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, ISSNls, etc.) for decades
• Digital publication created need for machine-readable PIDs
Persistent Identifiers (PID)

- Discoverable
- Accessible
- Useable
- Intelligible
- Interoperable
- Assessable
Why are PIDs important?

Because reliable web links are lacking
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
<table>
<thead>
<tr>
<th>Major causes of broken links, and some features</th>
<th>PURL</th>
<th>Handle</th>
<th>URN</th>
<th>DOI</th>
<th>ARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevents fire, war, flood, attack, bankruptcy, …</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Prevents human error</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Guarantees your links, or fixes them for you</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Decentralized admin plus interface syntax</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Flexible metadata and persistent statements</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Identifiers extensible during resolution</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Free, non-paywalled, in unlimited numbers</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*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/](https://arks.org/about/the-ark-origin-story/)

• Ten persistent myths about persistent identifiers
  • [https://escholarship.org/uc/item/73m910w8](https://escholarship.org/uc/item/73m910w8)

• Why Publishers Should Care About PIDs

• PIDapalooza
  • [https://www.pidapalooza.org/](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

https://n2t.net/ark:/12345/fk1234

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

\_________________/ \__/ \___/ \_____/ \_____/ \_______/

|          |     |      |      |       |
|          |     |      |      |       |
|          |     |      |      |       |
|          |     |      |      |       |
|          |     |      |      |       |

<table>
<thead>
<tr>
<th>ARK Label</th>
<th>Sub-parts Variants</th>
</tr>
</thead>
</table>

Name Mapping Authority (NMA) | Assigned Name

Name Assigning Authority Number (NAAN)
Why ARKs?

<table>
<thead>
<tr>
<th>Major causes of broken links, and some features</th>
<th>PURL</th>
<th>Handle</th>
<th>URN</th>
<th>DOI</th>
<th>ARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevents fire, war, flood, attack, bankruptcy, ...</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Prevents human error</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Guarantees your links, or fixes them for you</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Decentralized admin plus inferenceable syntax</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Flexible metadata and persistence statements</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Identifiers extensible during resolution</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Free, non-paywalled, in unlimited numbers</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
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/arcs-forum
• Email forum (French): framalistes.org/sympa/info/arcs-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/](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/)
- [https://digital.library.unt.edu/ark:/67531/metadc12345/m1/embed/](https://digital.library.unt.edu/ark:/67531/metadc12345/m1/embed/)
- [https://digital.library.unt.edu/ark:/67531/metadc12345/m1/sequence/](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/)
- [https://digital.library.unt.edu/ark:/67531/metadc12345/m2/sequence/](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)
<table>
<thead>
<tr>
<th>Namespace</th>
<th>Status</th>
<th>Internal/External Facing</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>metapth</td>
<td>Operational</td>
<td>External-Facing</td>
<td>Used for The Portal to Texas History</td>
</tr>
<tr>
<td>metadc</td>
<td>Operational</td>
<td>External-Facing</td>
<td>Used for the UNT Digital Library and the Gateway to Oklahoma History.</td>
</tr>
<tr>
<td>metarkv</td>
<td>Operational</td>
<td>External-Facing</td>
<td>Used for items that are “archive only” and do not rely on Aubrey for access. (Web Archives)</td>
</tr>
<tr>
<td>metacrs</td>
<td>Obsolete</td>
<td>External-Facing</td>
<td>Historically used for Congressional Research Service (CRS) Reports; discontinued use in 2007 in favor of metadc.</td>
</tr>
<tr>
<td>metatest</td>
<td>Operational</td>
<td>External-Facing</td>
<td>Testing namespace.</td>
</tr>
<tr>
<td>coda</td>
<td>Operational</td>
<td>Internal-Facing</td>
<td>Namespace used for the Coda repository system.</td>
</tr>
</tbody>
</table>

UNT Libraries: TRAC Conformance Document
https://digital.library.unt.edu/ark:/67531/metadc1132746/
Grant Proposal: Developing a Data Trust for Open Access E-book Usage

Description
Grant proposal narrative for a project to build a pilot data trust related to usage of open access (OA) monographs that will allow authors and institutions to analyze the data in a secured system. Includes appendices for principal related efforts, curriculum vitae for the principal investigator, descriptions for positions that will be hired with grant funds, budget information, and meeting agendas.

Physical Description

Creation Information
Hawkins, Kevin S. Autumn 2019.

Context
This text is part of the collection entitled: UNT Scholarly Works and was provided by the UNT Libraries to the UNT Digital Library, a digital repository hosted by the UNT Libraries. It has been viewed 1228 times, with 51 in the last month. More information about this text can be viewed below.
The Coda system acts as a digital archive for items in the UNT Libraries' Digital Collections. This dashboard presents a non-technical overview.
<table>
<thead>
<tr>
<th>Ark ID</th>
<th>Bagged Date</th>
<th>URLs</th>
<th>ATOM</th>
<th>Size</th>
<th># Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>ark://67531/coda1nq8</td>
<td>Dec. 10, 2019</td>
<td>📂urls</td>
<td>📂ATOM</td>
<td>332.7 MB</td>
<td>1,060</td>
</tr>
<tr>
<td>ark://67531/coda1o5i4</td>
<td>Jan. 31, 2020</td>
<td>📂urls</td>
<td>📂ATOM</td>
<td>333.8 MB</td>
<td>1,067</td>
</tr>
<tr>
<td>ark://67531/coda1p0ac</td>
<td>April 16, 2020</td>
<td>📂urls</td>
<td>📂ATOM</td>
<td>902.1 MB</td>
<td>1,067</td>
</tr>
<tr>
<td>ark://67531/coda1qs6m</td>
<td>July 3, 2020</td>
<td>📂urls</td>
<td>📂ATOM</td>
<td>332.9 MB</td>
<td>1,067</td>
</tr>
</tbody>
</table>
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.

  https://digital.library.unt.edu/ark:/67531/metadc28359/

  https://vphill.com/journal/post/2845/

  https://vphill.com/journal/post/5548/