Persistent Identifiers: Using Archival Resource Keys to keep it all together

Texas Conference on Digital Libraries

May 25, 2022

Overview

- Persistent Identifiers Introduction
  - Laura Waugh, Texas State University
- The ARK Alliance
  - Mark Phillips, University of North Texas
- ARKs at UH
  - Sean Watkins & Bethany Scott, University of Houston
- Minting ARKs at UNT
  - Mark Phillips, University of North Texas

Persistent Identifiers Introduction

Laura Waugh
Texas State University Libraries

Persistent Identifiers (PIDs)

Long-lasting reference

People (researchers)

Places (their organizations)

Things (research outputs)

New name for a long-standing concept

- Publishers used identifiers (ISBNs, ISSNs, etc.) for decades
- Digital publication created need for machine-readable PIDs

Persistent Identifiers (PIDs)

- Discoverable
- Accessible
- Useable
- Intelligible
- Interoperable
- Assessable

Why are PIDs important?

Because reliable web links are lacking

Whoops!

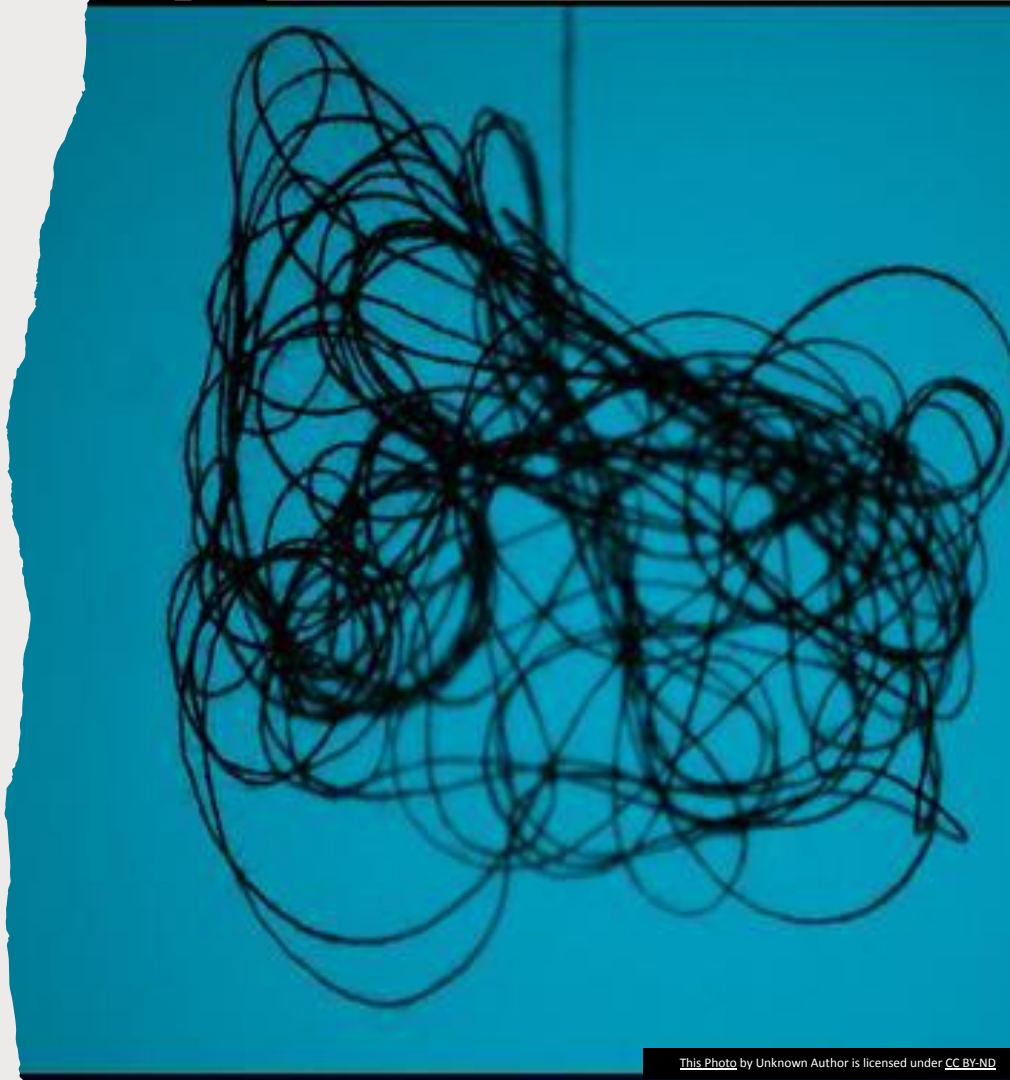
404 Page Not Found

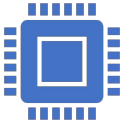
Looks like this page went on vacation.

Try our homepage or blog instead.

History: The Tangled Web

- Internet is launched
- URLs begin breaking
- URL-forwarding
- Internet indirection infrastructure
- IETF (Internet Engineering Task Force) tries URNs
- Fee-based DOIs introduced by publishers
- Handles are introduced as sole vendor/gatekeeper
- https://arks.org/about/the-ark-origin-story/





Commonalities of web-based PIDs

Examples: ARKs, DOIs, Handles, URNs

- Have been around more than 20 years
- Similar goals to address Internet indirection infrastructure
- Start with a string to identify the name assigning authority
- Require active updating of URL redirects

Organizational Commitment

PIDs are only as persistent as the organizations that provide and support them

- PIDs demonstrate a commitment to stewardship
- Rely on commitment and upkeep

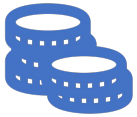
Do PIDs solve broken links?

No. There’s not a clean solution to the problem

PIDs are an important step toward addressing this

Major causes of broken links, and some features	PURL	Handle	URN	DOI	ARK
Prevents fire, war, flood, attack, bankruptcy, …	No	No	No	No	No
Prevents human error	No	No	No	No	No
Guarantees your links, or fixes them for you	No	No	No	No	No
Decentralized admin plus interface syntax	No	No	No	No	**Yes**
Flexible metadata and persistent statements	No	No	No	No	**Yes**
Identifiers extensible during resolution	**Yes**	No	**Yes**	No	**Yes**
Free, non-paywalled, in unlimited numbers	**Yes**	No	**Yes**	No	**Yes**

Capability defined for five types of persistent identifier (PID).



Considerations

- Broadly, what units are we trying to identify?
- When do we need to assign/mint a PID?
- Are PIDs minted before or after ingest?
- Technical implementation, system, and strategy required
- Cost versus Loss

More Info

- The ARK Origin Story
  - https://arks.org/about/the-ark-origin-story/
- Ten persistent myths about persistent identifiers
  - https://escholarship.org/uc/item/73m910w8
- Why Publishers Should Care About PIDs
  - https://scholarlykitchen.sspnet.org/2021/06/21/why-publishers-should-care-about-persistent-identifiers/
- PIDapalooza
  - https://www.pidapalooza.org/ | *@pidapalooza*

The ARK Alliance:
21 years
950 institutions
8.2 billion persistent identifiers

Mark Phillips, *University of North Texas Libraries*

May 2022

Digital preservation means

Long term *protection* for digital resources

- from human error, natural disaster, legal challenge, deliberate attack, social upheaval, bankruptcy, etc.

Long term *access* to those resources from unbroken links

- with *persistent identifiers (PIDs)*, also known as *permalinks*

Why persistent identifiers?

Because of “link rot” (broken references, 404 Not Found)

- Reliable, unbroken web links (URLs) are rare
- The average URL lifetime is only 100 days

But why not just search when you need a link?

- Because scholars and researchers take years to find their object references

Common types of persistent identifiers

- PURL, Handle, URN, DOI, ARK

What is an ARK (Archival Resource Key)?

A labelled URL with a globally unique identity inside it

makes ARK actionable (the resolver)

core globally unique identity (independent of web and hostname)

ARK anatomy

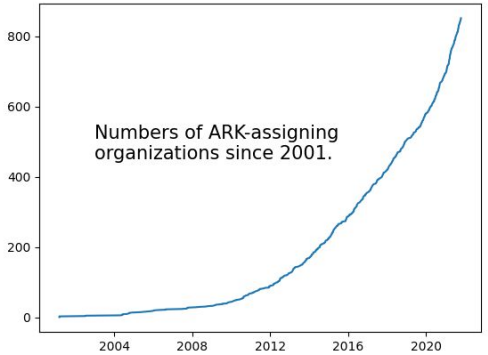
```
https://example.org/ark:/12345/x54xz321/s3/f8.05v.tiff
\_____________________/ \__/ \___/ \______/\____/\_______/
           |            |     |       |       |        |
           |      ARK Label   |       | Sub-parts Variants
           |                  |       |
Name Mapping Authority (NMA)  | Assigned Name
                              |
          Name Assigning Authority Number (NAAN)
```

Why ARKs?

Major causes of broken links, and some features	**PURL**	**Handle**	**URN**	**DOI**	**ARK**
Prevents fire, war, flood, attack, bankruptcy, ...	No	No	No	No	No
Prevents human error	No	No	No	No	No
Guarantees your links, or fixes them for you	No	No	No	No	No
Decentralized admin plus inferenceable syntax	No	No	No	No	**Yes**
Flexible metadata and persistence statements	No	No	No	No	**Yes**
Identifiers extensible during resolution	**Yes**	No	**Yes**	No	**Yes**
Free, non-paywalled, in unlimited numbers	**Yes**	No	**Yes**	No	**Yes**

Who is using ARKs?

- Libraries, data centers, archives, museums, publishers, government agencies, and vendors
- Example institutions:

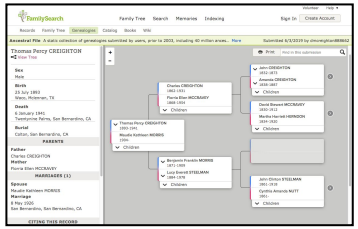


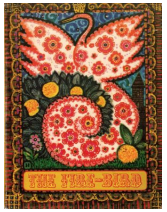
Internet Archive
Caltech Archives
Hawaii State Archives
French National Archives
Rockefeller Archive Center
Library and Archives Canada
Archives de la Ville de Genève
Silent Film Sound & Music Archive

University of California Berkeley
Smithsonian National Museum
National Library of France
University of Chicago
Musée du Louvre
Family Search
British Library
Google

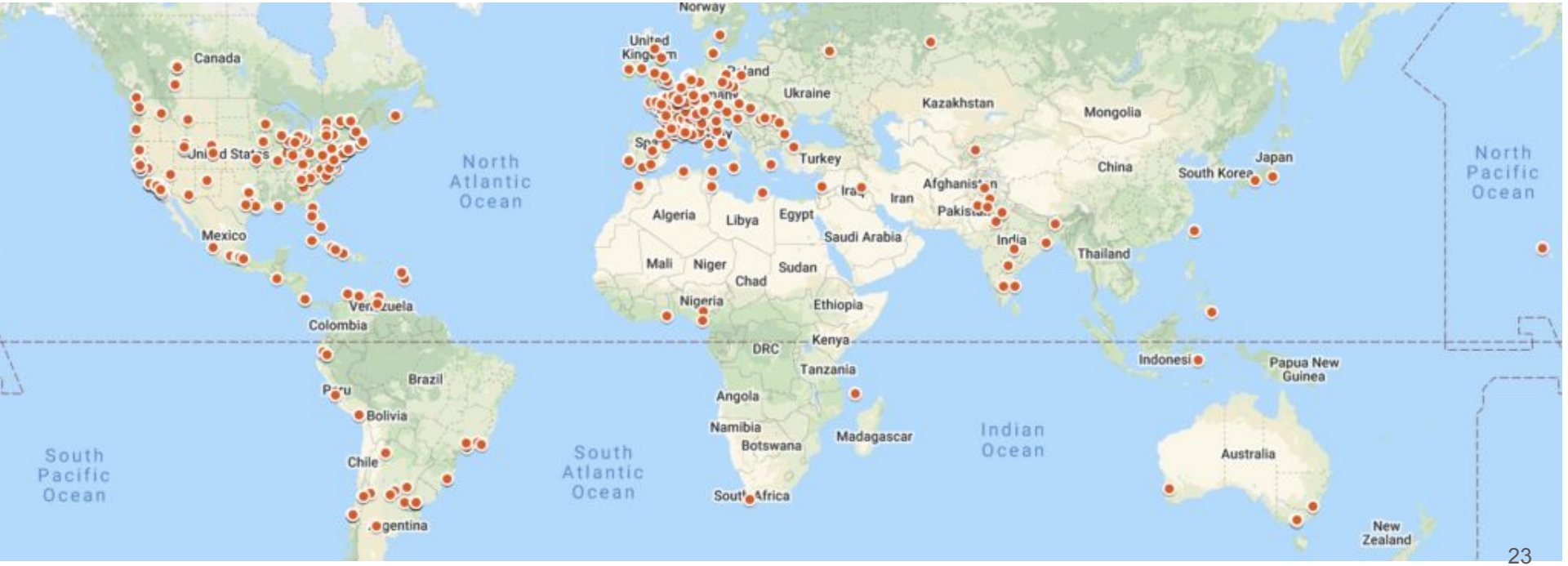
What are ARKs used for?

- genealogical records (8 billion FamilySearch)
- publisher content (100 million Portico)
- scientific datasets and records (22 million INIST)
- scanned books and texts 30 million Internet Archive)
- bibliographic records (15 million BnF main catalog)
- museum specimens (15 million Smithsonian Institution)
- public health documents (15 million UCSF IDL)
- historical documents (21 million CDL, 5 million BnF Gallica)
- historical authors and scholars (4 million SNAC)
- fine art museum collections (483,000 Louvre)
- vocabulary terms (9,000 Periodo, YAMZ)





ARK Alliance: 950 institutions and 8.2 billion ARKs in 21 years



The ARK Alliance

Home of the ARK Alliance

arks.org

Join one of our working groups: info@arks.org

Get started with ARKs by filling out:

n2t.net/e/naan_request

Stay in touch:

- Twitter: @arks_org
- Email forum (English): groups.google.com/group/arks-forum
- Email forum (French): framalistes.org/sympa/info/arks-forum-fr

ARKs @ UH

Bethany Scott
Sean Watkins

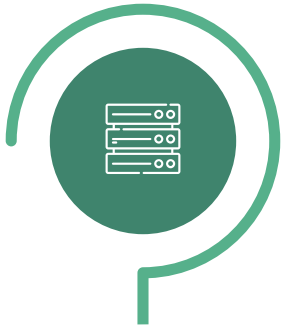
TCDL: May 25, 2022

What are we identifying

Digital Objects

Objects within our digital repositories

Digital Collections
A/V Repository



Preservation SIPs

Packages stored in preservation system

Archivematica

Vocabulary

Controlled vocabulary terms

UHL Vocabularies

When are identifiers assigned

BEFORE INGEST

ARKs are minted prior to ingesting into repositories

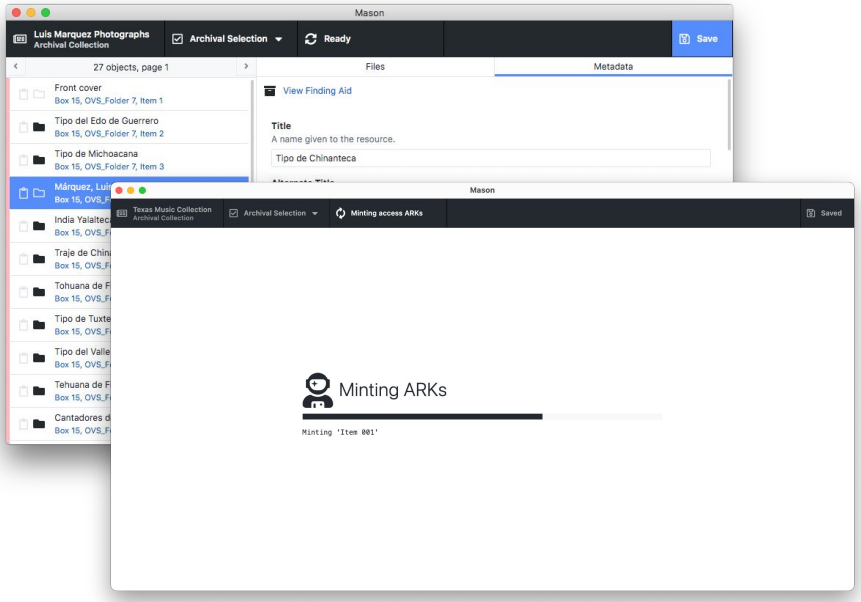
UPDATE ON PUBLISH

ERC where URL is updated during object publishing

How are we assigning identifiers

ARKs are minted and maintained through our custom identifier manager **Greens** using NOIDs

Digital Object and Preservation identifiers are assigned using our digital project application **Mason**



https://github.com/uhlibraries-digital


How did these decisions get made

Minting ARKs @ UNT Libraries

Mark Phillips

May 25, 2022

ARK identifiers @ UNT Libraries

- To uniquely identify a “digital object” in the UNT Libraries’ Digital Collections.
- An object may be a report, issue of a newspaper, photograph, book, dataset, ETD.
- We assign an External-Facing identifier for all items.
- All of the identifier resolution is built into our web applications as URL routing.
- Additionally, we assign Internal-Facing identifiers for all items that go into our preservation repository (to ensure uniqueness) but maintain the original External-Facing identifiers.

Digital Objects in our digital collections

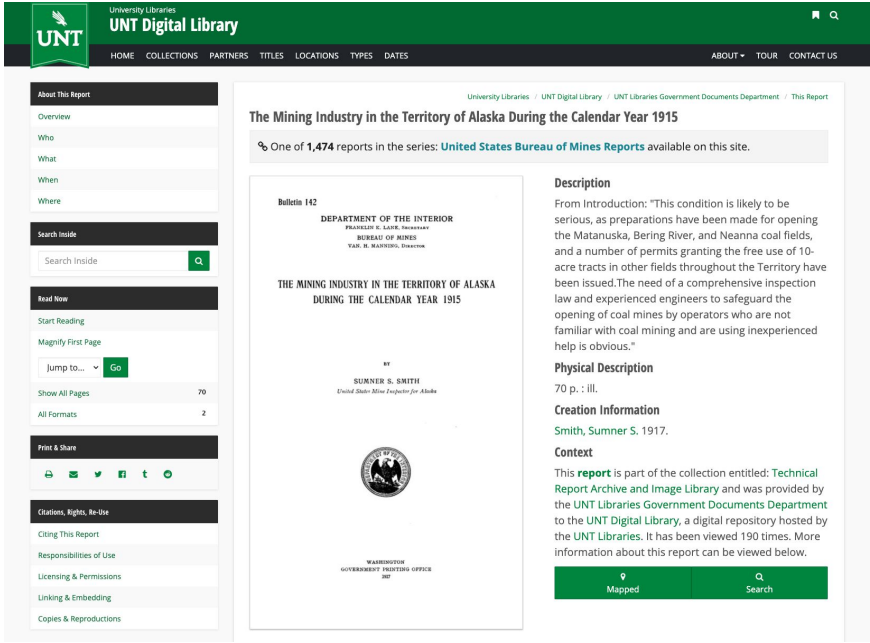
“Wear their Identifiers”

ark:/67531/metadc12345

https://digital.library.unt.edu/ark:/67531/metadc12345/

Standard bookmarking and tools just work.

No need for “use this URL to bookmark”



http://n2t.org/ark:/67531/metadc12345

Object level URLs

- https://digital.library.unt.edu/ark:/67531/metadc12345/
- https://digital.library.unt.edu/ark:/67531/metadc12345/?
- https://digital.library.unt.edu/ark:/67531/metadc12345/??

IIIF Manifest

- https://digital.library.unt.edu/ark:/67531/metadc12345/manifest/

Image URLs

- https://digital.library.unt.edu/ark:/67531/metadc12345/thumbnail/
- https://digital.library.unt.edu/ark:/67531/metadc12345/small/

Metadata URLs

- https://digital.library.unt.edu/ark:/67531/metadc12345/metadata/
- https://digital.library.unt.edu/ark:/67531/metadc12345/metadata.untl.xml
- https://digital.library.unt.edu/ark:/67531/metadc12345/metadata.mets.xml
- https://digital.library.unt.edu/ark:/67531/metadc12345/metadata.dc.xml
- https://digital.library.unt.edu/ark:/67531/metadc12345/metadata.dc.txt

Citation Page

- https://digital.library.unt.edu/ark:/67531/metadc12345/citation/

Mapping Manifestations

Manifestation List for Object

- https://digital.library.unt.edu/ark:/67531/metadc12345/m/

Image-based manifestation

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/embed/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/sequence/

PDF manifestation

- https://digital.library.unt.edu/ark:/67531/metadc12345/m2/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m2/sequence/

FileSet Interactions

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/

OCR Pages

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/ocr/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/ocr.txt

Zoom Interface

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/zoom/

Image URLs

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/thumbnail/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/med_res/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/high_res/

IIIF URLs

- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/canvas/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/image/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/annotations/
- https://digital.library.unt.edu/ark:/67531/metadc12345/m1/1/annotations/ocr/
- https://digital.library.unt.edu/iiif/ark:/67531/metadc12345/m1/1/full/max/0/default.jpg

Minting ARKs

- Number Server for Name generation
  - web.py service
  - Returns an identifier and increments a counter
  - Return integer-based or base36 encoded version
  - Run under mod_wsgi in Apache
- We assign an ARK for end user access (metapth, metadc, metarkv)
- We assign a different ARK for (coda)
- We have a “test” ARK namespace (metatest)
- We have an obsolete namespace (metacrs)

Namespace	Status	Internal/External Facing	Usage
metapth	Operational	External-Facing	Used for The Portal to Texas History
metadc	Operational	External-Facing	Used for the UNT Digital Library and the Gateway to Oklahoma History.
metarkv	Operational	External-Facing	Used for items that are “archive only” and do not rely on Aubrey for access. (Web Archives)
metacrs	Obsolete	External-Facing	Historically used for Congressional Research Service (CRS) Reports; discontinued use in 2007 in favor of metadc.
metatest	Operational	External-Facing	Testing namespace.
coda	Operational	Internal-Facing	Namespace used for the Coda repository system.

UNT Libraries: TRAC Conformance Document
https://digital.library.unt.edu/ark:/67531/metadc1132746/

https://digital.library.unt.edu/ark:/67531/metadc1596980/

Language

- English

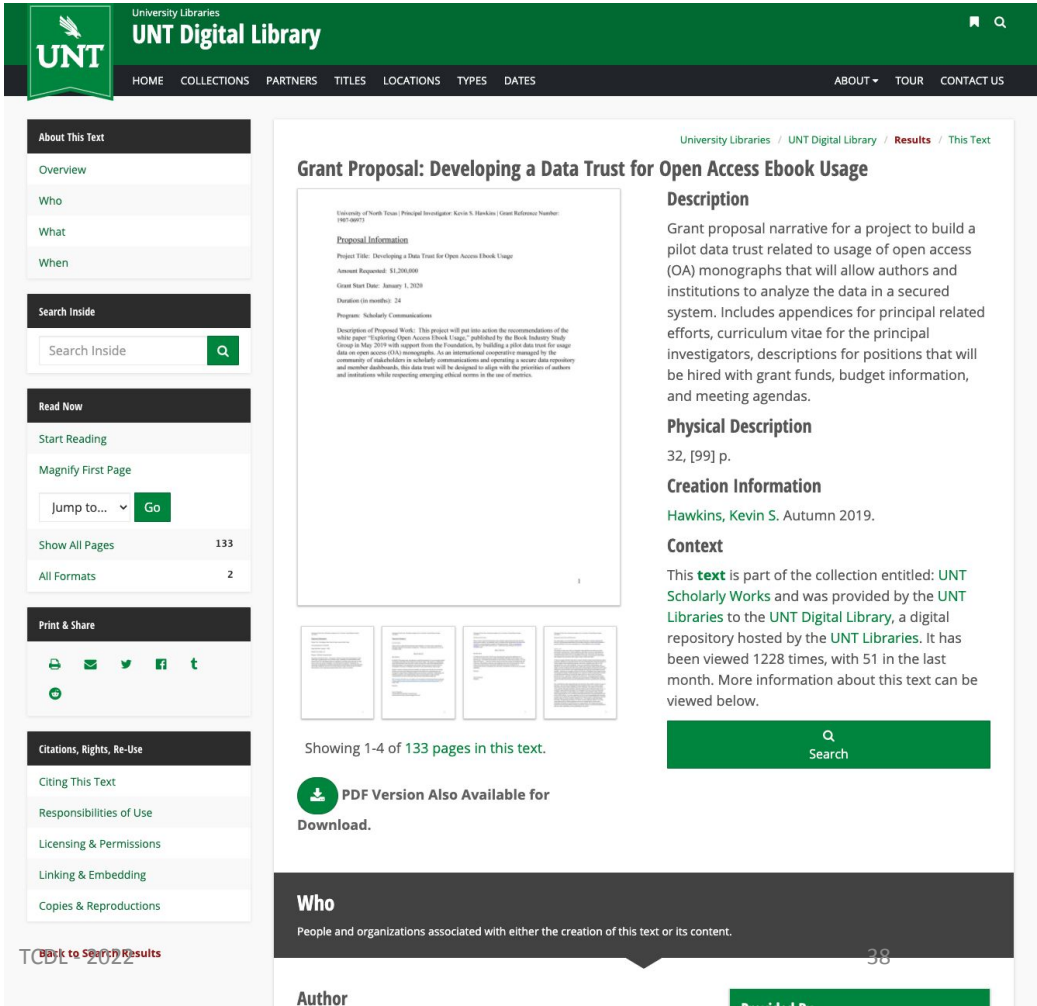
Item Type

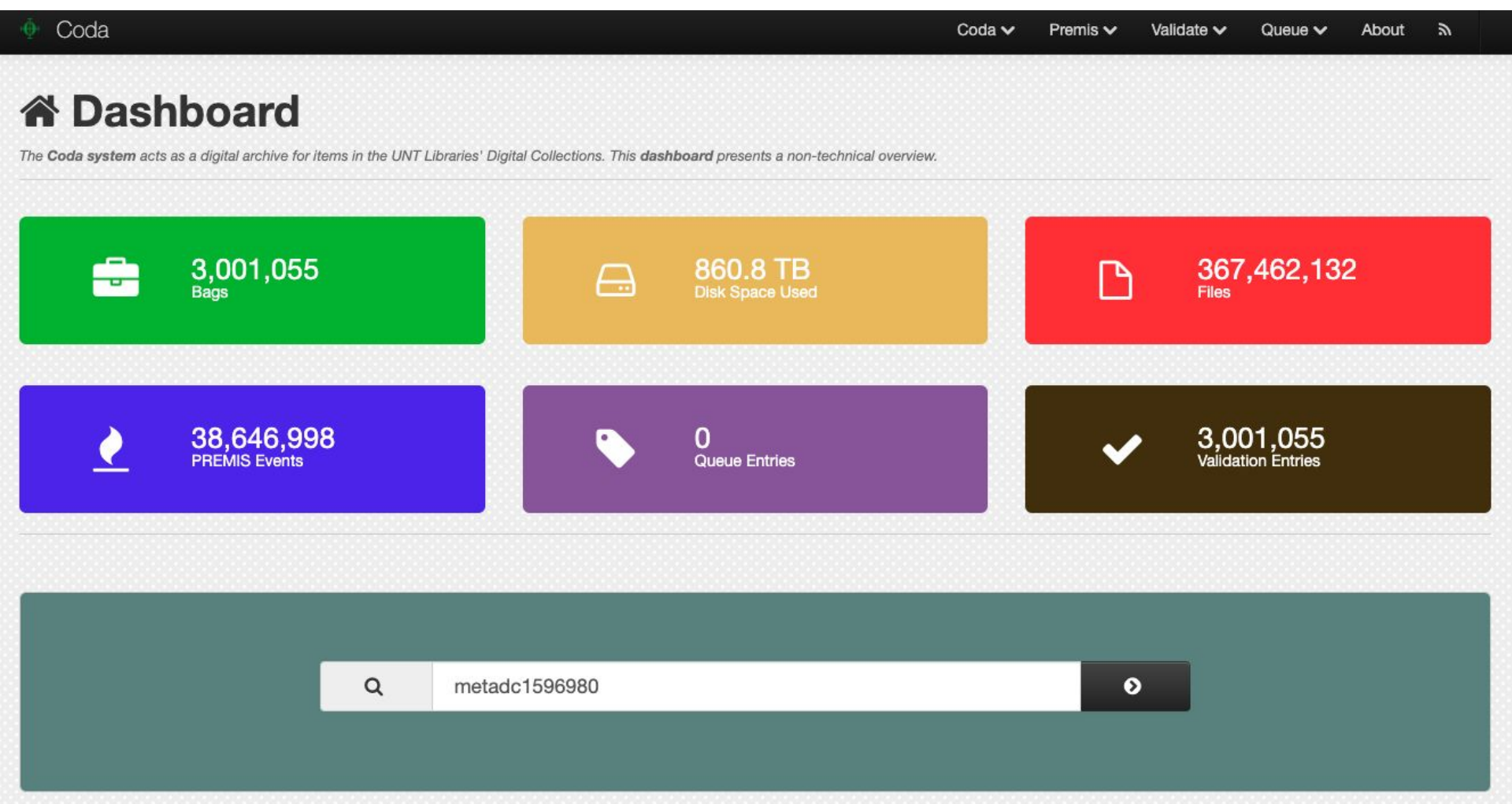
- Text

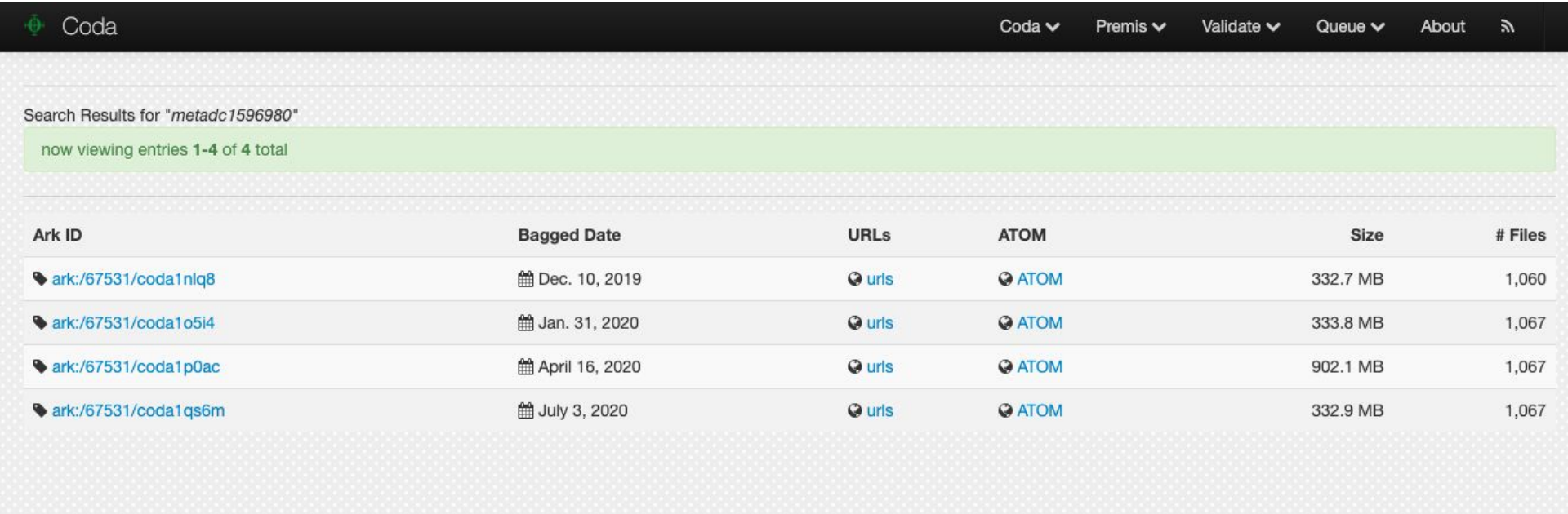
Identifier

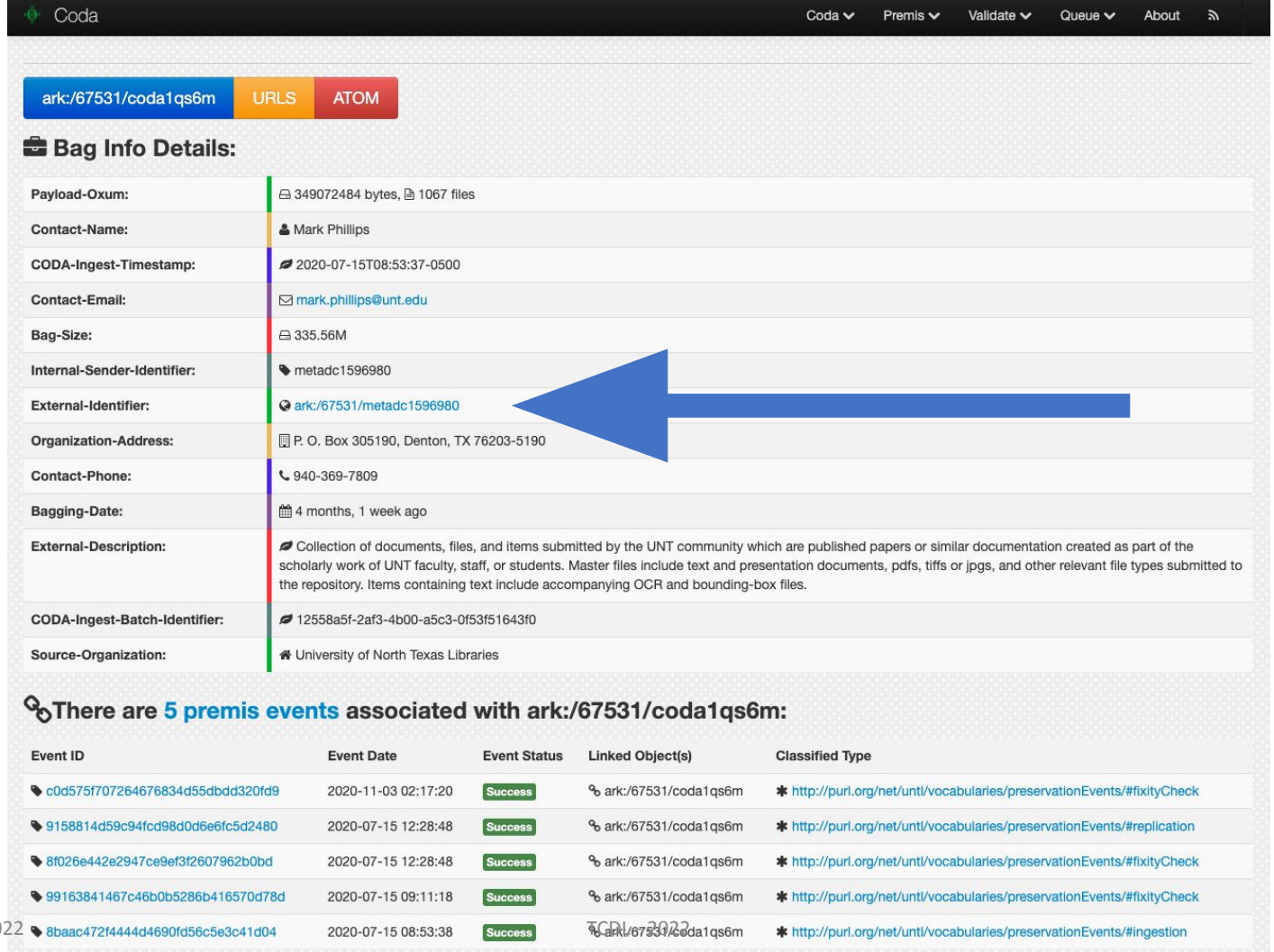
Unique identifying numbers for this text in the Digital Library or other systems.

- **Grant Number**: 1907-06973
- **Archival Resource Key**: ark:/67531/metadc1596980




Coda
Coda
Premis
Validate
Queue
About
Dashboard
The Coda system acts as a digital archive for items in the UNT Libraries' Digital Collections. This dashboard presents a non-technical overview.
3,001,055
Bags
860.8 TB
Disk Space Used
367,462,132
Files
38,646,998
PREMIS Events
0
Queue Entries
3,001,055
Validation Entries
metadc1596980


Coda
Coda
Premis
Validate
Queue
About
Search Results for "metadc1596980"
now viewing entries 1-4 of 4 total
Ark ID
Bagged Date
URLs
ATOM
Size
Files
ark:/67531/coda1nlq8
Dec. 10, 2019
urls
ATOM
332.7 MB
1,060
ark:/67531/coda1o5i4
Jan. 31, 2020
urls
ATOM
333.8 MB
1,067
ark:/67531/coda1p0ac
April 16, 2020
urls
ATOM
902.1 MB
1,067
ark:/67531/coda1qs6m
July 3, 2020
urls
ATOM
332.9 MB
1,067


Coda
Coda
Premis
Validate
Queue
About
ark:/67531/coda1qs6m
URLS
ATOM
Bag Info Details:
Payload-Oxum: 349072484 bytes, 1067 files
Contact-Name: Mark Phillips
CODA-Ingest-Timestamp: 2020-07-15T08:53:37-0500
Contact-Email: mark.phillips@unt.edu
Bag-Size: 335.56M
Internal-Sender-Identifier: metadc1596980
External-Identifier: ark:/67531/metadc1596980
Organization-Address: P. O. Box 305190, Denton, TX 76203-5190
Contact-Phone: 940-369-7809
Bagging-Date: 4 months, 1 week ago
External-Description: Collection of documents, files, and items submitted by the UNT community which are published papers or similar documentation created as part of the scholarly work of UNT faculty, staff, or students. Master files include text and presentation documents, pdfs, tiffs or jpgs, and other relevant file types submitted to the repository. Items containing text include accompanying OCR and bounding-box files.
CODA-Ingest-Batch-Identifier: 12558a5f-2af3-4b00-a5c3-0f53f51643f0
Source-Organization: University of North Texas Libraries
There are 5 premis events associated with ark:/67531/coda1qs6m:
Event ID
Event Date
Event Status
Linked Object(s)
Classified Type
c0d575f707264676834d55dbdd320fd9
2020-11-03 02:17:20
Success
ark:/67531/coda1qs6m
http://purl.org/net/untl/vocabularies/preservationEvents/#fixityCheck
9158814d59c94fcd98d0d6e6fc5d2480
2020-07-15 12:28:48
Success
ark:/67531/coda1qs6m
http://purl.org/net/untl/vocabularies/preservationEvents/#replication
8f026e442e2947ce9ef3f2607962b0bd
2020-07-15 12:28:48
Success
ark:/67531/coda1qs6m
http://purl.org/net/untl/vocabularies/preservationEvents/#fixityCheck
99163841467c46b0b5286b416570d78d
2020-07-15 09:11:18
Success
ark:/67531/coda1qs6m
http://purl.org/net/untl/vocabularies/preservationEvents/#fixityCheck
8baac472f4444d4690fd56c5e3c41d04
2020-07-15 08:53:38
Success
ark:/67531/coda1qs6m
http://purl.org/net/untl/vocabularies/preservationEvents/#ingestion

Challenges, things we might do differently

- Trailing slash or no trailing slash added by web framework
  - If you add the trailing slash, when do you do the ? and ?? Inflections
- We've had to jump through a few hoops to support the ? and ?? Inflections. Currently handled by mod_rewrite in Apache
- While minting on ingest is the ideal approach, we have found need to mint and assign identifiers before ingest for some workflows. This just needs to be planned for.
- We would likely reimplement with a more opaque betanumeric Name instead of our meta(pth|dc|crs|rkv|text) and coda shoulder plus integer approach.

Other writings on the topic.

- Phillips, M. E. (2008). *Using Archival Resource Keys (ARKs) for Persistent Identification.* https://digital.library.unt.edu/ark:/67531/metadc28359/

- Phillips, M. E. (2010). *Some examples of hackable identifiers in the UNT Digital Library.* https://vphill.com/journal/post/2845/

- Phillips, M. E. (2015). *How we assign unique identifiers.* https://vphill.com/journal/post/5548/