Technology Module:
Walk The Workflow

@digitalPOWRR

This POWRR Institute is generously funded by the INSTITUTE of Museum and Library SERVICES

Expected Outcomes

- ✓ **Become familiar with basic steps that will take you through a simple digital preservation workflow.**
- ✓ **Learn about common open source tools currently available to perform this work.**
- ✓ **Learn how to acquire and transfer digital files from a source using the tool DataAccessioner.**
- ✓ **Learn how to prepare files for upload to a preservation system using the tool Bagger.**
- ✓ **Learn how to create a checksum and check file fixity for digital materials using the tool Fixity to confirm they remain unchanged.**

Walk The Workflow

Walk, step-by-step, through an actual workflow for a sample case study, using simple tools on your laptops.

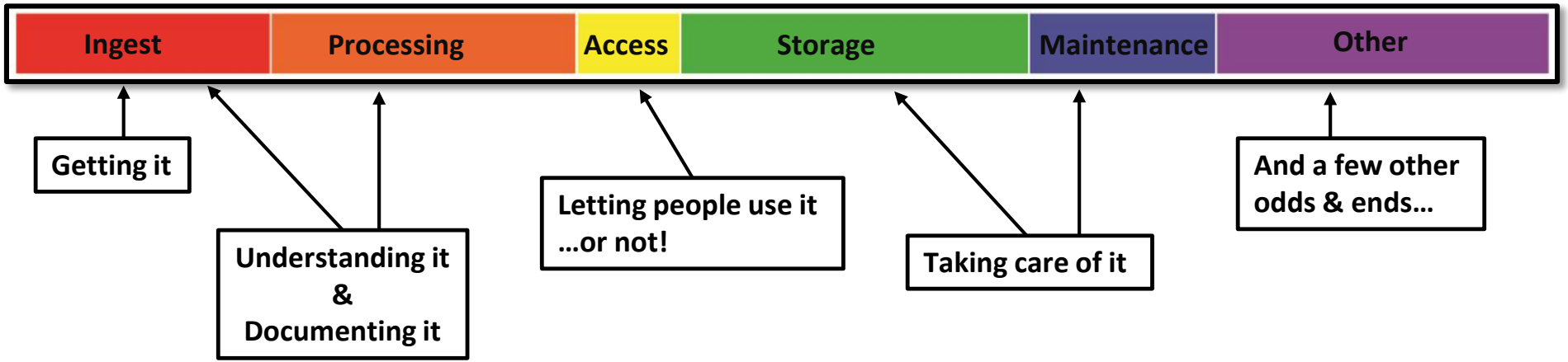
It’s ok (and important!) to triage what you have now

Remember: It’s a relay….not a marathon

Start with a simple workflow, with the expectation that it WILL change

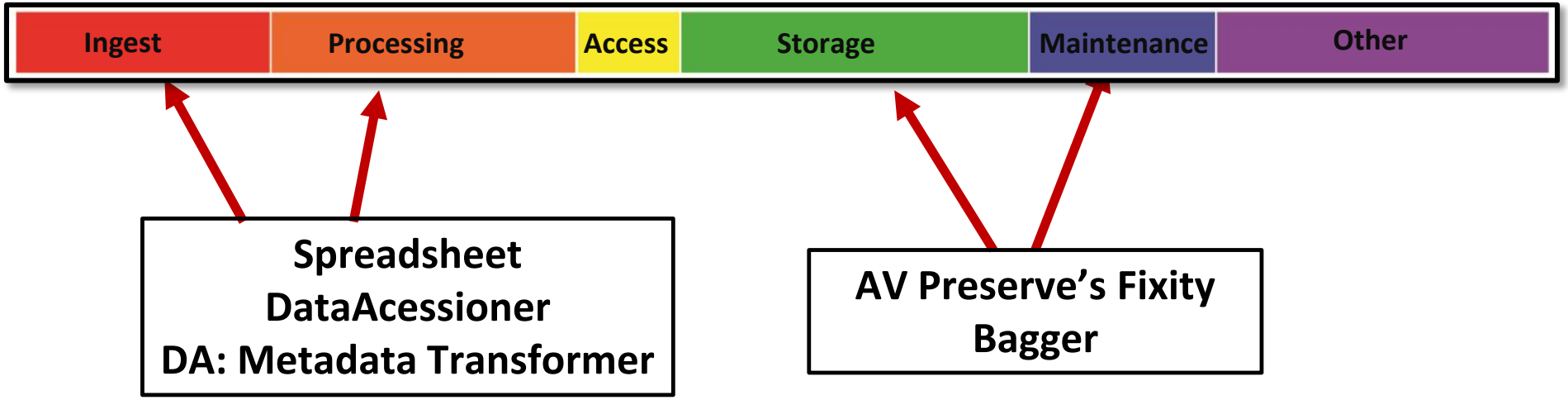
The POWRR Approach

Walk The Workflow



STEP 1 – *Don't Panic*

We've Acquired WHAT?!?!



There are other open-source tools that can perform these activities.

Use Cases

Backlog

What is THAT?

What is on it?

Digitization Workflow

Now what?

Born Digital Acquisitions

Huh?

"I'd like our institution to be the home for your literary papers."

~ gets handed flash drive ~

Actual Conversation, ca. 2004

Case Study

- Small, processed collection in The Archives entitled: "*The Archive's Furry Residents*"
- Contains CD's and floppy disks, among other things
- A collection record in Archon
- "*CD's and Floppy Disks – unknown content*"

Walk The Workflow

Starting from scratch:

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. Continue populating the Inventory **Spreadsheet.**
7. Once you've grabbed everything you can, place everything in standards-based packages using **Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

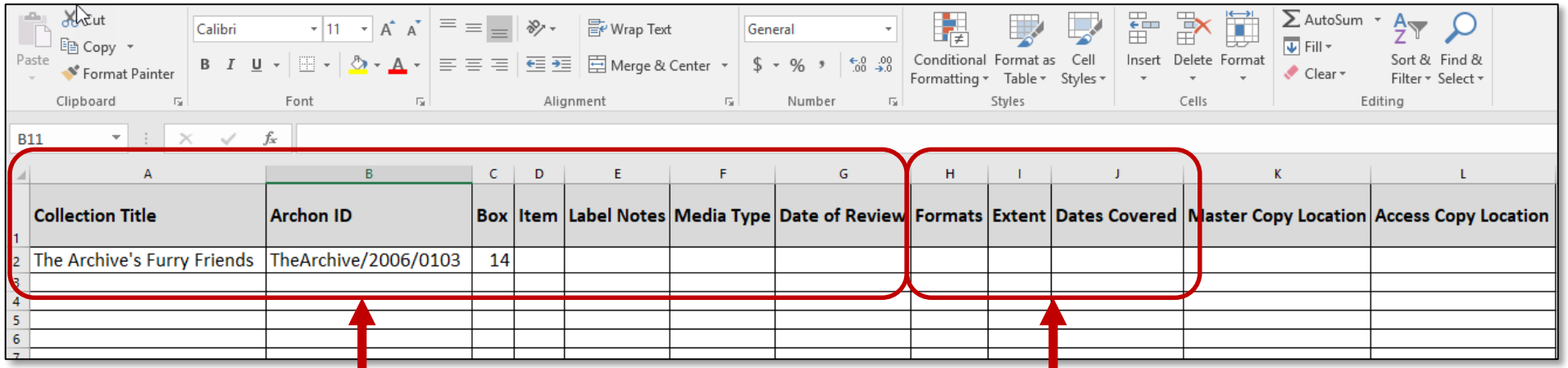
WTW – Spreadsheet

1. Collection Title
2. Archon ID
3. Box
4. Item
5. Label Notes
6. Media Type
7. Date of Review
8. Formats
9. Extent
10. Dates Covered
11. Master Copy Location
12. Access Copy Location

These are things we can't tell by just looking at the stuff

WTW – Spreadsheet

Ingest | Processing | Access | Storage | Maintenance | Other



	A	B	C	D	E	F	G	H	I	J	K	L
1	Collection Title	Archon ID	Box	Item	Label Notes	Media Type	Date of Review	Formats	Extent	Dates Covered	Master Copy Location	Access Copy Location
2	The Archive's Furry Friends	TheArchive/2006/0103	14									
3												
4												
5												
6												

Fill out what we can……

…and use DataAccessioner to discover this information

Walk The Workflow

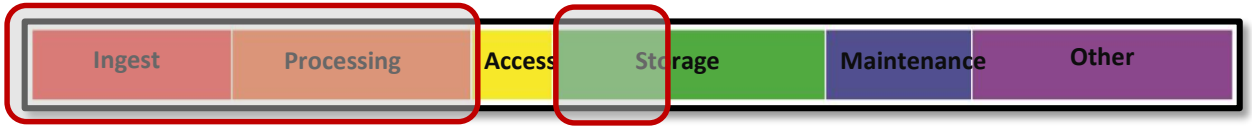
1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. **Run DataAccessioner.**

   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. Continue populating the Inventory **Spreadsheet.**
7. Once you've grabbed everything you can, place everything in standards-based packages using **Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

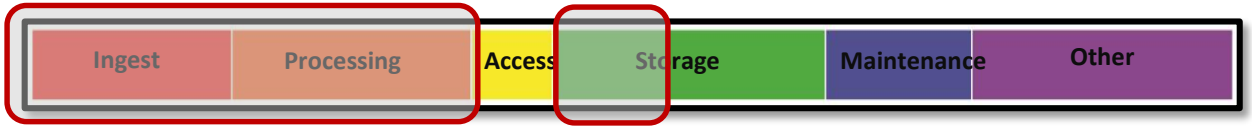
WTW – DataAccessioner



1. Collection Title
2. Archon ID
3. Box
4. Item
5. Label Notes
6. Media Type
7. Date of Review
8. Formats
9. Extent
10. Dates Covered
11. Master Copy Location
12. Access Copy Location

These are things we can't tell by just looking at the stuff

WTW – DataAccessioner



Who developed DataAccessioner?

- Originally created for tech services staff at Duke University RBMSC
- Updated by Seth Shaw for POWRR and other organizations
- dataaccessioner.org

What is DataAccessioner?

- It is a simple open-source tool with a user-friendly interface used to migrate content between media while also:
  - creating and validating checksums
  - gathering metadata (via FITS)
  - compiling an XML metadata file, with the option to include Dublin Core metadata as of v 1.0) for future reference.

On Your Flash Drives

- Digital_POWRR_Workshop_Tools_and_Hands_On_Activities
- Data Accessioner
- DataAccessioner_v1_1
- dataaccessioner-1.1
- Open the file named *dataaccessioner.jar*

XML – An Interlude

- XML = eXtensible Markup Language.
- Used to store and transport data
- Is readable by humans* and computers.
- Information in an XML file is stored in nested blocks that have opening and closing brackets.

**It actually is!!! You'll see…*

XML – Example 1

```
<?xml version="1.0" encoding="UTF-8"?>
 <note>
  <to>Jane</to>
  <from>John</from>
  <heading>A Note</heading>
  <body>Please bring the work files with you.</body>
 </note>
```

XML – Example 2

```
<?xml version="1.0" encoding="UTF-8"?>
  <books_to_purchase>
  <book>
     <name>What is XML?</name>
     <price>$35.95</price>
     <description> A book about XML. </description>
     <author>John Smith</author>
  </book>
  <book>
     <name>What is Digital Preservation?</name>
     <price>$55.95</price>
     <description> A book about Digital Preservation. </description>
     <author>Jane Doe</author>
  </book>
  </books_to_purchase>
```

Where is XML Used in Digital Preservation?

- XML files are used to store the metadata (a set of data that describes and gives information about other data) for the files in a digital collection.
- XML metadata files are produced by the various tools that are used to process and ingest the digital files to prepare them for long-term digital preservation.

  These XML files "describe" the properties of the original digital files that are ingested such as the

  - file format (including whether the file format is corrupted or not)
  - version of the file format (i.e. PDF file format version 2.0)
  - date the file was created
  - checksum of the file (to provide fixity)
  - description metadata you added yourself while ingesting the files…we added Dublin Core metadata using DataAccesssioner!

XML From *Furry Friends* Accession

Basic descriptive metadata you created

Extracted metadata:
Folder names
File names
Last modified
Size
…and more!

MD5 Checksum

Dublin Core Metadata

```
<?xml version="1.0" encoding="UTF-8"?>
- <collection name="The Archive's Furry Friends" xmlns="http://dataaccession[illegible].org/schema/dda-1-1">
  - <accession number="TheArchive/2006/0103_Box14_CD1">
      <ingest_note>The Archive's Furry Friends transferred by Jaim[illegible] Schumacher on Tue Nov 21
        10:10:36 CST 2017</ingest_note>
      <ingest_time>00:00:57.57178</ingest_time>
      <additional_notes>CD has a label that states "Orbit's stuff"</additional_notes>
    - <folder name="A Curator's Cat Collection" last_modified="2017-10-23T17:25:02.000">
        <dcx:description xmlns:dcx="http://purl.org/dc/xml/">
          <dc:description xmlns:dc="http://purl.org/dc/elements/1.1/">Contents of CD 1 located
            in Box 14 of collection TheArchive/2006/0103</dc:description>
          <dc:rights xmlns:dc="http://purl.org/dc/elements/1.1/">CC BY NC ND</dc:rights>
        </dcx:description>
        <folder name="Classic Kitties" last_modified="2017-10-23T17:24:58.000">
          - <file name="233_638576246007_2392_n.jpg" last_modified="2014-04-02T18:[illegible]:06.000"
            MD5="e285034d51e058a277b02132d2ffa11f" size="82873">
              <premis:object xsi:type="premis:file" xmlns:uuid="java:java.util.UUID"
              xmlns:fits="http://hul.harvard.edu/ois/xml/ns/fits/fits_output"
              xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
              xmlns:premis="info:lc/xmlns/premis-v2">
                - <premis:objectIdentifier>
                    <premis:objectIdentifierType>uuid</premis:objectIdentifierType>
                    <premis:objectIdentifierValue>49b856f3-1892-404d-a2bf-
                      7b8fbe40ee9f</premis:objectIdentifierValue>
                  </premis:objectIdentifier>
                - <premis:objectCharacteristics>
                    <premis:compositionLevel>0</premis:compositionLevel>
                  - <premis:fixity>
                      <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
                      <premis:messageDigest>e285034d51e058a277b02132d2ffa11f</premis:messag[illegible]
```

More XML From *Furry Friends* Accession

File Characterization Shenanigans!

```
<premis:size>82873</premis:size>
- <premis:format>
  - <premis:formatDesignation>
      <premis:formatName>JPEG File Interchange
        Format</premis:formatName>
      <premis:formatVersion>1.01</premis:formatVersion>
    </premis:formatDesignation>
  - <premis:formatRegistry>
      <premis:formatRegistryName>http://www.nationalarchives.gov.uk/pronom</premis:form
      <premis:formatRegistryKey>fmt/43</premis:formatRegistryKey>
    </premis:formatRegistry>
    <premis:formatNote>image/jpeg</premis:formatNote>
    <premis:formatNote>DROID Signature File Version: 88</premis:formatNote>
    <premis:formatNote>Identified by: Droid v6.1.5</premis:formatNote>
    <premis:formatNote>Identified by: Jhove v1.11</premis:formatNote>
    <premis:formatNote>Identified by: file utility v5.03</premis:formatNote>
    <premis:formatNote>Identified by: Exiftool v10.37</premis:formatNote>
    <premis:formatNote>Identified by: NLNZ Metadata Extractor
      v3.6GA</premis:formatNote>
  </premis:format>
</premis:objectCharacteristics>
<premis:originalName>233_638576246007_2392_n.jpg</premis:originalName>
```

Walk The Workflow

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

    Creates basic preservation metadata files in XML for you!

    Allows us to add descriptive metadata.

    Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).

4. **Make an Access Copy from the Master Copy.**
5. **Run the DA: Metadata Transformer Tool to make sense of the XML.**
6. Continue populating the Inventory Spreadsheet.
7. Once you've grabbed everything you can, place everything in standards-based packages using **Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

E:\Digital_POWRR_Workshop_Tools_and_Hands_On_Activities\Data Accessioner\da-mt-1.1\DAMetadataTransformer-1.1

WTW – DA Metadata Transformer Tool

Ingest | Processing | Access | Storage | Maintenance | Other

Coverts XML Into CSV (Comma Separated Value….a spreadsheet!)

	A	B	C	D	E	F	G	H	I
1	directory path	file name	file exten	last modified	size (bytes)	md5	file format	format ve	format re
2	A Curator's Cat Collection/Classic Kitties/	233_638576246007_2392_n.jpg	jpg	2014-04-02T18:27:06.000	82873	e285034d5	JPEG File Interchange Format	1.01	fmt/43
3	A Curator's Cat Collection/Classic Kitties/	252214_10101729763467027_865277491_n.jpg	jpg	2014-04-02T18:24:34.000	78193	2cf2ac108	JPEG File Interchange Format	1.02	
4	A Curator's Cat Collection/Classic Kitties/	396743_10101730863183187_1062547226_n.jpg	jpg	2014-04-02T18:24:50.000	32097	1d98f60a3	JPEG File Interchange Format	1.02	fmt/44
5	A Curator's Cat Collection/Classic Kitties/	475px-450px-50s_cats_large.jpg	jpg	2014-04-09T11:15:22.000	36620	108eaa26f	JPEG File Interchange Format	1.02	fmt/44
6	A Curator's Cat Collection/Classic Kitties/	543349_10101339612837017_413576055_n.jpg	jpg	2014-04-02T18:25:32.000	81353	346b060c0	JPEG File Interchange Format	1.01	
7	A Curator's Cat Collection/Classic Kitties/	59282_10102075760101997_1384684780_n.jpg	jpg	2014-04-02T18:22:50.000	69741	7dd3bc35	JPEG File Interchange Format	1.01	
8	A Curator's Cat Collection/Classic Kitties/	6-50s-cats-from-hubby.jpg	jpg	2014-04-09T11:22:40.000	142338	78b33b0c1	jpeg		
9	A Curator's Cat Collection/Classic Kitties/	76a77588d7d10632ba43a86a85f7efee.jpg	jpg	2014-04-09T11:15:46.000	381996	752f523fd	JPEG File Interchange Format	1.01	fmt/43
10	A Curator's Cat Collection/Classic Kitties/	Cats from the 1950's (3).jpg	jpg	2014-04-09T11:22:32.000	170740	076a92d23	JPEG File Interchange Format	1.01	fmt/43
11	A Curator's Cat Collection/Classic Kitties/	Cats from the 1950's (5).jpg	jpg	2014-04-09T11:22:26.000	277130	01877dac0	JPEG File Interchange Format	1.01	fmt/43
12	A Curator's Cat Collection/Classic Kitties/	champ-champ-pet-food-company-j-r-butland-cat-fo	jpg	2014-04-09T11:16:02.000	116145	d2292263(	JPEG File Interchange Format	1.02	fmt/44
13	A Curator's Cat Collection/Classic Kitties/	Thumbs.db	db	2014-04-09T11:54:40.000	22528	5441363cd	FPX		
14	A Curator's Cat Collection/Classic Kitties/	Thumbs.db.doc	doc	2014-04-15T12:06:00.000	22528	5441363cd	Microsoft Word Binary File Format		
15	A Curator's Cat Collection/Kitty Research/	animalplay.ppt	ppt	2014-04-15T12:06:04.000	4558848	1429e8cf5	Microsoft Powerpoint Presentation	97-2003 1	fmt/126
16	A Curator's Cat Collection/Kitty Research/	Behaviour_-_Cat_behaving_badly.pdf	pdf	2014-04-09T11:13:16.000	222865	ff4cd50d1	PDF/A	1b	
17	A Curator's Cat Collection/Kitty Research/	Cats Indoors! Slide Show - Wildlife Impacts.ppt	ppt	2014-04-09T11:11:34.000	7991808	58a86c588	Microsoft Powerpoint Presentation	97-2003	fmt/126
18	A Curator's Cat Collection/Kitty Research/	Cat_BasicCare.pdf	pdf	2014-04-09T11:12:38.000	107759	d2d3f866a	Portable Document Format	1.3	fmt/17
19	A Curator's Cat Collection/Kitty Research/	FelHusCh1.pdf	pdf	2014-04-09T11:13:00.000	7882504	252c45971	Portable Document Format	1.6	fmt/20
20	A Curator's Cat Collection/Kitty Videos/	1101810_10102319469226957_55721_n.mp4	mp4	2014-04-15T12:06:32.000	6837983	fb7bbe9fb	MPEG-4 Media File		fmt/199
21	A Curator's Cat Collection/Kitty Videos/	1253816_10102391726213377_17322_n.mp4	mp4	2014-04-15T12:06:36.000	2056636	5ad4c702(	MPEG-4 Media File		fmt/199
22	A Curator's Cat Collection/Kitty Videos/	Thumbs.db.doc	doc	2014-04-15T12:06:48.000	25088	6ad68295:	Microsoft Word Binary File Format		

Walk The Workflow

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

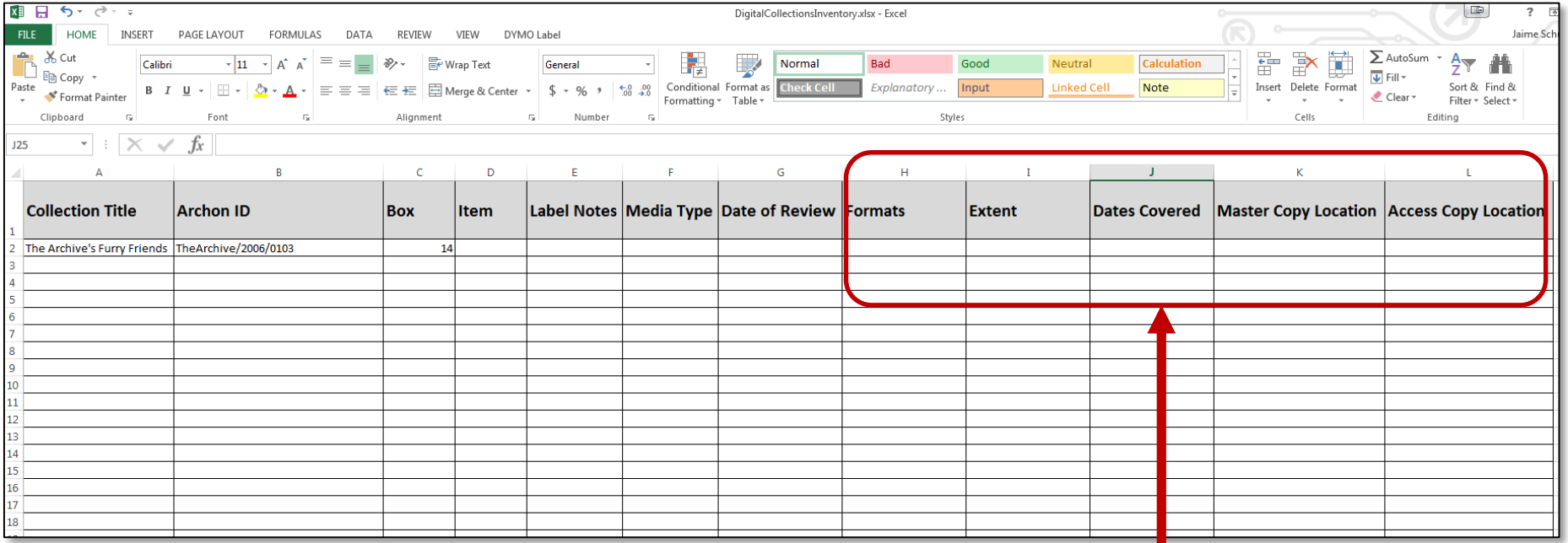
   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. **Continue populating the Inventory Spreadsheet.**
7. Once you've grabbed everything you can, place everything in standards-based packages using **Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

WTW – Spreadsheet

Ingest | Processing | Access | Storage | Maintenance | Other



Now we can fill this out!

Walk The Workflow

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. Continue populating the Inventory Spreadsheet.
7. **Once you've grabbed everything you can, place everything in standards-based packages using Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

Bags – An Interlude

What are bags?

Digital collection packed into a directory (the bag) along with a machine-readable manifest file (the tag) that lists the contents.

What can you store in bags?

Documents, pictures, music, movies, folders, etc. Anything digital.

What is the purpose of a bag?

To allow a sender to prepare a collection to send to a recipient that is off-site and allow the receiver to confirm all received contents.

Why use bags in digital preservation?

To help alleviate concern regarding the corruption or loss of files during transfer of content over a network.

Bags – Structure and Usage

Bags have 3 elements:

- A bag declaration text file, which acts as a seal of authenticity.
- A text-file manifest listing the files in the collection.
- A subdirectory – usually titled “data” filled with the digital content.

How Bags are used:

- The receiving computer analyzes the manifest file and then runs checksums on the contents in the bag.
- If the checksums match what is listed in the manifest, then the transfer is deemed successful.

WTW – Bagger

Who developed Bagger?

The Library of Congress

What is Bagger?

It is a digital records packaging and validation tool based on the BagIt specification.

How does Bagger work?

It allows creators and recipients of BagIt packages to verify that the files in the bag that was sent and received are complete and valid.

Manifests of the files that exist in the bag and their corresponding checksum values are created by Bagger and prepared for sending to a recipient.

The recipient uses those manifests to verify the bag and its content.

Walk The Workflow

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. Continue populating the Inventory Spreadsheet.
7. **Once you've grabbed everything you can, place everything in standards-based packages using Bagger.**
8. Setup the ongoing fixity monitoring of the Bags using **Fixity.**

E:\Digital_POWRR_Workshop_Tools_and_Hands_On_Activities\Bagger\bagger-2.7.6\bagger-2.7.6\bin\bagger.bat

Walk The Workflow

1. Begin an Inventory **Spreadsheet.**
2. Our PC has a CD drive, so we'll start with those while we look for a working floppy drive.
3. Run **DataAccessioner.**

   Creates basic preservation metadata files in XML for you!

   Allows us to add descriptive metadata.

   Moves everything to a stable carrier (The Archives has a network drive…we'll put stuff there).
4. Make an Access Copy from the Master Copy.
5. Run the **DA: Metadata Transformer Tool** to make sense of the XML.
6. Continue populating the Inventory Spreadsheet.
7. Once you've grabbed everything you can, place everything in standards-based packages using Bagger.
8. **Setup the ongoing fixity monitoring of the Bags using Fixity.**

WTW – Fixity

Who developed Fixity?

AV Preserve: https://www.avpreserve.com/products/fixity/

What is Fixity?

It is a simple open-source tool that automatically monitors and reports on the data integrity of selected digital content.

How does Fixity work?

It scans a folder or directory and creates a manifest of the files, including their file paths and their checksums, against which a regular comparative analysis can be run.

It monitors file integrity through the generation and validation of checksums, and file attendance through monitoring and reporting on new, missing, moved and renamed files.

Checksums – An Interlude

A **file checksum** is a calculated string of number and letters that acts as a fingerprint for the particular file that it was calculated from.

Why are they used?

To ensure the integrity of a file after it has been transmitted from one storage device to another

To confirm that a file has not degraded or corrupted after being stored on a device for a period of time (compare previous stored checksum to recalculated current value).

With some limitations, it can also provide assistance in determining if a file or files have been modified since they were ingested.

How are they calculated?

Checksums are calculated using hash functions. Hash functions are mathematical functions. (Md5, SHA-1, SHA-256, etc.)

Checksums – An Interlude

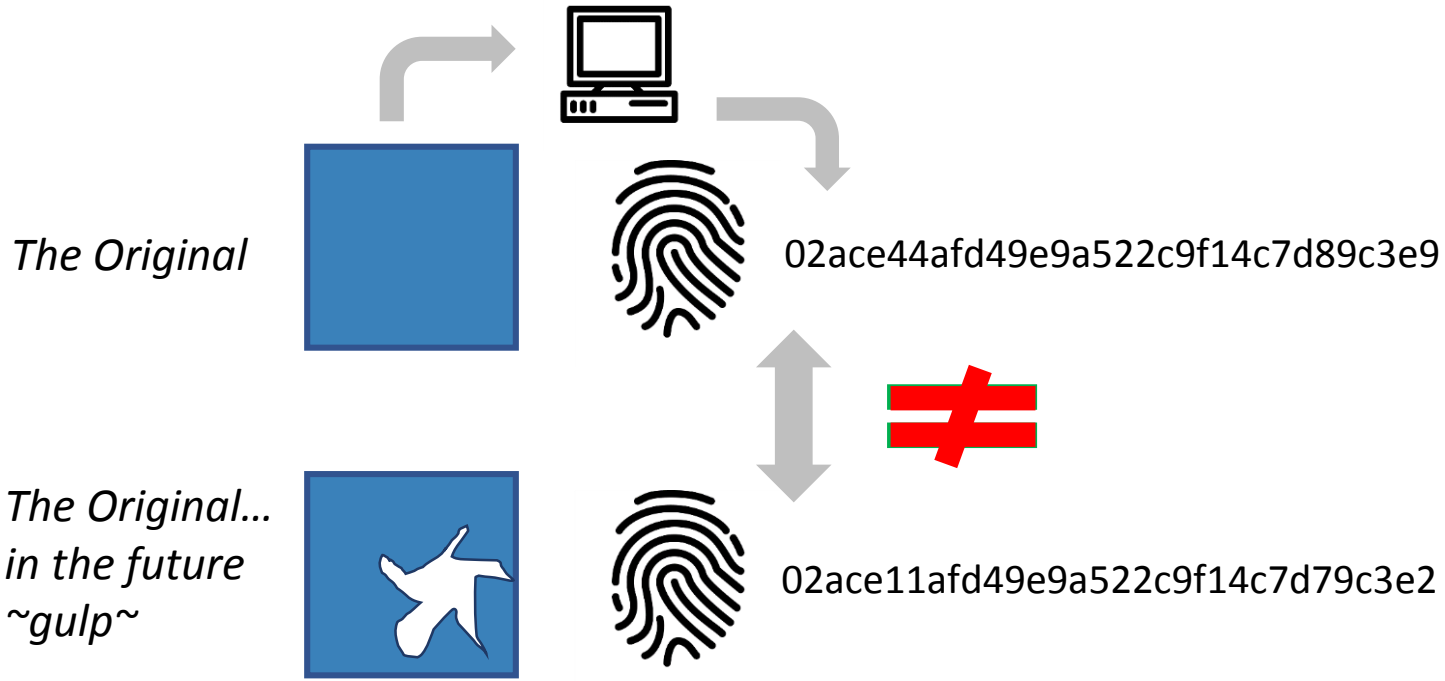


Image by Arthur Shlain from the Noun Project

WTW – Fixity

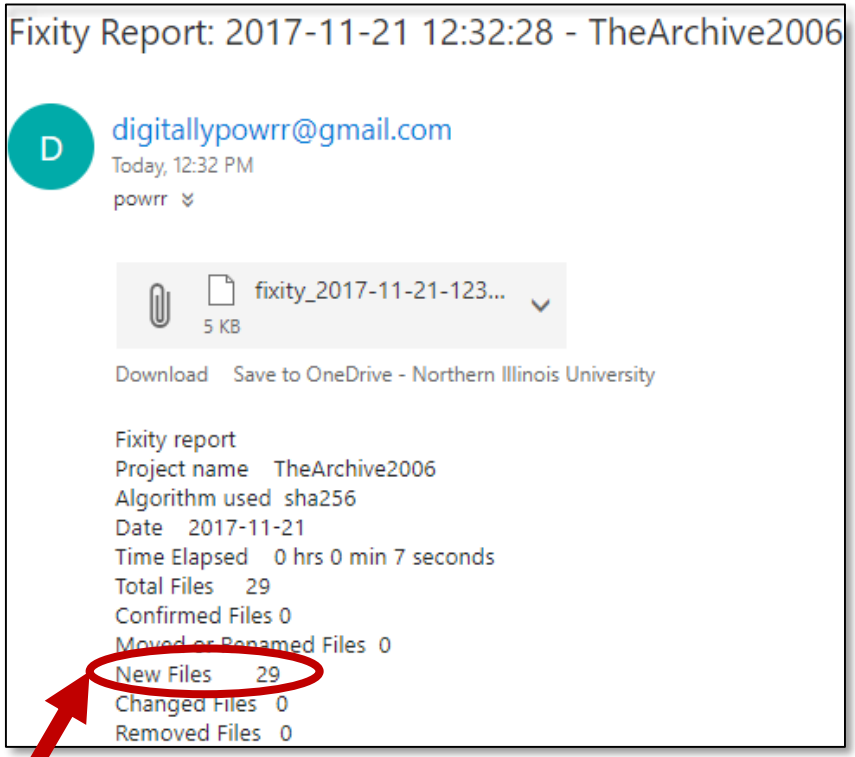
Let's give Fixity a whirl!

E:\Digital_POWRR_Workshop_Tools_and_Hands_On_Activities\Fixity\Fixity for Windows\fixity-win-0.5\fixity-win

OR

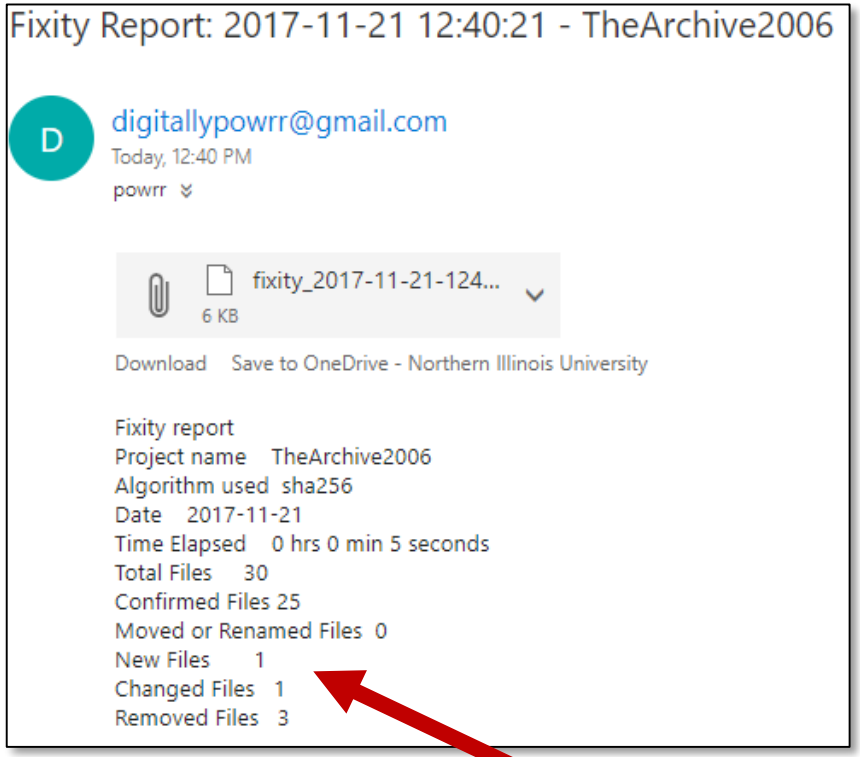
E:\Digital_POWRR_Workshop_Tools_and_Hands_On_Activities\Fixity\Fixity for Mac

New Project



Since it's a new project, *Total Files* and *New Files* are the same. *Confirmed Files* is 0.

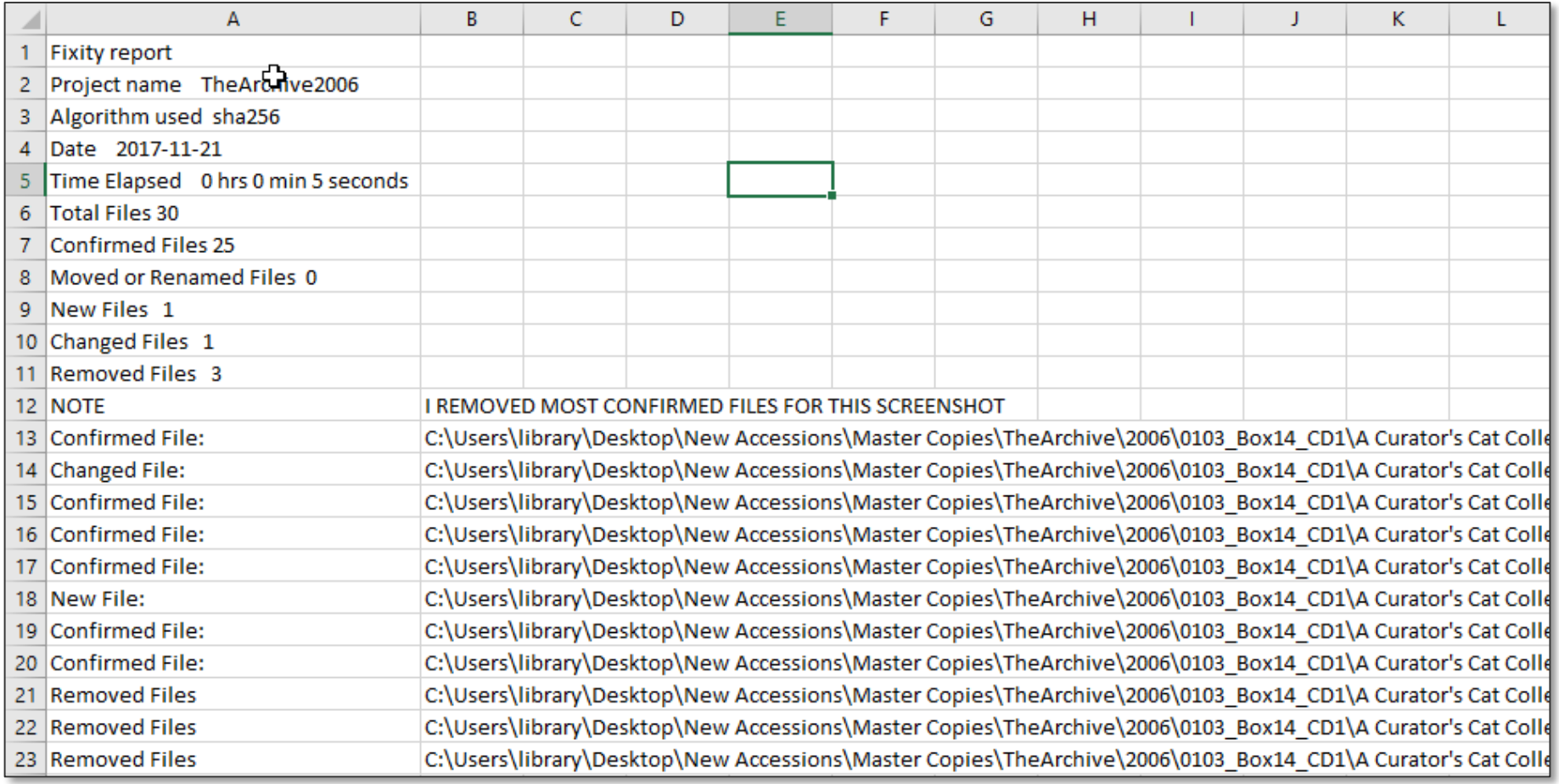
We Have a Problem



But if there is a problem, Fixity will tell you.

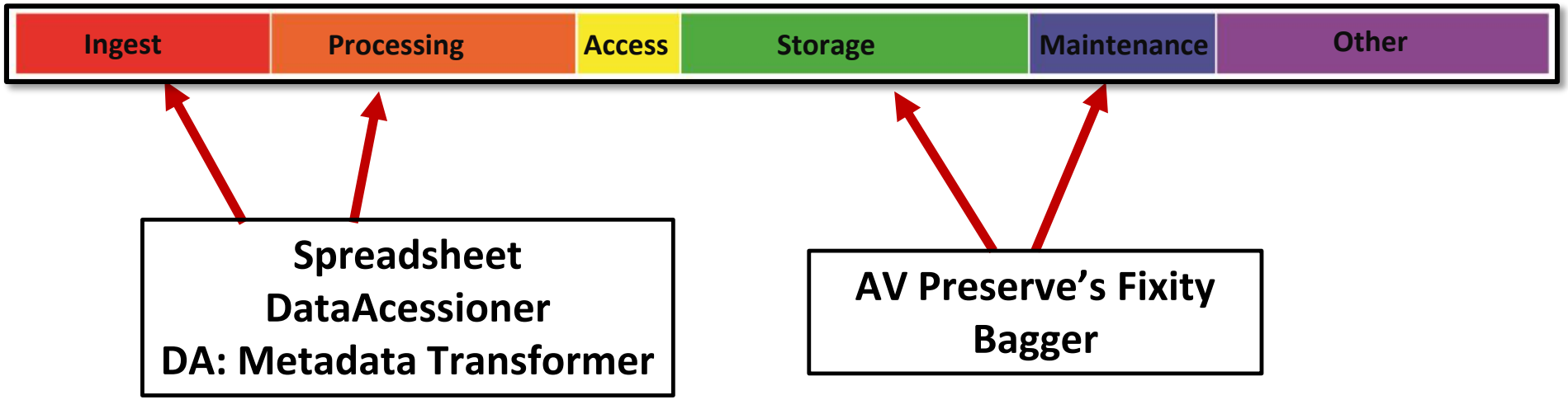
If the monitored files are unchanged
Total Files
and
Confirmed Files
will be the same.

Fixity Report



	A	B	C	D	E	F	G	H	I	J	K	L
1	Fixity report											
2	Project name TheArchive2006											
3	Algorithm used sha256											
4	Date 2017-11-21											
5	Time Elapsed 0 hrs 0 min 5 seconds											
6	Total Files 30											
7	Confirmed Files 25											
8	Moved or Renamed Files 0											
9	New Files 1											
10	Changed Files 1											
11	Removed Files 3											
12	NOTE	I REMOVED MOST CONFIRMED FILES FOR THIS SCREENSHOT										
13	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
14	Changed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
15	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
16	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
17	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
18	New File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
19	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
20	Confirmed File:	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
21	Removed Files	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
22	Removed Files	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										
23	Removed Files	C:\Users\library\Desktop\New Accessions\Master Copies\TheArchive\2006\0103_Box14_CD1\A Curator's Cat Colle										

We Walked The Workflow!!



Technology Module:
Walk The Workflow

QUESTIONS?