CAPTURE, INGEST, & CHECKSUM TOOL

DOCUMENTATION

Lead developer: Dean Farrell
Development support: Michael Kimsal
Technical support: Lisa Gregory, Jennifer Howard, Kathleen Kenney, & Amy Rudersdorf

Partner organizations:
Elon University
North Carolina State Archives
University of North Carolina at Charlotte

CINCH is a product of the State Library of North Carolina, a division of the North Carolina Department of Cultural Resources.

The development of the CINCH tool was supported through an Institute of Museum & Library Services 2011 Sparks! Ignition grant.

1. OVERVIEW

At a basic level, CINCH is a process involving many actions[1] designed to locate targeted files on the internet and download them in a preservation-ready state. This includes maintaining the files' integrity by virus checking and repeated checksumming, as well as enhancing the files' context with metadata extraction.

It is helpful to think of CINCH as an assembly line, with a series of deliberate actions being performed on a single file in progression. At each step, CINCH attempts to perform the desired action; whether successful or not, each action and its result is recorded for the user's review. The result is a zipped file or files that contains the original files targeted by the user, their associated metadata, and an audit trail document outlining every action taken upon the files.

2. TABLE OF CONTENTS

This guide contains the following sections:

[1] Definitions of many of these terms can be found at the end of this document.

2. HOW TO USE CINCH…IN 5 EASY STEPS!

1. Create a file that contains URLs of all of the files you'd like CINCH to download. We'll call this the "files list." You can create this list manually, use information from your web harvester to get those URLs, or use another tool like a file list utility or site map generator. Files should…
    a. have one URL per line;
    b. point to files with allowable file extensions (.csv, .doc, .docx, .gif, .jpg, .pdf, .png, .ppt, .pptx, .txt, .xls, .xlsx; files that end in .asp, .aspx, .jsp, .php, .cfm, and .cfml can also be included, if they lead to dynamically generated .pdf files);
        i. contain the full path to the file, including "http(s):// " [e.g., http://www.tryonpalace.org/pdfs/in_convention_8-1-1788.pdf];
    c. contain between 1 and 4500 URLs; and
    d. be formatted as .txt or .csv.
2. Log in to CINCH at http://cinch.nclive.org. If you do not have a CINCH log in, email **digital.info@ncdcr.gov** to obtain one.[2]
3. Upload your "files list" following the directions on the CINCH website. CINCH will attempt to download each file in your "files list" within 24 hours. Once CINCH begins processing the URLS, you can log out.
4. Once complete, you will receive an email from CINCH alerting you that your zip file(s) are ready for download. The zip file(s) will contain:
    a. A separate metadata file for each file type.
    b. If errors occurred, an error list that includes:
        i. the original URL you submitted,
        ii. the file name (if the error occurs on a local file), and
        iii. an error message.
    c. An event list that includes:
        i. the original URL you submitted,
        ii. the file name,
        iii. the event name, and
        iv. the date and time the event occurred.
    d. A manifest that lists all of the downloaded files in the zip file.
    e. All of the actual files from the "files list" that CINCH was able to locate.

---

[2] Users must be affiliated with a government agency, library, archive, historical society, or related organization or institution in North Carolina to obtain a log in. Otherwise, CINCH is available at github for download at https://github.com/SLNC-DIMP/Cinch.

f. A folder containing all of the "problem files."

Depending on the number and size of your requested files, CINCH may create more than one zip file.

5. Login to CINCH and go to the "download your files" link to retrieve your files within 30 days.

3. WHAT CINCH DOES...

The sections below describe CINCH's actions in detail. CINCH follows good preservation practice whenever it is performing an action. Any action, whether successful or not, is recorded. These recorded actions can be reviewed in the event log.

This document uses two symbols to help you understand what CINCH does when it encounters an error and how it continues:

 = An error. CINCH is unable to proceed with actions on that file.

 = Success! CINCH will process the file and continue to the next step.

...JUST AFTER A USER LOGS IN

1. When the user uploads their "files list," CINCH adds the URLs from the "files list" to its database for downloading and processing. Once CINCH begins processing the URLs, the user can log out of the system.

    If the "files list" isn't completely uploaded, CINCH will pick up where it left off the next time the action is run.

    CINCH continues.

...TO THE REMOTE FILES

2. CINCH tries to locate the URLs the user has requested in the "files list."

    If the file at a requested URL is not found, it is skipped and added to the error list.

    CINCH continues.

3. The remote files are checksummed using the algorithm SHA-1.

    If a checksum can't be computed, it is noted in the error list, and the file is not downloaded.

    CINCH continues.

4. The requested URLs are checked to see whether they are duplicates from a previous download by the current account to which the user is logged in.

   If it appears that a file has been downloaded previously with CINCH, it is noted in the error list, and the file is put in the "problem_files" folder. All actions performed on local files, mentioned below, will also take place on this file.

   CINCH continues.

5. The remote files are checked to see if they have allowable file extensions (.csv, .doc, .docx, .gif, .pdf, .png, .ppt, .pptx, .txt, .xls, .xlsx).

   If a file has a non-allowable extension, it is not downloaded and an error is noted in the error list.

   If it has no file extension or one of the following file extensions (.asp, .aspx, .jsp, .php, .cfm, and .cfml) CINCH creates a dynamically generated .pdf and continues.

6. The remote files are assigned modified names.
   a. Spaces and unusual characters are stripped out and replaced with underscores.
   b. If two or more files have the same name, a random number is appended to the end of the file name to prevent overwriting of files.

   CINCH continues.

7. The remote files' sizes are assessed to determine if each file is .4 GB (409.6 MB) or smaller.

   If a file is .4 GB (409.6 MB) or larger, it won't fit into a .zip archive and it is not downloaded.

   CINCH continues.

Actions on the remote file end here. To review, if a file can be checksummed, its file extension has been determined to be supported, and is within the specified file size limit, CINCH downloads and saves the file.

...TO THE DOWNLOADED FILES

8. CINCH checks whether the downloaded files have retained their original last modified dates/times.

   
   If the last modified date/time was not retained for a file, it is reset to the last modified date/time the downloaded file had on the remote server.

   
   CINCH continues.

9. Each file is checked for viruses using ClamAV.

   
   If the scan fails or ClamAV detects a virus, the file is deleted.

   
   CINCH continues.

10. Downloaded files are checksummed (SHA-1).

    
    If a file cannot be checksummed, the checksum creation error will appear in the error list and the file will be added to the problem_files folder of your zip download.

    
    CINCH continues.

11. Names and checksums are checked against all other files that user has downloaded through CINCH.

    
    If a duplicate checksum and/or a duplicate name is detected, an error is noted in the error list and the file is added to the problem_files folder of your zip download.

    
    CINCH continues.

12. Using Apache Tika, metadata is extracted. Each file type has slightly different extracted metadata. Below is the list of metadata fields extracted from text-based files:

Metadata Field	File Location Derived From
Author	The file header - embedded metadata
Creation date and time	The file header - embedded metadata
Last modified date and time	The file header - embedded metadata
Creator	The file header - embedded metadata
Producer	The file header - embedded metadata
File name	The original URL, stripped of special characters. May include a unique number as well, to avoid duplication.
Title	The file header - embedded metadata
Number of pages	The file header - embedded metadata (a null return indicates the number of pages could not be determined)
Subject	The file header - embedded metadata
Keywords	The file header - embedded metadata
Licensed to	The file header - embedded metadata
Possible title	Best guess, extracted from full text (if any)
Possible keywords	Best guess, extracted from full text (if any)
Checksum (SHA-1)	Computed by CINCH
Whether or not full text exists	Determined by CINCH

If the file's media type[3] is text-based, text is extracted and processed on a very basic level for a possible title and keywords.

13. The files are checksummed one more time to ensure that nothing went amiss between downloading the file and the current time.

If the currently generated checksum for a file doesn't match the checksum generated during the initial checksum creation an error is noted in the error list and the file is added to the problem_files folder of your zip download.

CINCH continues.

14. The files are zipped. Each zip file contains:
    a. A separate metadata file for each file type.
    b. If errors occurred, an error list that includes:
        i. the original URL submitted by the user,
        ii. the file name (if the error occurs on a local file), and
        iii. an error message.
    c. An event list that includes:
        i. the original URL submitted by the user,

[3] Formerly known as MIME types.

ii. the file name,
iii. the event name, and
iv. the date and time the event occurred.

d. A manifest that lists all of the downloaded files in the zip file.
e. All of the actual files from the "files list" that CINCH was able to locate.
f. A folder containing all of the problem files.

Each zip file can only contain 0.5 GB or 65,500 files. If the entire download is larger or has more files, the files will be split into multiple zip files.

15. The event file is placed in the very last zip file associated with a user's "files list."

...IN THE FINAL STEPS

16. CINCH emails the user that their requested files are available for download once it has completed all of the processes.
17. To download files, the user logs in to CINCH and clicks on the "Download your files" link.
18. After 30 days all user files that are at least 30 days old are purged from the file system. Even after the files are deleted, however, CINCH retains all of the files' checksums, original URL paths, metadata, and event history in order to check future uploads for file duplicates. To gain access to this data, email digital.info@ncdcr.gov.

APPENDIX A: CINCH GLOSSARY

DEFINITIONS OF TERMS IN THE TEXT AND/OR WHY THIS IS IMPORTANT TO THE CINCH PROCESS.

- **action**: step in the process in which a file is acted upon to achieve a desired outcome.
- **Apache Tika**: software tool that analyzes text, extracts metadata, and detects media types.
- **ClamAV**: software that detects malware and viruses.
- **checksum/med/ming**: mathematical value determined by a specific algorithm, used to verify data.
- **.csv**: comma-separated values, a type of file where data is separated by commas.
- **file extensions**: suffix added to the end of a computer file name to indicate the type of file.
- **integrity**: the quality of being whole and unaltered through loss, tampering, or corruption.
- **last modified date/time**: time stamp set by software programs to mark when a file was last altered.
- **media type**: standardized two-part identifier to classify file formats on the Internet.
- **preservation-ready**: file that is in a suitable state for digital preservation, a series of managed activities necessary to ensure continued access to digital materials for as long as they are needed.
- **remote**: refers to a file stored at a location distant from the user.
- **SHA-1**: secure hash algorithm, used to ensure data integrity. Output from checksum process.
- **zipped file, or .zip**: file format used for data compression in order to reduce file size.

APPENDIX B: CINCH OUTPUTS, ERROR MESSAGES, EVENTS, & DATA TABLES

A. DOWNLOADED FILES DIRECTORIES

CINCH places all files directly into the .zip file, with the exception of problem files. These will be found in their own problem_files directory within the .zip file.

B. ERROR MESSAGES

Types of errors that are recorded and their corresponding error messages:

Error Message found in Error File	What the Message Means
Unable to download file	CINCH either could not find, was unable to complete the file download, or the file was too large (over .4GB)
Could not create checksum	A remote file can't be checksummed
Duplicate checksum	CINCH finds two files with the same checksum
Unable to extract file metadata	CINCH is unable to extract file metadata
Corrupt File. Checksum mismatch	Checksum generated for the file doesn't match the checksum stored in the database for the file
Unable to add file to Zip download	
Unknown error	
Virus detected	CINCH detects a virus. The file is deleted.
Unsupported file type	A file does not have an allowable extension, either natively or after attempting to convert dynamically to a.pdf.
Unable to delete file	
Unable to determine full text status	CINCH cannot determine whether or not the text-based file has full text.
Virus check couldn't scan file	Virus check could not be completed.
Duplicate filename	Multiple files have the same filename.
File media-type doesn't match file extension	The type of file noted in the metadata does not match the file extension.

C. FILE EVENTS AS LISTED IN THE EVENT LIST

These events may be found in the event list document created by CINCH.

Event	Where in the process this occurs
Downloaded	After Step 7
Renamed	Step 6
Download last modified time corrected	Step 8
Virus check	Step 9
Checksum creation	Step 3 and step 10
Deleted - virus detected	Step 9
Metadata extraction	Step 12
File integrity check	Step 13

Zipped for download	Step 14
Deleted - expired	30 days after the files was downloaded
Full text check	Step 12
Download failed	After Step 7

E. TABLES

These are the tables in the CINCH database. Users won't ever see these; they are provided here for general documentation purposes and for individuals interested in working with the CINCH code.

- file_event_history: This is an audit trail document, which records file errors before download, downloads, virus check results, and checksum.
- file_info: When requested files are checked remotely, this table keeps track of file checksums, whether or not the file exists, whether or not the file extensions are allowable, whether or not the files are too large, and whether all CINCH tasks have been run on the file. It includes the file's last modified time.
- files_for_download: contains the files listed in files lists. As files are downloaded, they are marked as "processed" in this table.
- problem_files: Tracks download errors, such as whether or not the file exists, whether or not the file extensions are allowable, whether or not the files are too large, and problems with checksumming.
- uploads: tracks the files lists uploaded by CINCH users.

APPENDIX C: GRAPHICAL REPRESENTATION OF CINCH PROCESSES

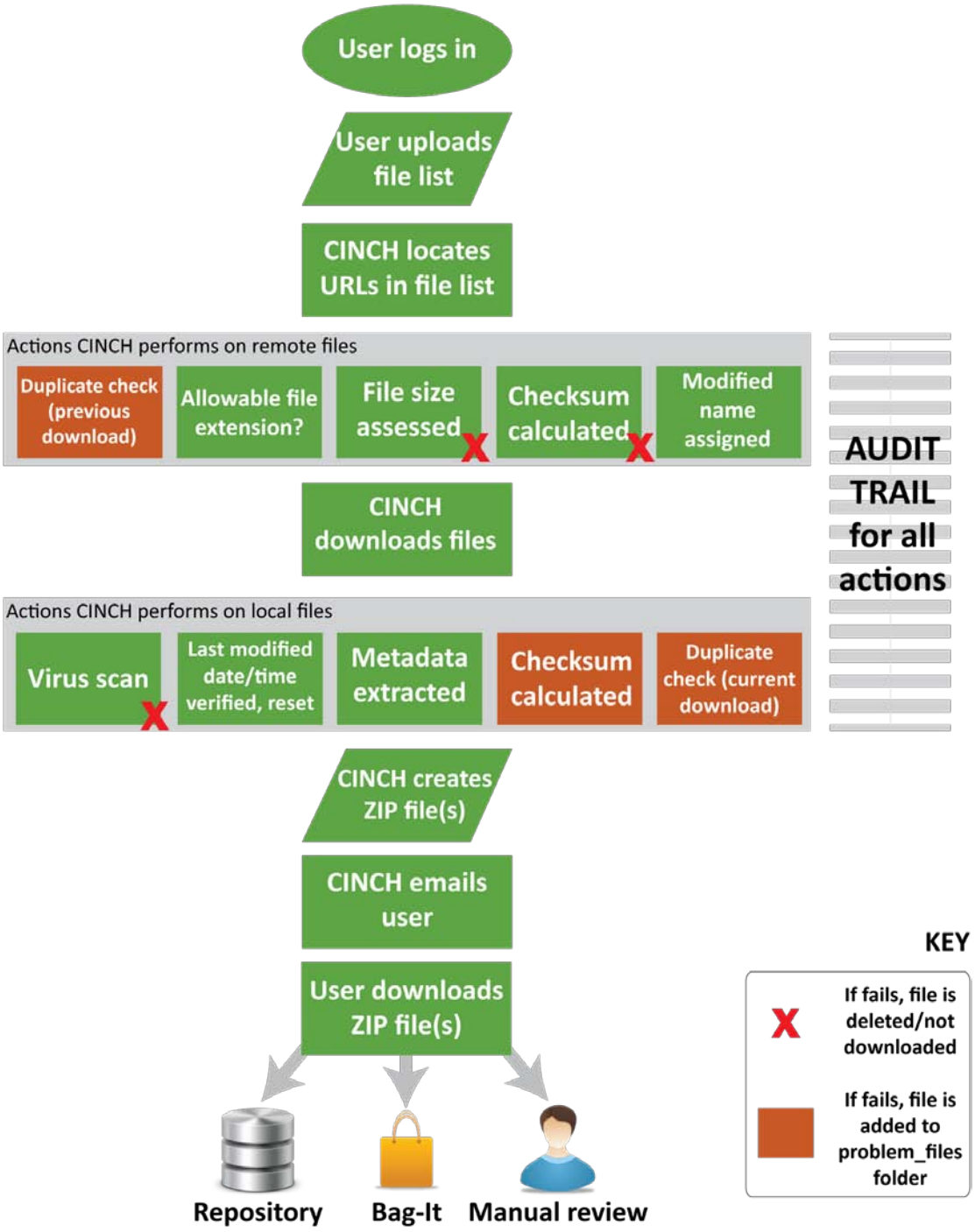


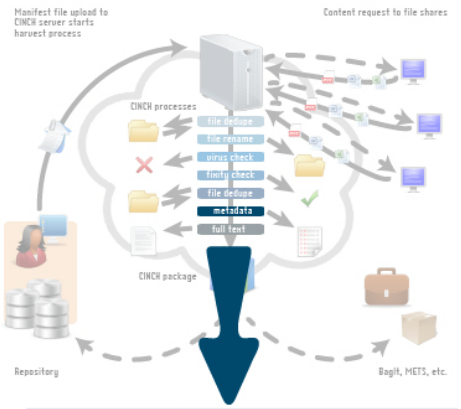
Image 1. Process begins with a user logging in and manually uploading a files list to the CINCH server. All processing actions are automated. A SIP is produced for ingest or to repurpose according to the repository's specific needs.

curl_files.zip - ZIP archive, unpacked size 1,642,349 bytes

Name	Size	Packed	Type	Modified	CRC32
..			File Folder		
abcs.ncpublicschools.org_abcs_ser...	243,080	224,896	Adobe Acrobat D...	2/17/2012 3:...	EA6272EE
abcs.ncpublicschools.org_abcs_ser...	243,080	224,899	Adobe Acrobat D...	2/17/2012 3:...	BF547B02
abcs.ncpublicschools.org_abcs_ser...	243,080	224,894	Adobe Acrobat D...	2/17/2012 3:...	C27248A7
abcs.ncpublicschools.org_abcs_ser...	243,080	224,894	Adobe Acrobat D...	2/17/2012 3:...	EB387D7C
abcs.ncpublicschools.org_abcs_ser...	243,080	224,896	Adobe Acrobat D...	2/17/2012 3:...	CBF849ED
abcs.ncpublicschools.org_abcs_ser...	243,080	224,898	Adobe Acrobat D...	2/17/2012 3:...	F550EBC9
daq.state.nc.us_rules_rules_D1701...	10,384	5,284	Adobe Acrobat D...	3/10/2008 8:...	3152D6AE
daq.state.nc.us_rules_rules_D1701...	10,384	5,284	Adobe Acrobat D...	3/10/2008 8:...	3152D6AE
daq.state.nc.us_rules_rules_D1702...	11,123	6,009	Adobe Acrobat D...	3/10/2008 8:...	DB1DF2...
daq.state.nc.us_rules_rules_D1702...	11,123	6,009	Adobe Acrobat D...	3/10/2008 8:...	DB1DF2...
daq.state.nc.us_rules_rules_D1703...	11,924	6,829	Adobe Acrobat D...	3/10/2008 8:...	1AC8B490
daq.state.nc.us_rules_rules_D1703...	11,924	6,829	Adobe Acrobat D...	3/10/2008 8:...	1AC8B490
daq.state.nc.us_rules_rules_D1704...	10,551	5,435	Adobe Acrobat D...	3/10/2008 8:...	EC9FAC...
daq.state.nc.us_rules_rules_D1704...	10,551	5,435	Adobe Acrobat D...	3/10/2008 8:...	EC9FAC...
daq.state.nc.us_rules_rules_D1705...	11,441	6,358	Adobe Acrobat D...	3/10/2008 8:...	33271F78
daq.state.nc.us_rules_rules_D1705...	11,441	6,358	Adobe Acrobat D...	3/10/2008 8:...	33271F78
daq.state.nc.us_rules_rules_D1706...	10,831	5,723	Adobe Acrobat D...	3/10/2008 8:...	109080B0
daq.state.nc.us_rules_rules_D1706...	10,831	5,723	Adobe Acrobat D...	3/10/2008 8:...	109080B0
daq.state.nc.us_rules_rules_D1707...	11,150	6,057	Adobe Acrobat D...	3/10/2008 8:...	D58D9E...
error_files.csv	2,100	331	Microsoft Office ...	2/17/2012 4:...	F08DF5E1
event_list.csv	20,650	1,302	Microsoft Office ...	2/17/2012 4:...	DA4134F0
file_manifest.csv	1,252	342	Microsoft Office ...	2/17/2012 4:...	D6A0FA...
PDF_Metadata.csv	16,209	1,340	Microsoft Office ...	2/17/2012 4:...	DB71A289

Total 1,642,349 bytes in 23 files

Image 2. Contents of example ingest package. PDFs are downloaded files; .csv files contain event list, error list, a manifest of the files downloaded, and automatically generated metadata.



METADATA FIELD	SAMPLE VALUE
author	DPI
creation_date	2012-01-17T11:31:57Z
last_modified	2012-01-17T11:31:57Z
creator	JFreeReport version 0.8.3e
producer	iText by lowagie.com (r1.00 - ps122)
resource_name	[CINCH-unique-filename].pdf
title	2010-2011 Growth and Performance
pages	63
subject	[blank]
keywords	[blank]
licensed_to	[blank]
possible_doc_title	Climate Ready EstuariesA Blueprint for Change
possible_doc_keywords	sea-level, rise, climate, local, water
Checksum	b2b764fa0d71fba9b48c8af4e9dbcf2459ce38d8
Full text	yes

Image 3. Example of auto-generated metadata. "Subject", "keyword", and "licensed-to" can be populated based on the value in the "author" field. In this example, "DPI" (Department of Public Instruction) might correspond to the LC name authority file for the state agency in the "subject" field. "Education" or "state agency" could be tied to DPI and automatically populate the "keyword" field and "licensed-to" might be auto-populated with a rights statement based on the author's publishing procedures.

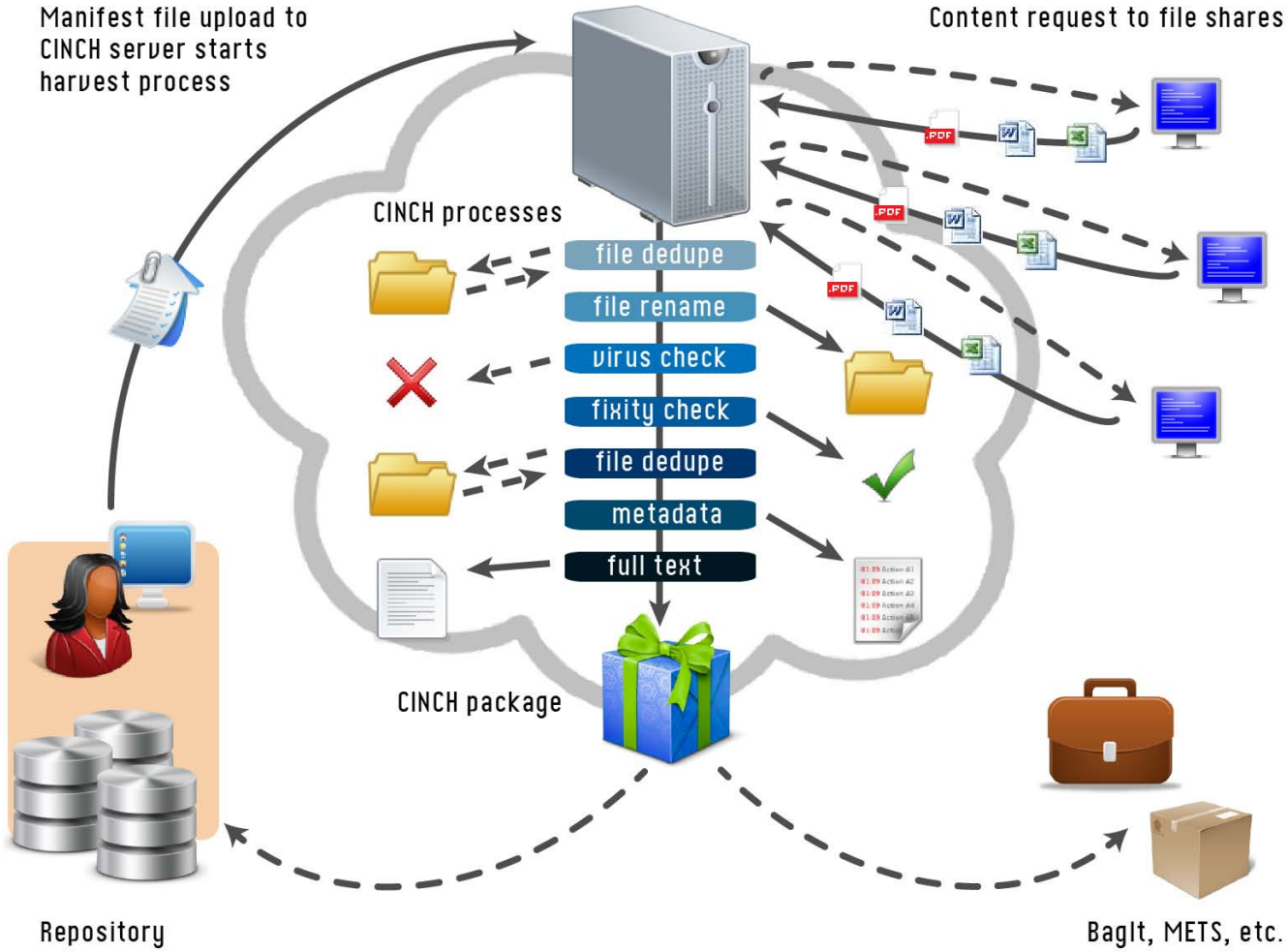


Image 4. The entire CINCH process.