Collecting born-digital materials from the web: It's a CINCH!

Lisa Gregory
State Library of North Carolina
Digital Preservation Interest Group, June 24, 2012

CINCH (Capture INgest CHecksum) is a tool that automates the transfer of online content to a repository, using ingest technologies appropriate for digital preservation.

More familiarly, CINCH

- grabs freely available online content,
- authenticates it,
- extracts metadata, and
- readies it for repository ingest.

- Modular
- Flexible
- Easy to use
- Repository-neutral
- Open source
- (For North Carolina libraries) Hosted

State Library of North Carolina

Collection: State government, genealogy, North Caroliniana

Support: All North Carolina libraries

Services & Materials: for those with visual/physical disabilities

State Library of North Carolina

Collection: State government, genealogy, North Caroliniana

Support: All North Carolina libraries

Services & Materials: for those with visual/physical disabilities

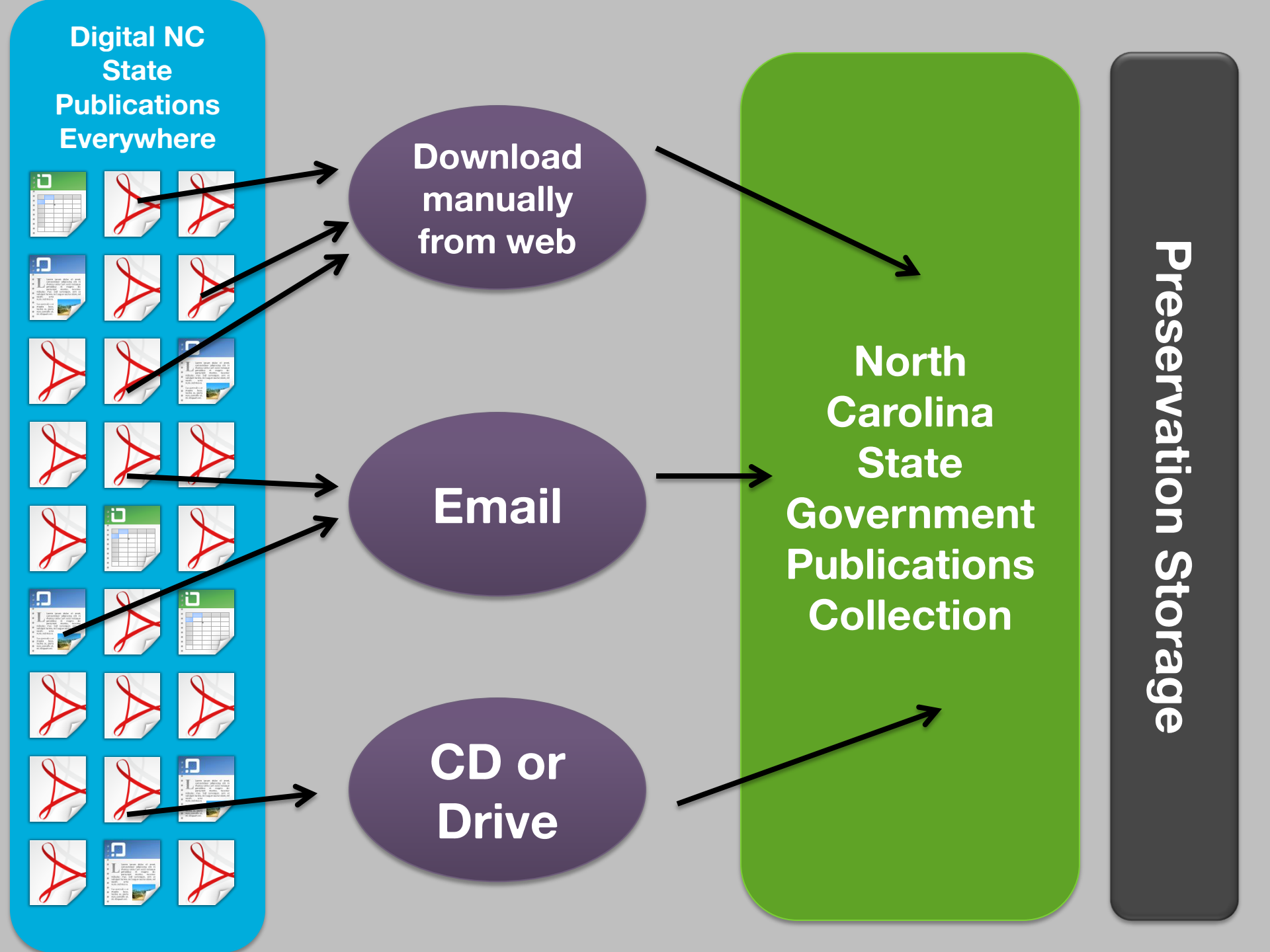
North Carolina General Statute 125

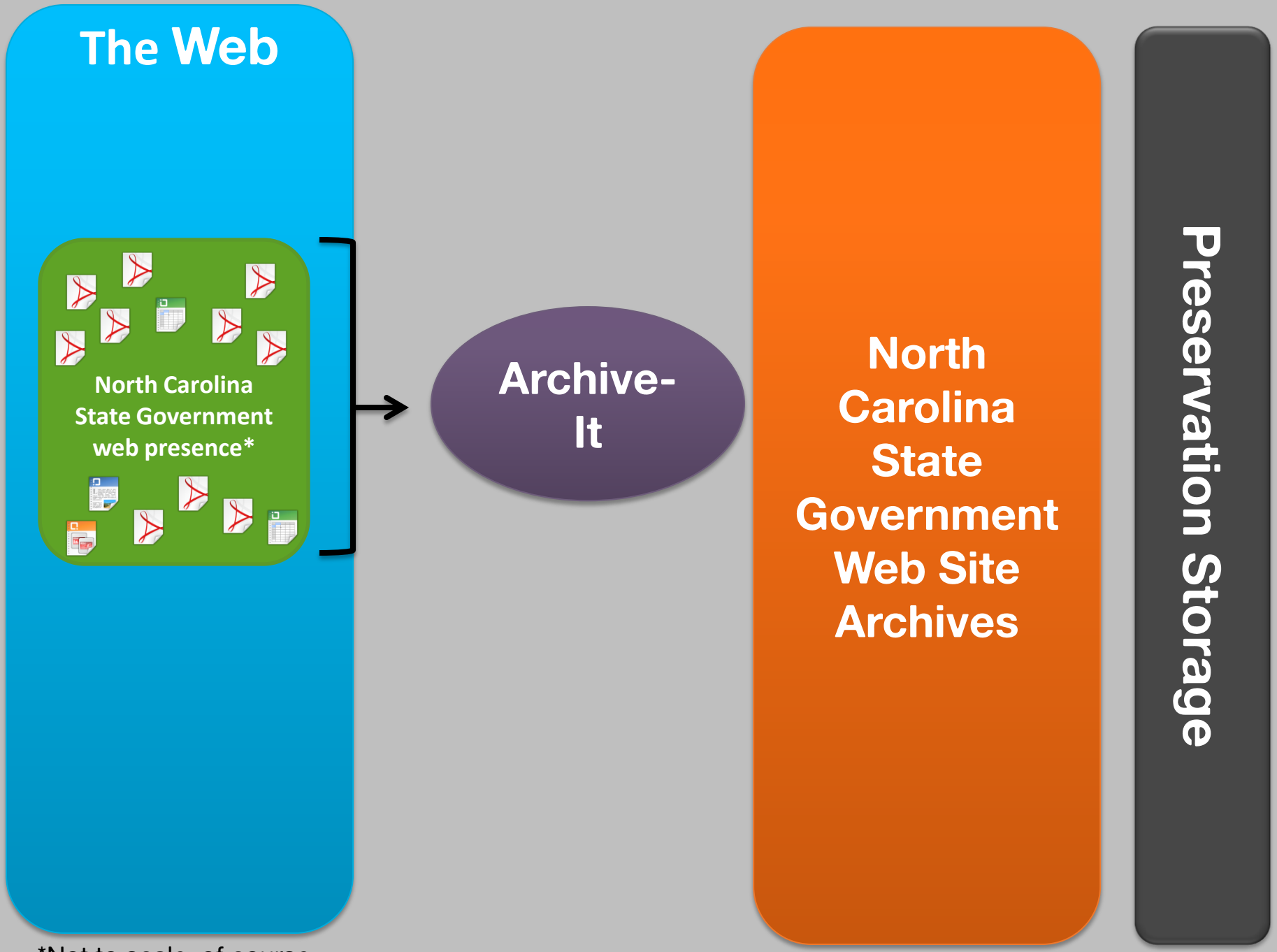
The State Library shall be the official, complete, and permanent depository for all State publications…

- Session laws
- Annual reports
- Technical reports
- Newsletters
- Websites
- and more …

North Carolina General Statute 125

The State Library shall be the official, complete, and permanent depository for all State publications…


Digital NC State Publications Everywhere
Download manually from web
Email
CD or Drive
North Carolina State Government Publications Collection
Preservation Storage



*Not to scale, of course

Drawbacks

Manual collection

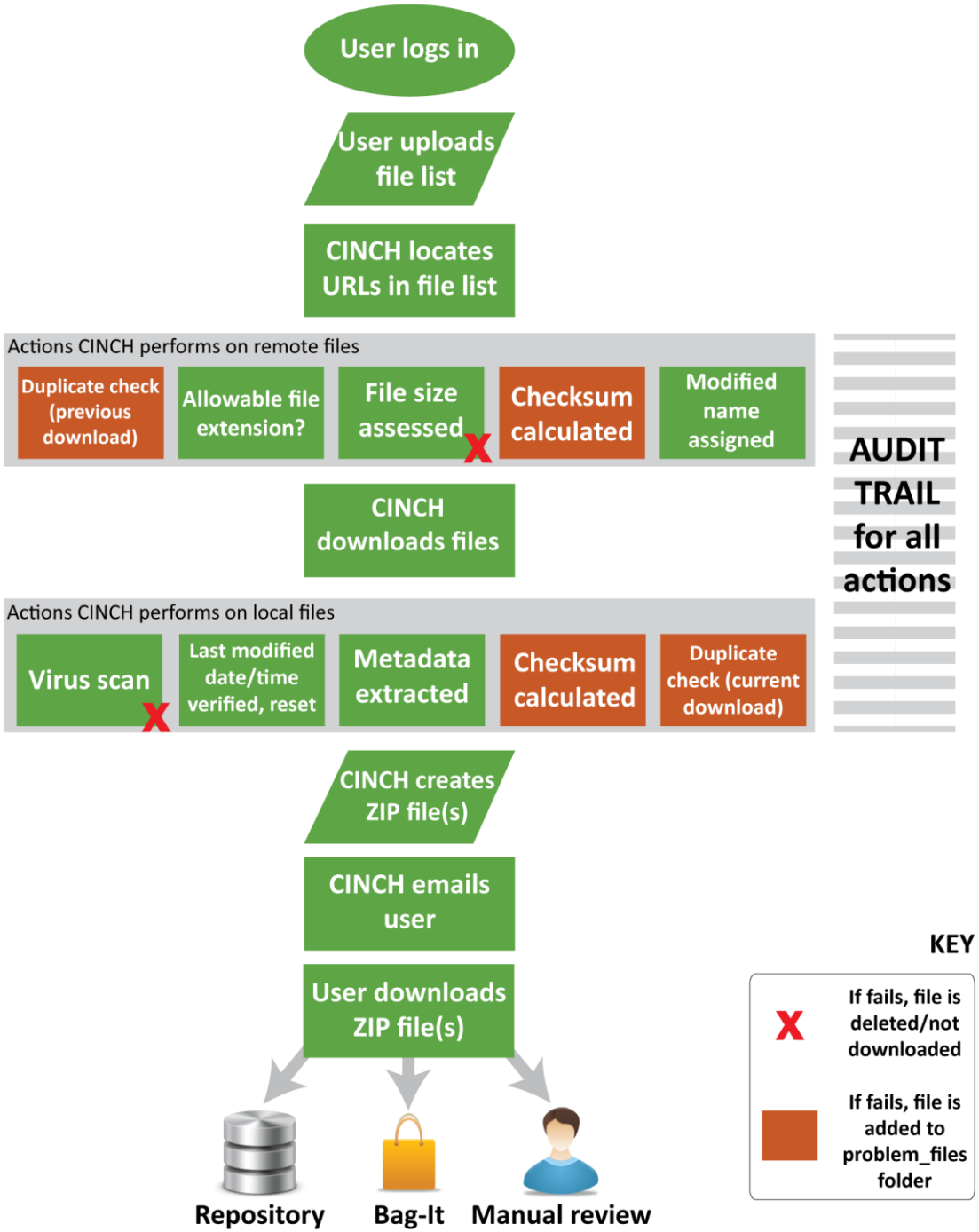
- We're not getting it all
- Our staff could be doing value-add tasks instead
- The ingested object may not be "authentic"
- We have to ~~badger~~ encourage contributors

Website archiving

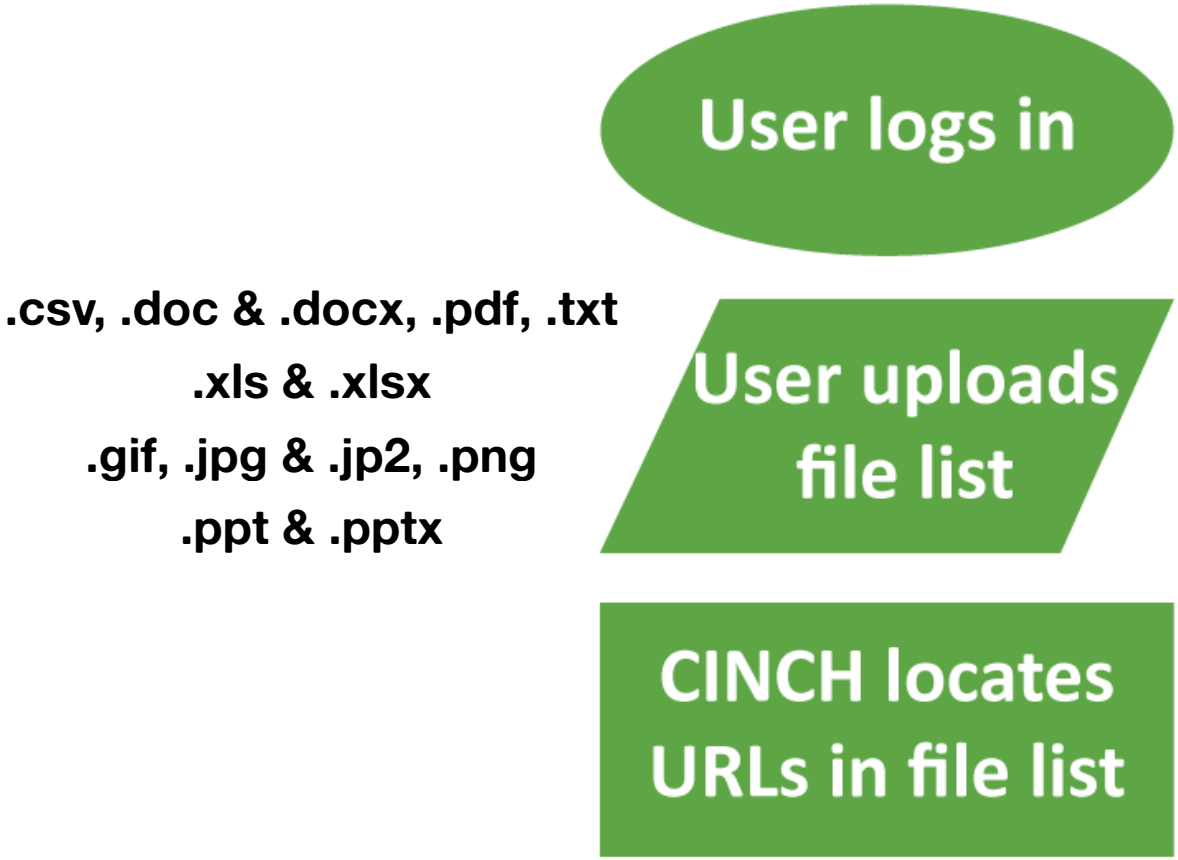
- A web archive is hard for users to understand
- We can't provide the continuity from digitized to digital
- We have tons o' data

How can we extract, use, & preserve the publications found throughout our web site archives, in an automated and preservation-responsible way?

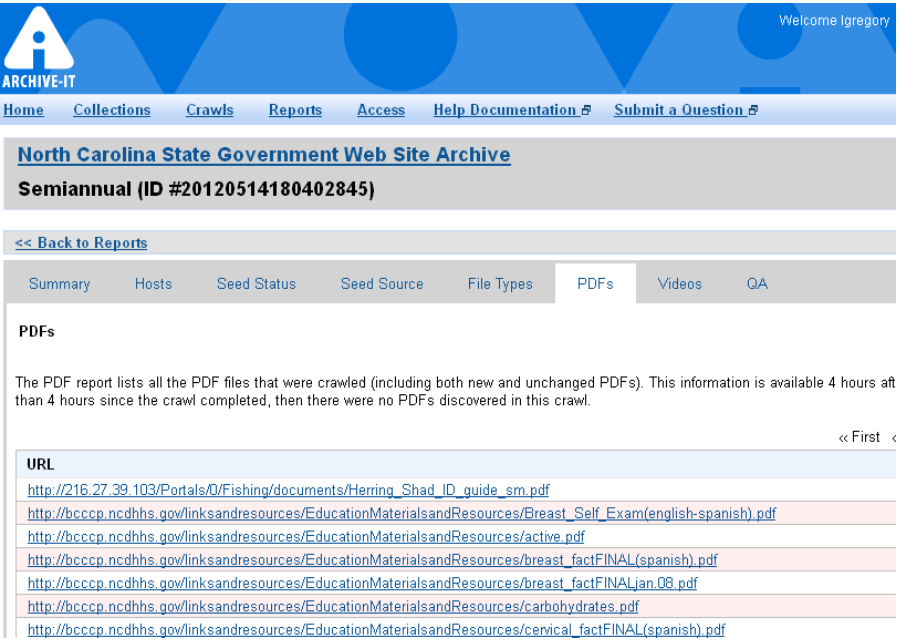
Preservation Storage


User logs in
User uploads file list
CINCH locates URLs in file list
Actions CINCH performs on remote files
Duplicate check (previous download)
Allowable file extension?
File size assessed
Checksum calculated
Modified name assigned
CINCH downloads files
Actions CINCH performs on local files
Virus scan
Last modified date/time verified, reset
Metadata extracted
Checksum calculated
Duplicate check (current download)
AUDIT TRAIL for all actions
CINCH creates ZIP file(s)
CINCH emails user
User downloads ZIP file(s)
Repository
Bag-It
Manual review
KEY
If fails, file is deleted/not downloaded
If fails, file is added to problem_files folder

First steps


User logs in
.csv, .doc & .docx, .pdf, .txt
.xls & .xlsx
.gif, .jpg & .jp2, .png
.ppt & .pptx
User uploads file list
CINCH locates URLs in file list

User uploads file list



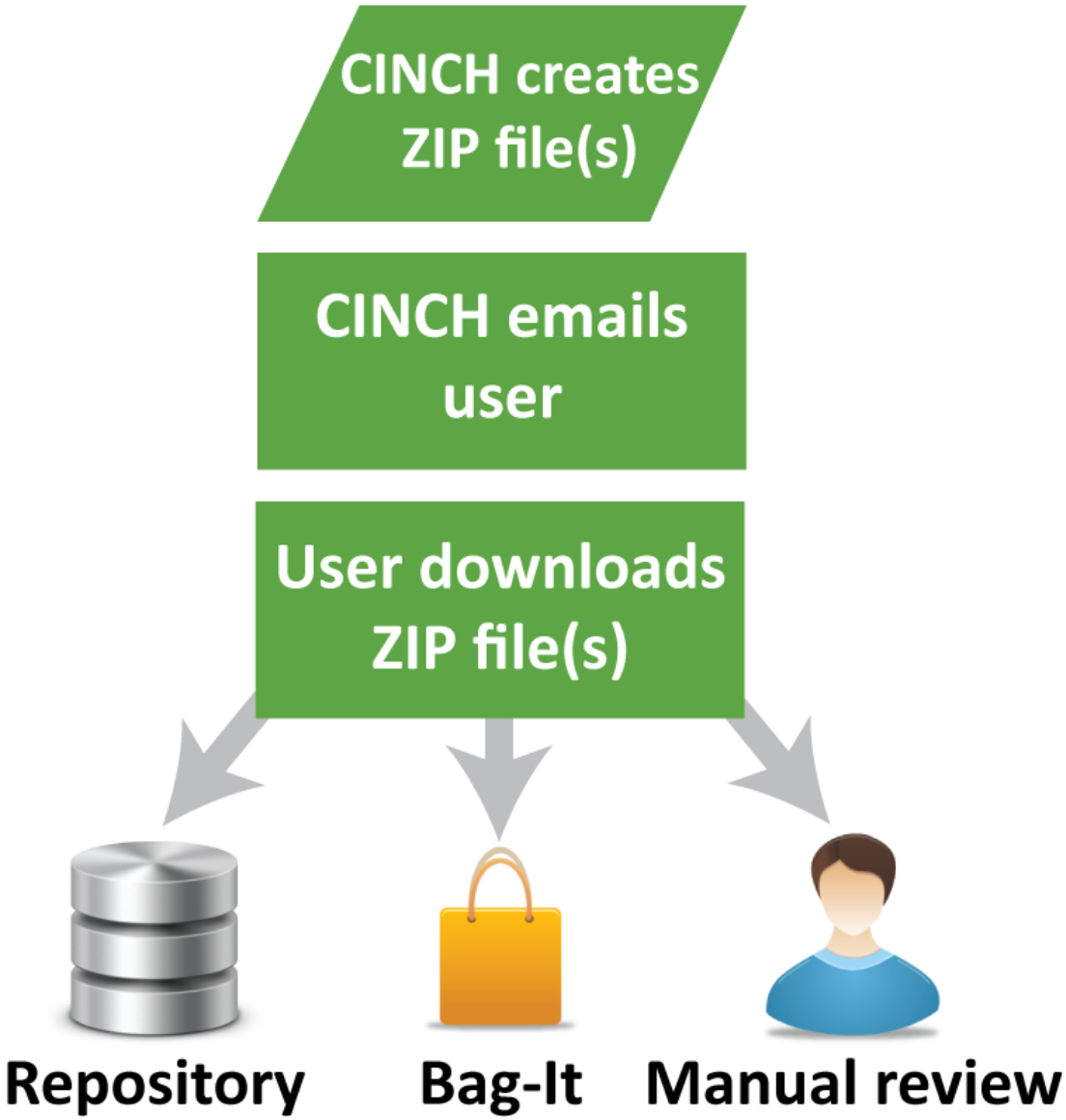
Report from Internet Archive's Archive-It service

Sitemap generator
Ex: www.xml-sitemaps.com or A1 Sitemap Generator

Actions CINCH performs on remote files

Actions CINCH performs on local files

Final steps


CINCH creates ZIP file(s)
CINCH emails user
User downloads ZIP file(s)
Repository
Bag-It
Manual review

CINCH: Capture, INgest, & CHecksum Tool

Home | About | FAQ | Contact | Upload | Download Your Files | Change Password | Logout (test)

Home » Download Your Files

Your Available Downloads

Note: the contents of each file should be unique

Displaying 1-1 of 1 result(s).

Download zip file (created on 2012-06-19 18:30:01):
Download this file: test_downloads_2012_06_19.zip

NORTH CAROLINA DEPARTMENT OF CULTURAL RESOURCES
WWW.NCCULTURE.COM

STATE LIBRARY of NORTH CAROLINA

NC live .org

Problem files

Metadata, audit trail

Files

Name	Size	Type	Date Modified
problem_files		File Folder	6/19/2012 6:30 PM
event_list_2012_06_19.csv	31 KB	Microsoft Office Exc...	6/19/2012 6:30 PM
file_manifest_2012-06-19.csv	2 KB	Microsoft Office Exc...	6/19/2012 6:30 PM
PDF_Metadata_2012_06_19.csv	6 KB	Microsoft Office Exc...	6/19/2012 5:30 PM
statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	18 KB	Adobe Acrobat Doc...	5/27/2011 8:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classb2.PDF	21 KB	Adobe Acrobat Doc...	3/27/2007 2:09 PM
statelibrary.ncdcr.gov_govdocs_classification_classb.PDF	14 KB	Adobe Acrobat Doc...	4/26/2012 11:25 AM
statelibrary.ncdcr.gov_govdocs_classification_classc.PDF	37 KB	Adobe Acrobat Doc...	4/7/2008 10:45 AM
statelibrary.ncdcr.gov_govdocs_classification_classd2.PDF	60 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classd.PDF	97 KB	Adobe Acrobat Doc...	3/18/2004 9:44 AM
statelibrary.ncdcr.gov_govdocs_classification_classf.pdf	16 KB	Adobe Acrobat Doc...	5/27/2011 8:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classg2.PDF	22 KB	Adobe Acrobat Doc...	4/4/2006 11:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classg3.pdf	107 KB	Adobe Acrobat Doc...	2/7/2012 1:40 PM
statelibrary.ncdcr.gov_govdocs_classification_classg4.PDF	110 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classg.PDF	109 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classh.PDF	55 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi2.PDF	80 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi3.PDF	83 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi4.PDF	11 KB	Adobe Acrobat Doc...	7/6/2006 9:17 AM
statelibrary.ncdcr.gov_govdocs_classification_classi.PDF	26 KB	Adobe Acrobat Doc...	4/26/2012 11:25 AM
statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	159 KB	Adobe Acrobat Doc...	3/18/2004 9:29 AM

Name	Size	Type	Date Modified
problem_files		File Folder	6/19/2012 6:30 PM
event_list_2012_06_19.csv	31 KB	Microsoft Office Exc...	6/19/2012 6:30 PM
file_manifest_2012-06-19.csv	2 KB	Microsoft Office Exc...	6/19/2012 6:30 PM
PDF_Metadata_2012_06_19.csv	6 KB	Microsoft Office Exc...	6/19/2012 5:30 PM
statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	18 KB	Adobe Acrobat Doc...	5/27/2011 8:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classb2.PDF	21 KB	Adobe Acrobat Doc...	3/27/2007 2:09 PM
statelibrary.ncdcr.gov_govdocs_classification_classb.PDF	14 KB	Adobe Acrobat Doc...	4/26/2012 11:25 AM
statelibrary.ncdcr.gov_govdocs_classification_classc.PDF	37 KB	Adobe Acrobat Doc...	4/7/2008 10:45 AM
statelibrary.ncdcr.gov_govdocs_classification_classd2.PDF	60 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classd.PDF	97 KB	Adobe Acrobat Doc...	3/18/2004 9:44 AM
statelibrary.ncdcr.gov_govdocs_classification_classf.pdf	16 KB	Adobe Acrobat Doc...	5/27/2011 8:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classg2.PDF	22 KB	Adobe Acrobat Doc...	4/4/2006 11:12 AM
statelibrary.ncdcr.gov_govdocs_classification_classg3.pdf	107 KB	Adobe Acrobat Doc...	2/7/2012 1:40 PM
statelibrary.ncdcr.gov_govdocs_classification_classg4.PDF	110 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classg.PDF	109 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classh.PDF	55 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi2.PDF	80 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi3.PDF	83 KB	Adobe Acrobat Doc...	3/18/2004 9:30 AM
statelibrary.ncdcr.gov_govdocs_classification_classi4.PDF	11 KB	Adobe Acrobat Doc...	7/6/2006 9:17 AM
statelibrary.ncdcr.gov_govdocs_classification_classi.PDF	26 KB	Adobe Acrobat Doc...	4/26/2012 11:25 AM
statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	159 KB	Adobe Acrobat Doc...	3/18/2004 9:29 AM

URL

original file name

Original metadata retained in file header

Advanced

- PDF Properties (pdf, http://ns.adobe.com/pdf/1.3/)
  - pdf:Producer: Acrobat Distiller 5.0 (Windows)
  - pdf:Author: Sally Ensor
  - pdf:CreationDate: 2004-03-17T17:27:48-05:00
  - pdf:Creator: Acrobat PDFMaker 5.0 for Word
  - pdf:ModDate: 2004-03-17T17:28:01-05:00
  - pdf:Title: CLASSIFICATION SCHEME FOR
- XMP Core Properties (xmp, http://ns.adobe.com/xap/1.0/)
  - xmp:CreateDate: 2004-03-17T17:27:48-05:00
  - xmp:ModifyDate: 2004-03-17T17:28:01-05:00
  - xmp:MetadataDate: 2004-03-17T22:27:48Z
  - xmp:CreatorTool: Acrobat PDFMaker 5.0 for Word
  - xmp:Author: Sally Ensor
  - xmp:Authors (seq container)
  - xmp:Title (alt container)
- Dublin Core Properties (dc, http://purl.org/dc/elements/1.1/)
  - dc:creator (seq container)
  - dc:title (alt container)
- http://ns.adobe.com/illustrator/1.0/
  - illustrator:StartupProfile: Print
- Adobe Photoshop Properties (photoshop, http://ns.adobe.com/photoshop/1.0/
  - photoshop:Author: Sally Ensor
  - photoshop:Title: CLASSIFICATION SCHEME FOR
- http://ns.adobe.com/png/1.0/
  - png:Author: Sally Ensor
  - png:CreationTime: 2004-03-17T17:27:48-05:00
  - png:ModificationTime: 2004-03-17T17:28:01-05:00

Event List (Audit Trail)

Url **Filename** **Event Name** **Event Time**

	A	B	C	D
1	Url	Filename	Event Name	Event Time
2	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Download Last Modified time corrected	6/18/2012 22:15
3	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Downloaded	6/18/2012 22:15
4	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Renamed	6/18/2012 22:15
5	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	File integrity check	6/18/2012 23:00
6	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Virus check	6/19/2012 1:00
7	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Local checksum creation	6/19/2012 1:45
8	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	File integrity check	6/19/2012 1:45
9	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Metadata Extraction	6/19/2012 2:30
10	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Full text check	6/19/2012 2:30
11	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	File integrity check	6/19/2012 15:42
12	http://statelibrary.ncdcr.gov/govdocs/classification/Classa.pdf	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Zipped for download	6/19/2012 18:30
13	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Renamed	6/18/2012 22:15
14	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Download Last Modified time corrected	6/18/2012 22:15
15	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Downloaded	6/18/2012 22:15
16	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	File integrity check	6/18/2012 23:00
17	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Virus check	6/19/2012 1:00
18	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Local checksum creation	6/19/2012 1:45
19	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	File integrity check	6/19/2012 1:45
20	http://statelibrary.ncdcr.gov/govdocs/classification/classIntro.pdf	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	Metadata Extraction	6/19/2012 2:30

PDF_Metadata

Author	Creation Date	Last Modified Date	Creator	Producer	Resource name	Title	Pages
A	B	C	D	E	F	G	H
author	creation_date	last_modified	creator	producer	resource_name	title	pages
ENeville	2011-05-23T16:32:14Z	2011-05-23T16:32:14Z	PScript5.dll Version 5.2.2	Acrobat Distiller 9.0.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_Classa.pdf	Microsoft Word - Classa.doc	8
ENeville	2012-04-25T17:48:54Z	2012-04-25T17:48:54Z	PScript5.dll Version 5.2.2	Acrobat Distiller 9.2.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classb.PDF	Microsoft Word - classb.doc	4
eneville	2007-03-27T19:07:24Z	2007-03-27T19:07:24Z	PScript5.dll Version 5.2.2	Acrobat Distiller 7.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classb2.PDF	Microsoft Word - classb2.doc	10
ENeville	2008-04-07T15:13:14Z	2008-04-07T15:13:14Z	PScript5.dll Version 5.2	Acrobat Distiller 7.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classc.PDF	Microsoft Word - classc.doc	18
Sally Ensor	2004-03-18T14:53:13Z	2004-03-18T14:53:17Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classd.PDF	D	4
Sally Ensor	2004-03-17T23:01:14Z	2004-03-17T23:01:19Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classd2.PDF	D	2
ENeville	2011-05-23T16:36:09Z	2011-05-23T16:36:09Z	PScript5.dll Version 5.2.2	Acrobat Distiller 9.0.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classf.pdf	Microsoft Word - classf.doc	8
Sally Ensor	2004-03-17T23:08:32Z	2004-03-17T23:08:37Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classg.PDF	G	4
kmartin	2006-04-04T15:07:04Z	2006-04-04T15:07:04Z	PScript5.dll Version 5.2	Acrobat Distiller 6.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classg2.PDF	Microsoft Word - classg2.doc	11
ENeville	2012-02-03T16:44:08Z	2012-02-03T16:44:08Z	PScript5.dll Version 5.2.2	Acrobat Distiller 9.2.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classg3.pdf	Microsoft Word - classg3.doc	50
Sally Ensor	2004-03-17T23:44:10Z	2004-03-17T23:44:17Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classg4.PDF	G	8
Sally Ensor	2004-03-17T23:46:30Z	2004-03-17T23:46:33Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classh.PDF	H	2
ENeville	2012-04-25T17:52:36Z	2012-04-25T17:52:36Z	PScript5.dll Version 5.2.2	Acrobat Distiller 9.2.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classi.PDF	Microsoft Word - classi.doc	9
Sally Ensor	2004-03-18T13:18:25Z	2004-03-18T13:18:44Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classi2.PDF	I	2
Sally Ensor	2004-03-18T13:18:53Z	2004-03-18T13:18:56Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classi3.PDF	I	2
kmartin	2006-07-06T13:00:51Z	2006-07-06T13:00:51Z	PScript5.dll Version 5.2	Acrobat Distiller 6.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classi4.PDF	Microsoft Word - classi4.doc	4
Sally Ensor	2004-03-17T22:27:48Z	2004-03-17T22:28:01Z	Acrobat PDFMaker 5.0 for Word	Acrobat Distiller 5.0 (Windows)	statelibrary.ncdcr.gov_govdocs_classification_classIntro.pdf	CLASSIFICATION SCHEME FOR	10

Subject	Keywords	Licensed To	Possible Title	Possible Keywords	Checksum	Fulltext
I	J	K	L	M	N	O
subject	keywords	licensed_to	possible_doc_title	possible_doc_keywords	Checksum	Fulltext
			A:1	Dept., Division, series, Division,, Agricultural	4c38e61c7d6aaec288bea54c4176e9ff0b8698ba	Yes
			DEPARTMENT OFINSURANCEB1 Dept. of Insurance (c. 54, s. 3)	Division, Insurance, Dept., exists], Fire	e3a4dcfb7fadbe7a91d7743a666bcb9446de8ac9	Yes
			B:5	Dept., Division, publications, before, Development	2e77e6986448147eecfd5bf425ad59e1299399df	Yes
			C:1	Division, publications, Dept., Health, before	6a6ec35349602714d044da3e5bfead6a4f0dd4bf	Yes
			D DEPARTMENT OF CORRECTIOND1 Parole Commission (143B-266) [1973-]	School, State, Dept., Training, Office	576dafbb70b4733aa09eaf323addb949b7e47471	Yes
			D DEPARTMENT OF JUVENILE JUSTICE ANDDELINQUENCY PREVENTION	Juvenile, Dept., Services, Justice, Prevention	a7bb28fc2d25d55a3f4860398f2ebd2ac22949f2	Yes
			F:1	Services, State, Library, Commission, Division	3ad9c82942a08ae391734dfbd5af6c41ffed0d9e	Yes
			G PUBLIC EDUCATIONG1 State Board of Education	Services, Division, Education, School, Program	17b11bae708e33e98257c21d81ea65a761226771	Yes
			G:5	Division, publications, Education, Section, Services	eff9f28b9cd8360bfb2d5f09d5045f3891ebf4fa	Yes
			THE UNIVERSITY OF NORTHCAROLINA SYSTEM	Dept., Carolina, North, School, publications	874b761deceddb5a0b67e8eff0e81688d43f6037	Yes
			G THE COMMUNITY COLLEGE SYSTEMG145 GASTON COLLEGE (Dallas) [1963-	Technical, COLLEGE, COMMUNITY, Institute,, College,	083073ea80eff612d0397b52503089ddd0f57f44	Yes
			H STATE BOARD OF ELECTIONSH1 State Board of Elections [1901-]	Board, State, Elections, operates, there,	80bbca186c589f4310bd3a725b1b5f009bf2e601	Yes
			DEPARTMENT OFADMINISTRATIONI1 Dept. of Administration	State, Office, Dept., publications, Division	efba3efe4266ffda7def0eb2e02eda7439ffdefe	Yes
			I OFFICE OF THE STATE AUDITORI25 Dept. of State Auditor [1971-]	State, Auditor, audits, audit, Financial	11dff255e0d06d96b5c0a32572206f356fdc6f53	Yes
			I DEPARTMENT OF STATE TREASURERI35 Dept. of State Treasurer [1971-]	State, Retirement, Dept., publications, before	43f4fb08c79fce4bdd55ace704f8be616ddc62ec	Yes
			I:13	Division, Property, 1973,, Valorem, Dept.	595875919443b8f5e3c061c585ccd946b0db6942	Yes
			CLASSIFICATION SCHEME FORNORTH CAROLINA STATE	agency, North, Carolina, Dept., State	e9075a1724b4a049d13c6b9ce12793e1dc2d502f	Yes

Where can I get it?

slnc-dimp.github.com/Cinch/

North Carolina institutions will be able to use a hosted version in August, 2012.

I want more information!

cinch.nclive.org

Feedback welcome!

Lisa Gregory
lisa.gregory@ncdcr.gov

Dean Farrell
dean.farrell@ncdcr.gov

Funding for the CINCH: Capture, INgest, & CHecksum tool is made possible through an IMLS Sparks! Ignition grant.