Getting running with ARK persistable identifiers

ARK Tutorial 2023

*John Kunze, Donny Winston*
Tutorial overview

Where you are
- Tutorial – Getting running with ARK persistable identifiers
- Six segments with built-in break time
- Logistics first: wifi, bathrooms
- Introductions

Goals
- Know when to use Archival Resource Keys (ARKs)
- Learn ARK Anatomy
- Be able to create and resolve ARKs with confidence
Why care about ARK identifiers?

- Because robust web links are rare – the average URL lifetime is 100 days
- ARKs can be “persistent” identifiers (PIDs), but we prefer “persistable”
- “Ten persistent myths about persistent identifiers”
  
  [https://n2t.net/ark:/13030/c7gb1xh09](https://n2t.net/ark:/13030/c7gb1xh09)

Introduced in 2001: the ARK (Archival Resource Key) identifier scheme
ARK anatomy

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)
N2T.net is a global “name” to “thing” resolver

Why not “ARKresolver.net” like the exclusionary practice of every other PID scheme?

Because ARKs are inclusive and resolvers generalize easily.
ARK organizations

8.2 billion ARKs created by 1100 institutions – libraries, archives, museums, publishers, educators, etc. For example,

- Internet Archive
- Bodleian Libraries
- Berkeley Law Library
- Bibliothèque Mazarine
- New York Public Library
- French National Archives
- National Library of Austria
- Library and Archives Canada
- University of California Berkeley
- Smithsonian National Museum
- National Library of France
- University of Chicago
- Musée du Louvre
- Family Search
- British Library
- Google

https://n2t.net/ark:/53355/c1010066723
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- historical documents (21 million CDL, 5 million BnF Gallica)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 (490,000 Louvre)
- vocabulary terms (9,000 Periodo, YAMZ)
Case studies

NAAN  Organization
13960  Internet Archive
65665  Smithsonian
12148  French National Library (BnF)
99152  YAMZ.net metadata terms
21547, etc  iSamples physical samples
1st Break – 5 minutes
History of “persistable” id schemes

- PURL (Persistent URL) – “URLs are fine if you redirect from purl.org”
- URN (Uniform Resource Name), DOI (Digital Object Identifier) & Handle
  ○ “URLs and domain names are bad, except for ours, and we redirect”
- Tim Berners-Lee – “cool URLs don’t break”
- ARK (Archival Resource Key) – “URLs are fine if managed well, but please tell us which of your URLs are meant for what kind of persistence”
# PID schemes – pessimist view

<table>
<thead>
<tr>
<th>Helps with major causes of broken links?</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 or service provider 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>Best practices guard against copy/paste errors</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Global resolver downtime less than 1 day per year</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Identity independence from lost domain/server name</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

[Source: arks.org]
Web access – direct

1. click
2. URL
3. page
4. render page
Web access – indirect

Example: archive.example.org/photo123 → photos.example.org/vault/123

A redirect is like sending a request to a forwarding address
PID schemes – optimist view

<table>
<thead>
<tr>
<th>Features and costs</th>
<th>PURL</th>
<th>Handle</th>
<th>URN</th>
<th>DOI</th>
<th>ARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decentralized resolution</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Inferenceable syntax (variants, containment)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Flexible metadata by design, including none</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Inflections (...?info) and content negotiation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Nuanced persistence statements by design</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Path extensions during resolution (suffix passthrough)</td>
<td><strong>Yes</strong></td>
<td>No</td>
<td>Yes?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><em>Free, non-paywalled, in unlimited numbers</em></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## PID schemes – ecosystem view

<table>
<thead>
<tr>
<th>Identifiers in an Internet context</th>
<th>PURL</th>
<th>Handle</th>
<th>URN</th>
<th>DOI</th>
<th>ARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appear in Data Citation Index, HathiTrust, Wikipedia, Wikidata, Internet Archive, ORCID profiles</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Major adoption by most academic publishers outside the global South</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Free (subsidized) account and admin interface for one-off use, e.g., purl.org, zenodo.org, archive.org</td>
<td>Yes?</td>
<td>No?</td>
<td>No?</td>
<td>Yes</td>
<td>Yes?</td>
</tr>
<tr>
<td>IETF standard URI, validated by web browsers</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Replicated global resolver architecture</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Exercise: some situations calling for PIDs

Q: You have no PID or repository, but want to preserve 25 tech reports per year. Which approach would you take and which PIDs would work well?

Q: You have 99 semantic web terms to embed in PIDs. Which would you use?
Summary: ARK benefits

ARKs can serve as persistable identifiers with metadata
- found in the Data Citation Index, HathiTrust, Wikipedia, Wikidata, Internet Archive, ORCID profiles, etc.

In contrast to other id schemes, ARKs have
- no fees, no limits, no walled gardens (decentralized)
- very flexible metadata, including none
- can be assigned to anything digital, physical, or conceptual
Smithsonian ARKs: 65665

The Smithsonian Libraries & The Smithsonian Institution
- ARKs for collection metadata & multimedia objects
- Started in 2015
- By 2020 over 15 million ARKs and counting....

“ARKs are a perfect fit for our [Smithsonian] collections”
- Project size
- Cost
- Ease of implementation
- Permanence
Scientific specimens from the National Museum of Natural History
http://n2t.net/ark:/65665/381440f27-3f74-4eb9-ac11-b4d633a7da3d

Cultural artifacts from the National Museum of American History
http://n2t.net/ark:/65665/ng49ca746b2-42dc-704b-e053-15f76fa0b4fa

Sculpture from the Freer Gallery of Art & Arthur M. Sackler Gallery
http://n2t.net/ark:/65665/ye3080ce305-a705-49cc-a70d-99aff8cb65da

Photographs from the National Museum of African American History and Culture
http://n2t.net/ark:/65665/fd5ad97cb86-caaf-4209-8fde-98d70f52f072

Paintings from the Smithsonian American Art Museum
http://n2t.net/ark:/65665/vk7a466371d-0413-451f-bd76-ca0becc46f94
2nd Break – 5 minutes
ARK anatomy revisited

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)
Name Assigning Authority Number (NAAN)

A 5-digit number for an assignment stream within an organization

- Opaque good for longevity – 12148, 13960, 88238, 48225
- To try in browser: n2t.net/…
  - ark:, doi:, isbn:, pdb:
  - ark:12148, ark:13030

NAAN as

- Resolution reference point
- Isolates assignment responsibility (autonomy, uniqueness, re-use)
Opacity pros and cons

Can be generated ("minted") from any source:
- Counter, Noid, UUID, ULID, even content digest
- Anything unique – but best to keep it short
- With Noid (Nice Opaque Identifiers), you get check characters

Opaque ids are a pain for humans
- Difficult to enter correctly (no clues to correct spelling)
- No clues for humans to check for transcription errors
Object life stages

ARK metadata is uniquely flexible – none to any – and supports birth

- Planning phase, moment of birth, first analysis,
- Creating lots crazy metadata, then normalized metadata,
- Pre-release feedback and insights based on limited sharing,
- Corrections, abandonment,
- … plus archiving, public release, revision, enhancement, etc.
<table>
<thead>
<tr>
<th>Scheme</th>
<th>Name assigning authority number (NAAN)</th>
<th>Name</th>
<th>Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>delivery</td>
<td></td>
<td>the world</td>
<td></td>
</tr>
<tr>
<td>service</td>
<td></td>
<td>the naming authority</td>
<td></td>
</tr>
<tr>
<td>Name mapping</td>
<td></td>
<td>the resource</td>
<td></td>
</tr>
<tr>
<td>authority</td>
<td></td>
<td>page</td>
<td>variant</td>
</tr>
</tbody>
</table>

French National Library (BnF) ARKs: 12148

http://gallica.bnf.fr/ark:/12148/bpt6k103039f/f26.thumbnail
BnF ARKs in times of change

Originally: ARKs for
- digitized items
- bibliographic records from the main catalogue

New applications
- for new objects
- for existing objects:
  preservation
  repository,
  linked data
  services

New objects
- finding aids
- illuminations
- museographic descriptions
- born digital documents
- virtual exhibitions

Existing apps, other features
- full text OCR
- full text search
- audio rendering
Changing domain names

Changing technical environment

Changing organization
3rd Break – 15 minutes
ARK anatomy parallel with IIIF

Resolver Service  Compact ARK

/ \ / \  
https://example.org/ark:/12345/x6np1wh8k/c3/s5.v7.xsl 
/ \ \ / \ \ / \ Prefixes  Base Name  Suffixes
ARK + IIIF example

https://gallica.bnf.fr/iiif/ark:/12148/btv1b8449691v/f29/2131,4016,1467,948/full/0/default.jpg
People talk about persistence, but seldom define it

Proof of non-persistence: 404 Not Found

- Hey, the bits changed – therefore non-persistence? Nope.

Preservation ≠ unchanging content

- The more valuable the content, the more subject to human curation
- Changing technology drives changing preservation experience

Exercise: can you think of any content that never changes?
Preservation is not binary

Persistence is not “on” or “off”. It is nuanced.

- Rapidly changing files (earth observation sensor files that grow every 6 seconds, databases that are annotated regularly)
- Similarly, this journal is preserved – why does it keep growing?
- That abstract changed – why?
- Valuable objects are often complex, volatile, human-curated clusters

What does this volatility mean for Merkle DAG trees?
Abstract

In this paper we present a draft vocabulary for making “persistence statements.” These are not arcane notions, but simple tools for pragmatically addressing the concern that anyone feels upon experiencing a broken web link. Scholars increasingly use scientific and cultural assets in digital form, but choosing which among many objects to cite for the long term can be difficult. There are few well-defined terms to describe the various kinds and qualities of persistence that object repositories and identifier resolvers do or don’t provide. Given an object’s identifier, one should be able to query a provider to retrieve human- and machine-readable information to help judge the level of service to expect and help gauge whether the identifier is durable enough, as a sort of long-term bet, to include in a citation. The vocabulary should enable providers to articulate persistence policies and set user expectations.
Setting user expectations, part 1

Terms for content variance

- **frozen** – unchanging bitstream
- **keeping** – unchanging content
- **fixing** – subject to correction
- **rising** – subject to active enhancement
- **molting** – unchanging theme
Setting user expectations, part 2

Terms for object availability

- **finite** – ends at known date or event
- **indefinite** – no special commitment
- **lifetime** – as long as the provider exists
- **subinfinite** – beyond provider’s lifetime
Setting user expectations, part 3

A term for objects that grow in a certain way

- *waxing* – non-disruptive growth

Examples

- live sensor data feeds
- serial publications
Why should we believe you?

Terms specifying the nature of the provider

- *name* – of organization
- *identifier* – unique organizational identifier
- *mission* – is preservation in your mission?
- *succession* policy
Persistence in presence of versions

Terms for content referencing

- **extraversionsed** – “10.2345/67, Version 4”
- **intraversionsed** – “10.2345/67.V4”
- **introversionsed** – “10.2345/6789”
The landing page debate

What if you could get either experience?

- **plunging** – for machine consumption
- **landing** – for human consumption
Naming policy

Forming identifier strings

*NR* – non-reassignment

*OP* – opaque identifiers

*CC* – check character added
4th Break – 5 minutes
Object types

Digital, Physical, Conceptual

Questions: Surrogacy, Naming, Resolution

Commonly, “logical” objects are useful things to commit to

- Opaque ARKs to the object name level, with highest commitment
- Less opaque name suffix extensions, with lower commitments
YAMZ.net ARKs: 99152/h1

- ARKs for metadata terms
- Note: shared NAAN with reserved “shoulder”: h1
- Vocabulary creation, sharing, and standards
  - better, faster, cheaper
Glacier

Definition: Precipitation of ice crystals, isolated or as part of a cluster, falling from a cloud.

Created 2022.03.08
Last Modified 2022.04.08
Contributed by GCW
Glossary
Permalink: https://n2t.net/ark:/99152/h5966

Contributions to the YAMZ metadictionary are dedicated to the public domain under the terms of CC0. By using this site, you agree to Terms of Use and Privacy Policy statements similar to wikimediafoundation.org.
Tools

Documentation and Software: arks.org/resources
Minters: Noid, UUID, ULID, …
Resolvers: Noid, OJS Plugin, IA, ARKs Service UTScarborough
N2T, EZID – Suffix Passthrough
Suffix Passthrough in Action

Registered ARK
http://n2t.net/ark:/12345/x98765

Baseline redirection

Registered target URL
http://datazoo.example.com/carbon288
Final Break – 10 minutes
Obtain a NAAN so you can create ARKs

Fill out this form (linked, in case you forget, from the arks.org homepage):

n2t.net/e/naan_request
Wrap up – final questions?

John Kunze, jakkbl@gmail.com
Donny Winston, donny@polyneme.xyz