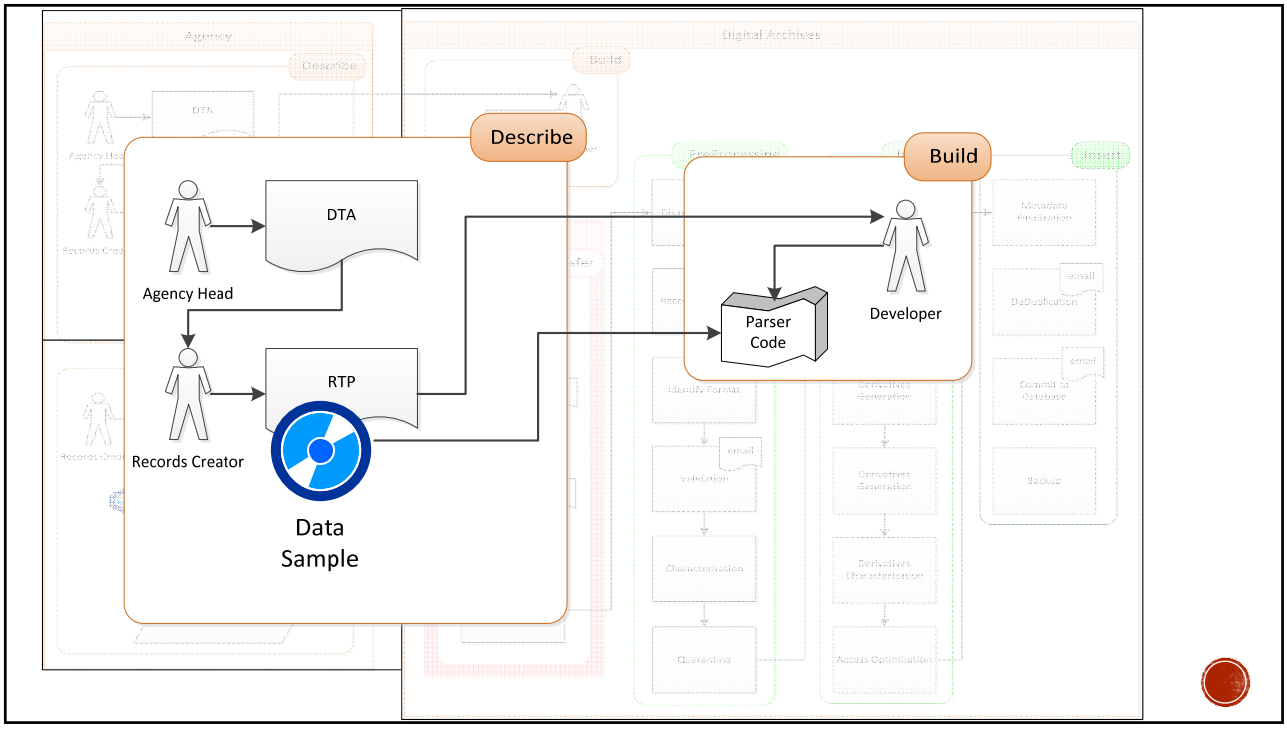
Identity	Integrity
Name of person creating	Name of person handling
Date of Creation	Name of custodian
Matter or Action	Indication of annotation
Relationship to other records	Technical changes
Documentary form	Presence of digital signature
Digital Signature	Location of duplicates
Name of person responsible	Hash value

HOW TO SEND RECORDS TO THE ARCHIVES

NEW PROCESS

To reap the greatest benefit, provide strong security = new way of 'doing business':

- Evolution of current archival processes to leverage technology, cutting edge research
- Digital Records cannot simply be 'dropped off'
  - Interrelatedness, creation process much more complex
- Requires more preparation in advance of transfer
  - BUT, one time effort* (*until things change)
- Starts with Digital Transmittal Agreement
- Then Records Transmittal Plan
- Then Testing
- Finally, all systems 'Go'



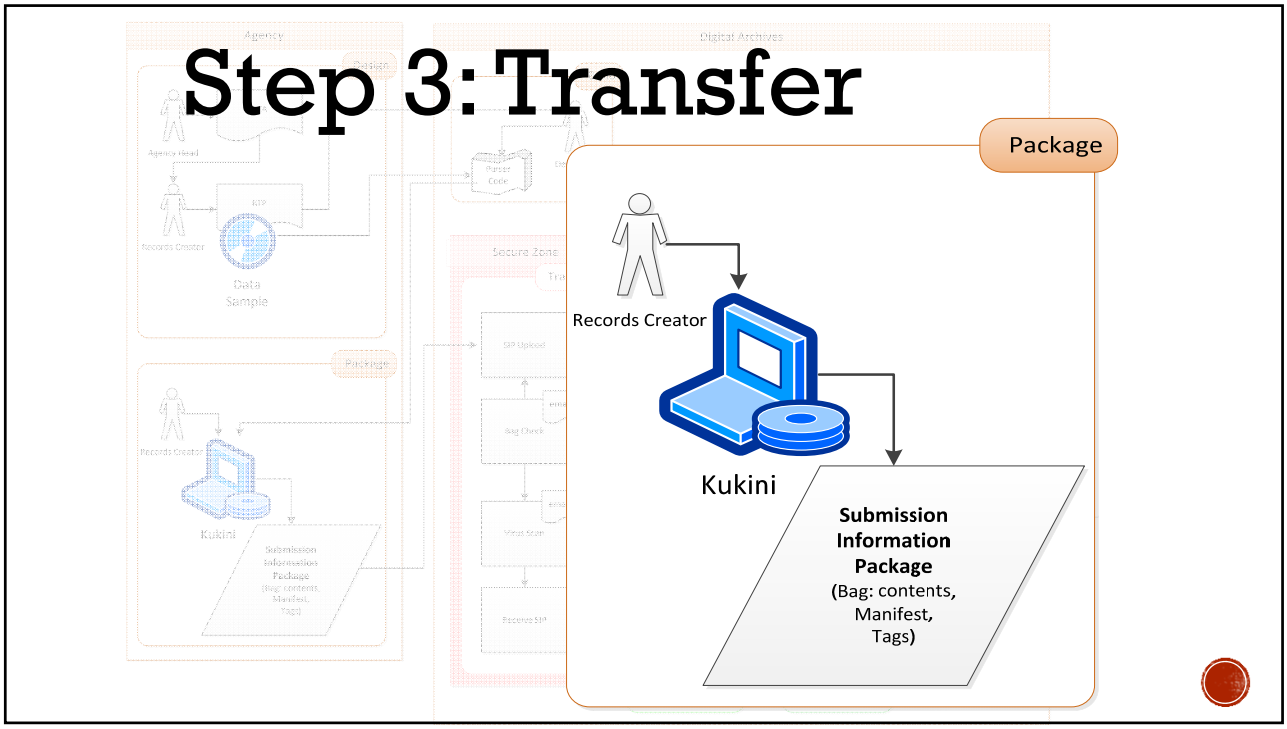
STEP 1: DIGITAL TRANSMITTAL AGRMT (DTA)

- Contains acknowledgement that this is a cooperative process
- Acknowledge need for Public Access (92-F)
  - Recognize that not ALL content is disclosable
- Asks partner agencies to provide input and feedback on protocols, interfaces
  - Provide information on agency records
  - Sample records
  - Access to subject matter experts

STEP 2: RECORDS TRANSMITTAL PLAN (RTP)

- 'Contract' for records transmission
  - Who are contacts
  - What type of records
  - What formats, number of each
  - Restrictions from public disclosure
- Accompany Data Sample for testing

Step 3: Transfer

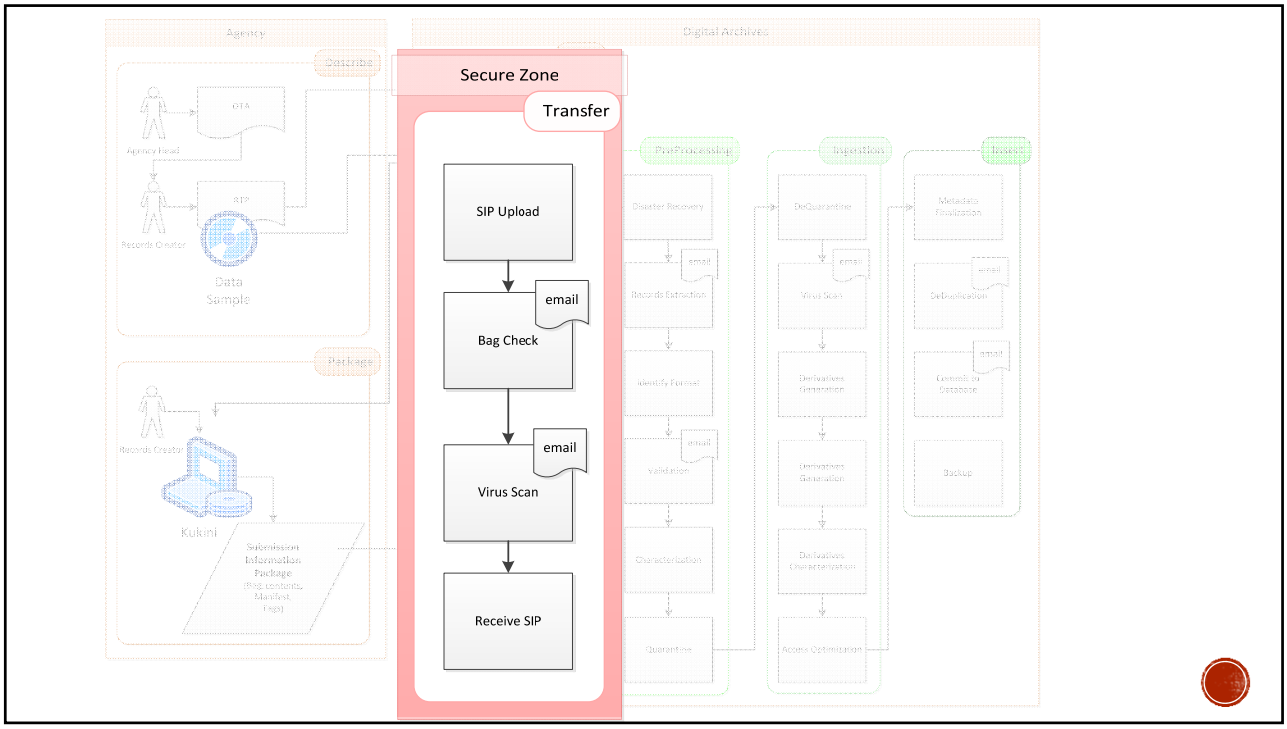


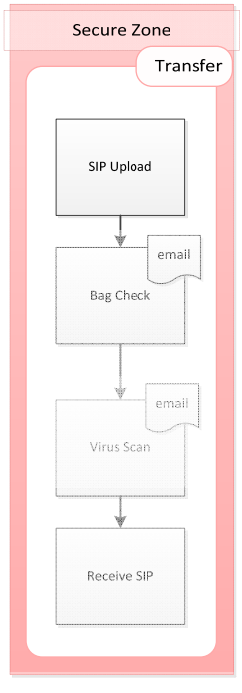
KUKINI IN ACTION

PRESERVING DIGITAL RECORDS

PERSISTENT IDS

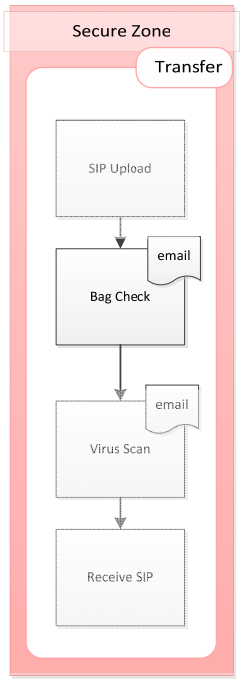
- Digital Archives assigns unique identifiers to everything in the repository:
  - Accessions
  - Records
  - Digital Objects
  - Archival Processing Events
- Protects against mixing up similarly named files
- Tracks what identifiers have been assigned to ensure uniqueness
- Archives is committed to maintaining these unique identifiers through technology upgrades and format migrations





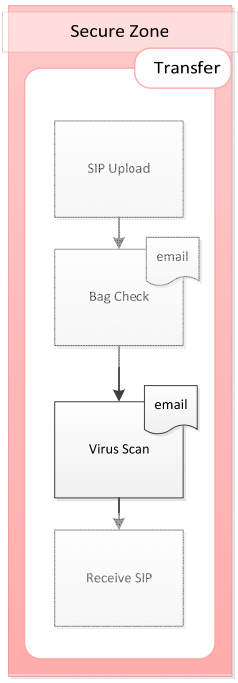
SIP UPLOAD

- Brings transferred files into the control of the Archives
- Transferred packages contain not only the records, but important authenticity metadata
- BagInfo
  - List of all the files contained in the bag
- Manifest
  - Listing of all files and their unique 'signature'
- Provenance
  - Listing of who sent, from what account, on what computer, in which network



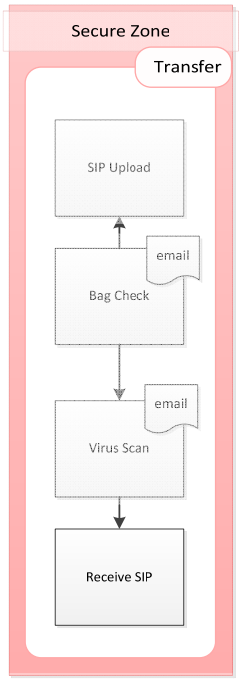
BAG CHECK

- Confirms that all records that were intended to be sent have been received
- Checks all the transferred files to ensure that none were damaged, corrupted in transit
- In the event that transfer is incomplete/corrupted, sender and archivist notified – transfer moved to error directory



VIRUS SCAN

- Scans all transferred files for viruses, malware or other potentially dangerous code
- Uses multiple virus scanners from different companies to achieve greater overlap and efficiency in virus detection
- In the event of a virus is detected, sender, IT and archivist notified – transfer moved to error directory

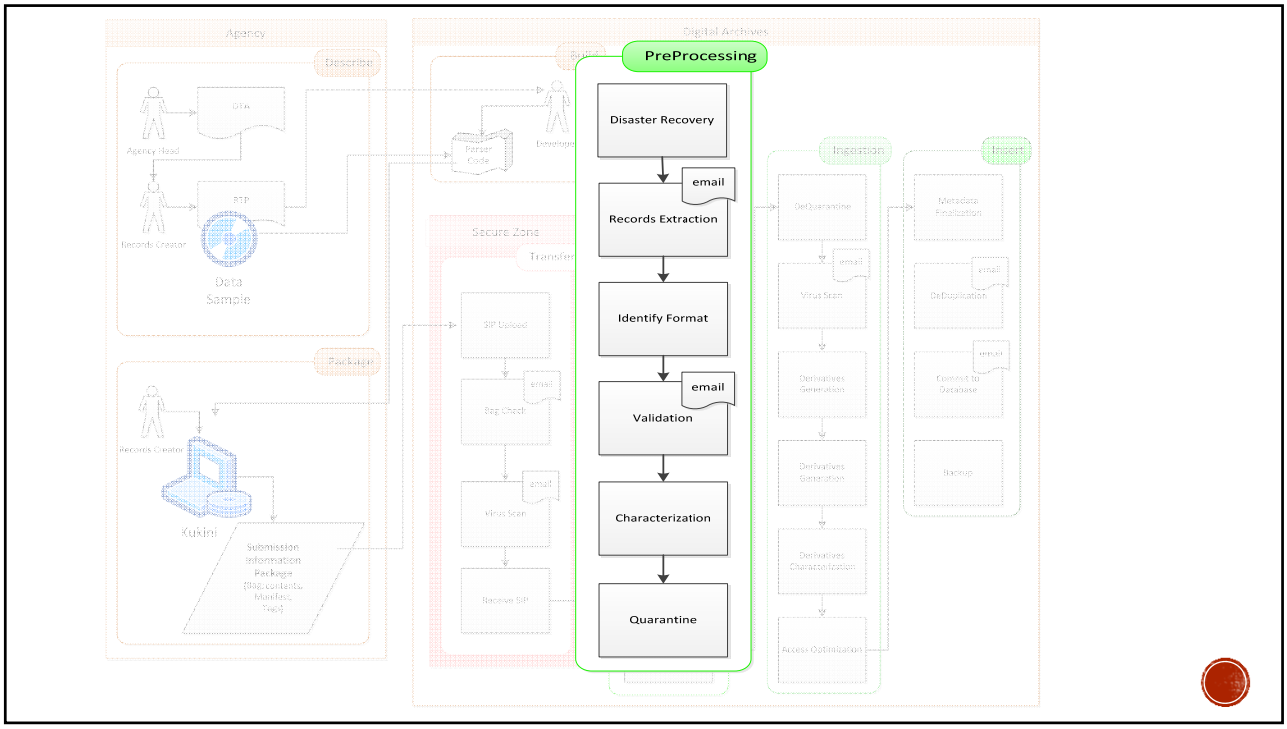


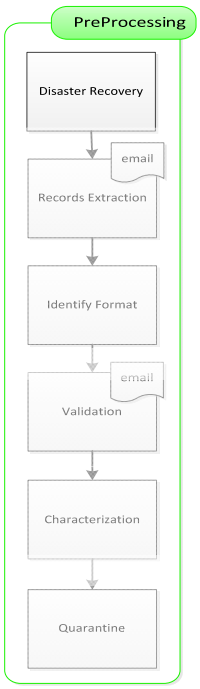
RECEIVE SIP

- The transferred material has been determined to have been complete and safe for handling
- Moves transferred records into the processing area of the Archives

After this point 'Archival' operations begin…

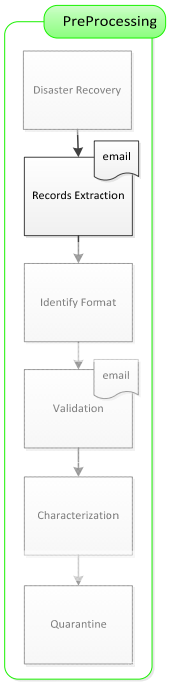
PRE-PROCESSING





DISASTER RECOVERY

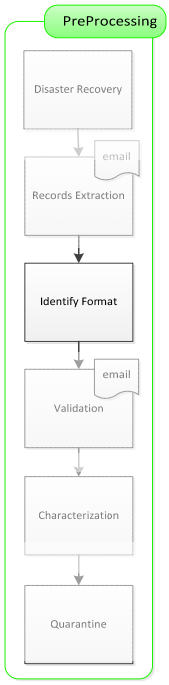
- Moves a copy of the records, as they were sent, off to backup location
- Nightly process moves records off to tape for transfer to offsite location
- Preserves original material
  - For reuse in the event of an error or disaster
  - Return to original agency in case of system error



RECORDS EXTRACTION

- Separates the single, bulk transfer into discrete identifiable records
- Associates index and metadata information with their corresponding digital objects

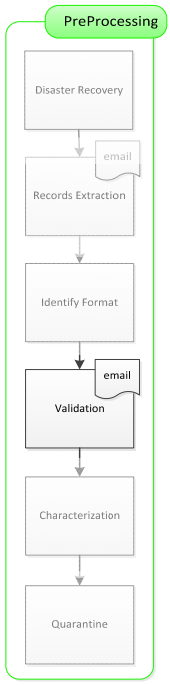
- Example of Accession and its Related Records, Digital Objects
- In the event that the records cannot be extracted, or are missing required information/files, sender and archivist notified – transfer is moved to error directory



IDENTIFY FORMAT

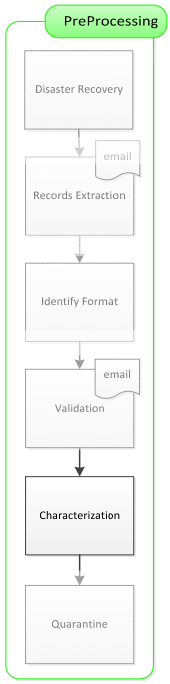
- Determines not only file extension, but version of file format
- Not all .PDFs are created equal
  - Same extension since the introduction of the Portable Document Format in 1993
- Format identification used for future preservation planning

- Example of Format Identification



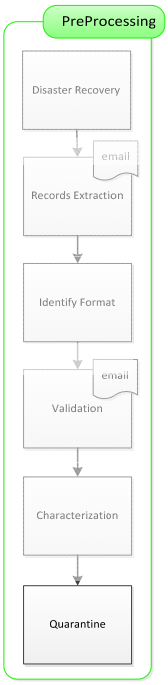
VALIDATION

- Confirms that file format conform to their published standard
- Make sure that PDFs, TIFs, DOCs, etc are 'good' and able to be read and preserved into the future
- Different formats, versions require different tools to validate
  - Research and use 'best tool' for the job
- Example of Properly Formatted Digital Objects
- In the event that file is 'not valid', sender and archivist notified – transfer is moved to error directory



CHARACTERIZATION

- Extracts important metadata about the digital objects
- Used for advanced searching
- Provides important information for preservation planning
- Different formats, versions require different tools to validate
  - Research and use 'best tool' for the job
- Example of Digital Objects Metadata



QUARANTINE

- Sends all records into a 'holding' period
- Allow virus scanning companies to update virus signature files

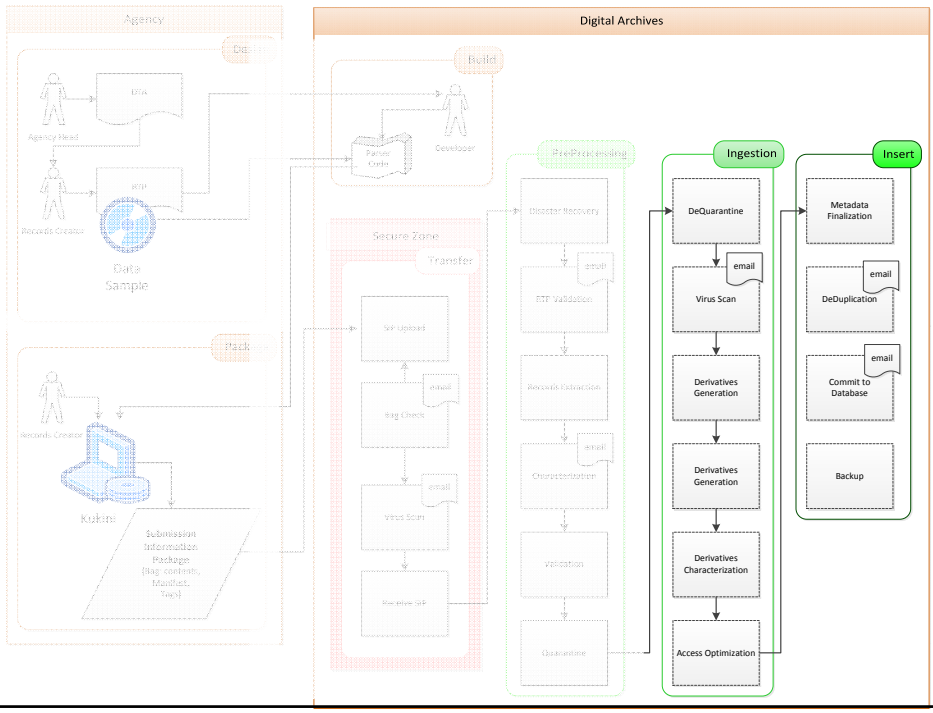
- After a pre-determined amount of time,
  - Records pulled out of Quarantine,
  - Virus Scanned again
  - Continue on to Ingestion Stage

NEXT STEPS

Adam Jansen

REFINE PRE-PROCESSING

- Put polish on visual interfaces
- Provide more detail in notification emails
- Continue to research best tools based on file format
- Optimize, stress test environment
- Finalize format of Records Transmittal Plan



GETTING CLOSE…

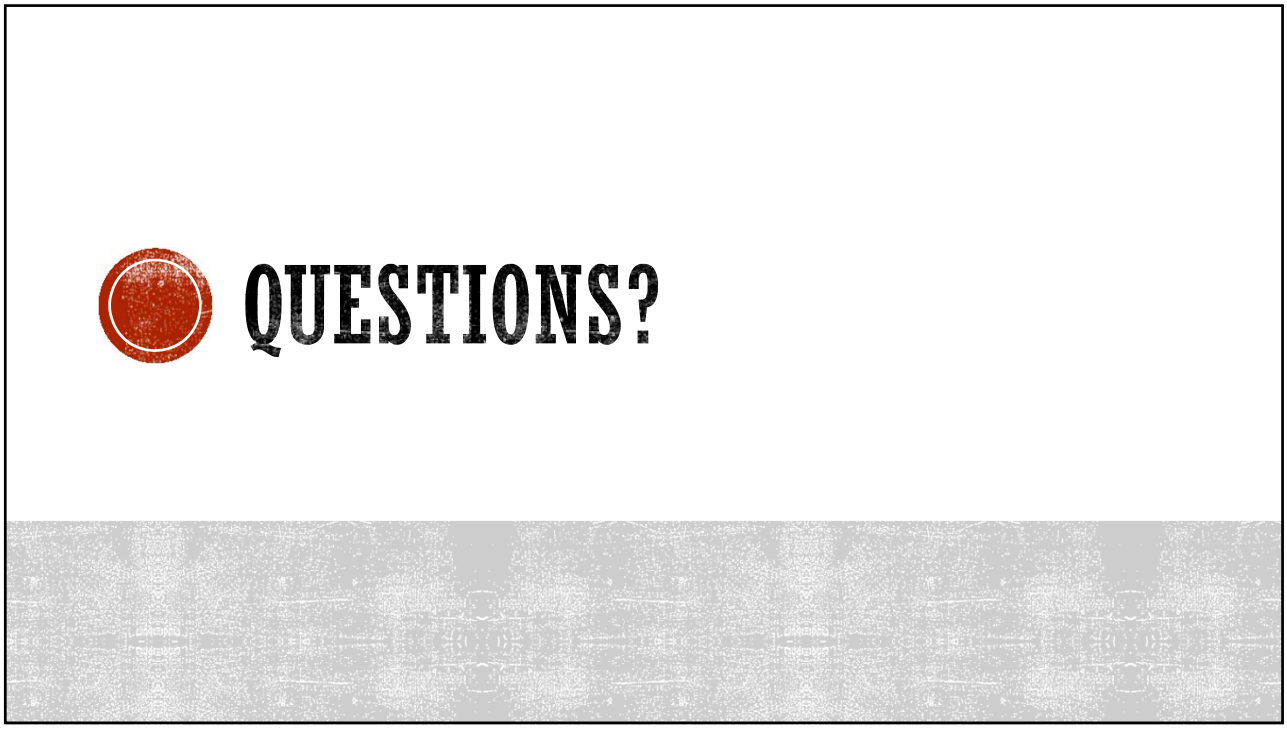
- Finish the Preservation Pipeline
- Lock down the back end
- Develop the search and display interfaces
- Test, Test, TEST

- Gather more data samples and work with more Partners to onboard records
  - Solicit feedback from Agency, refine processes and interfaces

- By this time next year, provide access to the public

DIGITAL PRESERVATION… THE LONG VIEW

- Digital Preservation REQUIRES migration and conversion
- Digital preservation is a continual process requiring:
  - Research on emerging technologies
  - Constant verification of integrity
  - Test of update options
  - Verification of upgrades
- Replacement of underlying infrastructure every 4-6 years
- Inclusion of evolving technology, methods of practice


QUESTIONS?