ARKs in the Open: 3.2 Billion Persistent Identifiers

1. John Kunze, California Digital Library
2. Bess Missell, Smithsonian Libraries
3. Karen Hanson, Portico, ITHAKA
4. Tom Creighton, Family Search International

April 2020
Why care about ARK identifiers?

Because persistent, reliable web links are lacking.

- The average URL lifetime is 44 days
- Maybe ok for the rest of the world, but not for archives and libraries
- URLs in vendor content management tools (a) can break between major system releases and (b) aren’t generally portable to another vendor’s system

Wanted: a flexible, low cost, vendor- and software-independent persistent identifier
ARK (Archival Resource Key)

- ARK: a persistent link for any kind of thing
- 3.2 billion ARKs created by 600 institutions – libraries, archives, museums, publishers, educators, etc. A sample:

  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
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)
What are ARKs used for?

- genealogical records (3 billion FamilySearch)
- publisher content (100 million Portico)
- scientific datasets and records (22 million INIST)
- scanned books and texts (23 million Internet Archive)
- bibliographic records (15 million BnF main catalog)
- museum specimens (15 million Smithsonian Institution)
- public health documents (14 million UCSF IDL)
- historical documents (21 million CDL, 5 million BnF Gallica)
- historical authors and scholars (4 million SNAC)
- vocabulary terms (9,000 Periodo, YAMZ)
Why ARKs and not DOIs (or Handles or PURLs or URNs)?

- **G****gle** for "ten persistent myths about persistent identifiers"
- **Flexible resolution**: centralized (n2t.net) or via your own server
- **University of California’s history of open**
  - 1968 Free speech movement
  - 1982 Open source Berkeley UNIX – FreeBSD – Mac OS X
  - 2013 Groundbreaking open access research policy
  - 2019 Termination of Elsevier journal subscription
  - 2001–2020 Non-paywalled, decentralized persistent identifiers – ARKs
ARKs and DOIs

DOIs (Digital Object Identifiers) – publishing industry solution

- requires membership, per-identifier fees, rigid metadata requirements

ARKs – cultural heritage solution

- no fees or membership, and highly flexible creation and metadata policies
- like DOIs, ARKs are also stable, linked to metadata, and found in the Data Citation Index (linked to the Web of Science), HathiTrust, Wikipedia articles, Wikidata records, Internet Archive collections, ORCID researcher profiles, etc.
The Covenant of the ARK

The ARK scheme
will not charge fees to create or use ARKs
will not limit the number of ARKs you assign
will not limit the kind of content you identify
will not require metadata, nor even persistence
will not mandate use of any particular resolver
Getting involved in ARKs

- Learn more: “ARK Identifiers FAQ”
- Start assigning: n2t.net/e/naan_request

Support open infrastructure
- Join us at ARKsInTheOpen.org
- Community owned infrastructure
- Collaboration between CDL and LYRASIS
ARFs at the Smithsonian Institution

Bess Missell
Metadata Librarian
Smithsonian Libraries
missellb@si.edu

Smithsonian Libraries
Who we are: The Smithsonian Libraries & The Smithsonian Institution.

What we assign ARKs to: Collection metadata & multimedia objects.

When did we start and how many are assigned so far:
- 2015: The Natural Museum of Natural History began assigning ARKs to their collections (over 10 million metadata & 3 million multimedia records).
- January, 2020: new datasetIDs for collection systems were registered in EZID & assigning ARKs in our collection systems began.
- February 26, 2020: Smithsonian Open Access launch with 11,486,102 ARKs on CC0 metadata records and 2,794,786 ARKs on CC0 multimedia records.
- We have assigned over 15 million ARKs and counting ....

Why we chose ARKs
- Project size.
- Cost.
- Ease of implementation.
- Permanence.

Why we are involved in ARKsInTheOpen
- ARKs are a perfect fit for the Smithsonian collections.
Smithsonian Libraries is a network of 21 specialized research libraries, as well as central support services which include Smithsonian Research Online, a bibliography of Smithsonian publication citations and the Institution’s repository. [library.si.edu](http://library.si.edu)

The Smithsonian Institution is the world’s largest museum, education, and research complex, with 19 museums and the National Zoo. [www.si.edu](http://www.si.edu)

- 21 Libraries
- 2.2M Library Volumes
- 80K Smithsonian Research Online
- 772K Website visitors
- 239K Social Media Followers

19 Museums + 1 Zoo
23.2M Visits by Public
155.5M Museum Objects & Specimens
2.2M Library Volumes
2,633 Scholarly Publications
154M Website Visitors
16.6M Social Media Followers
The Smithsonian is Assigning ARKs to our Collection Systems

Examples include records and images for:

**Scientific specimens** from the National Museum of Natural History
http://n2t.net/ark:/65665/381440f27-3f74-4eb9-ac11-b4d933174a3d

**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
The Natural Museum of Natural History began assigning ARKs to their metadata collections.

The Natural Museum of Natural History later began assigning ARKs to their multimedia collections.

Over 10 million ARKs have been assigned to NMNH metadata records.

Over 3 million ARKs have been assigned to NMNH multimedia records.

The Smithsonian released 11.5 million metadata records and 2.8 million multimedia records into the Public Domain.

The Smithsonian chose ARKs to be the global unique identifier (GUID) for these open access images and records.

https://www.si.edu/OpenAccess

#SmithsonianOpenAccess

Over 15 million ARKs and counting...
ARKs were chosen because

- A large number of ARKs will be needed: over 15 million ARKs and growing;
- cost;
- ease of implementation;
- the growth of the ARKs in the Open project encouraged the Smithsonian to choose ARKs as a viable, sustainable identifier.
Smithsonian Libraries

- issues, registers, maintains ARKs for Smithsonian collection systems;
- registers DOIs for Smithsonian publications and research;
- maintains a Smithsonian GUID webpage for SI staff and researchers
  https://library.si.edu/research/guids-help-make-your-data-findable
http://n2t.net/ark:/65665/vk7a466371d-0413-451f-bd76-ca0becc46f94

http://n2t.net/ is the resolver that takes the web call to the EZID service, who then uses the Name Assigning Authority Number (NAAN) to identify who is the registered naming authority. The Smithsonian also has registered datasetIDs (or shoulders) so that EZID passes the web traffic to a specific Smithsonian collection system.

Each Smithsonian collection system is configured to receive the web call from EZID, read the datasetID, and direct the call to the correct server for metadata records or multimedia.

vk7 in the ARK above is registered to metadata records in the Smithsonian American Art Museum (SAAM) collection management system. If vk7 were replaced with bj9 the call would change and go to the image delivery server for SAAM.

vc9 resolves to the Cooper Hewitt image server
https://collection.cooperhewitt.org/ark/vc9

ye3 resolves to the Freer Sackler metadata server
https://collections.si.edu/search/record/ark:/65665/ye3

jy5 resolves to the Freer Sackler image server
https://ids.si.edu/ids/deliveryService?id=ark:/65665/jy5
Using EZID, I register each Smithsonian collection system with our NAAN AND a datasetID with a URL to where the datasetID should resolve.

The Smithsonian wrote a datasetID schema which I follow when I create and register new collection systems:

Two randomly selected lowercase letters (no lowercase L, rm, nm, or fu) + One randomly selected number (2–9)
Each Smithsonian collection system is now configured to automatically generate an ARK when a metadata or multimedia record is saved. The ARK includes the SI NAAN and the datasetID assigned to the collection system.
Challenges for the ARK implementation included…

• tight schedule to meet the February 2020 SI Open Access launch;

• multiple collection management systems, administrators, and IT support;

• encountering IT problems such as:
  
  o identifying the correct syntax for the URL which needs to be registered with EZID:

    The datasetID needs to be included in the URL:
    https://collections.si.edu/search/record/ark:/65665/ye3

  o how to configure each system to receive the URL with ARK datasetID
Phase II of Open Access

Implementing ARKs with archival management systems

- What if collections are split between two collection systems?

- What if objects get moved from one collecting unit to another?

Implementing ARK inflections

Image from the website: https://n2t.net/e/ark_ids.html
http://n2t.net/ark:/65665/ng49ca746aa-75f3-704be053-15f76fa0b4fa

Commercial resolver: n2t.net

EZID Identifiers made easy

Resolve to: https://americanhistory.si.edu/collections/search/object?ark=ng4

Plus 12 more systems ...
Thank you!

Bess Missell
Metadata Librarian
Smithsonian Libraries
missellb@si.edu

ARKs at the Smithsonian Institution
ARKs in the Portico Archive

Karen Hanson, Senior Research Developer
April 23rd 2020
Overview

- **Who we are**
  - Portico - a community supported preservation archive. Work with libraries and publishers to preserve electronic scholarly publications

- **What we assign ARKs to**
  - Every package going into the archive, and a lot of other things (more on that later)

- **When did we start and how many are assigned so far**
  - Started ~2006; Assigned >2 billion ARK IDs

- **Why we chose ARKs**
  - Flexible, opaque, unique, easy to generate, recognized by the community

- **Why we are involved in ARKsInTheOpen**
  - Use ARKs extensively; may adopt some of the new specifications
Portico workflow

Batch of files received e.g. PDF and XML version of articles in a journal issue

Files checked, normalized, and packaged to prepare for preservation

Resulting “archival units” deposited into archive. Each unit = e.g. 1 article
Portico workflow

Batch of files received e.g. PDF and XML version of articles in a journal issue

Files checked, normalized, and packaged to prepare for preservation

Resulting “archival units” deposited into archive. Each unit = e.g. 1 article
Archival unit content structure

Archival Unit

Content Units

Functional Units

Storage Units

Article A

Article A: Version 1

Article A: Version 2

Marked up full text

Page images rendition

Figure graphic component

Publisher supplied XML

Normalized XML (JATS)

PDF

JPEG (high resolution)

PNG (low resolution)
Structure described in metadata

Archival unit: phc5qbrw2a.zip
Structure described in metadata

Archival unit: phc5qbrw2a.zip

Storage Units
- Publisher supplied XML
- Normalized XML (JATS)
- PDF
- JPEG (high resolution)
- PNG (low resolution)

Open BagIt “Bag”

Preservation Metadata
Archival unit content structure

- Archival Unit
  - Article A
    - Article A: Version 1
    - Article A: Version 2

- Content Units
  - Article A: Version 1
  - Article A: Version 2

- Functional Units
  - Marked up full text
  - Page images rendition
  - Figure graphic component

- Storage Units
  - Publisher supplied XML
  - Normalized XML (JATS)
  - PDF
  - JPEG (high resolution)
  - PNG (low resolution)
Use of ARKs supports a self describing archive

- The files are the archive
- The system manages the archive, but the files can exist independently
- ARKs are assigned to abstract concepts and sections of metadata, as well as digital objects
Fig. 1

Example figure!
What did we assign billions of ARKs to?

- Archival units (~110 million) – the “interesting” ones
- Versions of the content (~121 million)
- Archived files... including metadata files (~1.8 billion)
- Sections of metadata (technical metadata, event metadata)

... over 2 billion ARKs
ARK resolver use case: Enhanced Monographs

● “Enhancing Services to Preserve New Forms of Scholarship” – Mellon funded project, a collaboration with NYU Libraries, CLOCKSS and university presses

● Looks at monographs that go beyond text and images (embedded multimedia, interactive features etc.)

● Identify what can be preserved at scale
EPUB Challenge: Remote Resources

Remote resource visually embedded or linked
Problem of external content embedded in EPUBs
Problem of external content embedded in EPUBs
What if we could resolve an ARK to the video?

https://ids.portico.org/ark:/12345/srmkrq29x8
Thank you!

karen.hanson@ithaka.org

Thanks also to my colleague Amy Kirchhoff for helping me put together this presentation.
Who we are
• FamilySearch International

What we assign ARKs to
• Digital images of genealogically significant documents (eg. census records)
• Transcriptions of the data from the digital images
• ‘Persona’ data from those transcriptions
• Genealogies collected from patrons and interviews

When did we start and how many are assigned so far
• We started minting ARKs in 2012, but it took several months to switch over in full.
• We have minted several billion so far.

Why we chose ARKs
• We chose ARK because we wanted something recognizable as an industry effort to standardize long-lived URIs.
• Minting so many identifiers made other options cost prohibitive.
• We needed to control URI, resolution, redirects, etc.

Why we are involved in ARKsInTheOpen
• We are involved in ARKsInTheOpen to contribute our experience.
FamilySearch International - A Brief Introduction

- Originally The Genealogical Society of Utah
- We help people connect with their families through:
  - Providing engaging discovery experiences
  - Publishing guidance on how to do family history research
  - Acquiring and publishing billions of source records from around the world
  - Creating software systems to aid in researching and collaborating on family history
  - Providing all of this at no cost to our patrons - It’s free!
- We also maintain significant long-term digital preservation systems
  - Two independently implemented and maintained systems each holding two copies of all artifacts
  - One system in the public cloud; one system literally in a cave in the mountains
  - Tens of petabytes and billions of artifacts
- Open to anyone at no cost; Fully supported by the Church of Jesus Christ of Latter-day Saints

www.familysearch.org
Artifact Processing Abstraction
## Searching For Ancestors

### Records

<table>
<thead>
<tr>
<th>Name</th>
<th>Events</th>
<th>Relationships</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas P Creighton</td>
<td>birth: 1894</td>
<td>spouse: Maudie M Creighton</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>residence: 1940;1935</td>
<td>children: David M Creighton, Robert T Creighton</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>Texas</td>
<td>110th Company 8th Regiment Fort Crockett Galveston Texas</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>Ward 4, San Bernardino, San Bernardino Judicial Township, San Bernardino, California</td>
<td>military service: 29 June 1917 Galveston, Texas, United States</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>residence:</td>
<td>110th Company 8th Regiment M C Fort Crockett Galveston Texas</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>military service: 29 June 1917</td>
<td>Galveston, Texas, United States</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>residence: 1927</td>
<td>father: J R Creighton</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>Mineral Wells, Palo Pinto, Texas, United States</td>
<td>mother: Flora Boynton</td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td>Thomas William Creighton</td>
<td>birth: 23 February 1927</td>
<td></td>
<td><img src="#" alt="View" /></td>
</tr>
<tr>
<td></td>
<td>Texas Birth Certificate</td>
<td></td>
<td><img src="#" alt="View" /></td>
</tr>
</tbody>
</table>
Maudie M. Creighton -- ...
.../ark:/61903/1:1:K98H-2GL
David M. Creighton -- ...
.../ark:/61903/1:1:K98H-2GG
Robert T. Creighton -- ...
.../ark:/61903/1:1:K98H-2GP

Thomas Percy Creighton Details -- ...
.../ark:/61903/4:1:25V8-3J5

Census page with context -- ...
.../ark:/61903/3:1:3QSQ-G9MT-N9Z
personaUrl=%2Fark%3A%2F61903%2F1%3A1%3AK98H-2G2
<table>
<thead>
<tr>
<th>Name</th>
<th>Titles and Terms</th>
<th>Last Place of Residence</th>
<th>Gender</th>
<th>Age</th>
<th>Marital Status</th>
<th>Race</th>
<th>Birth Year (Estimated)</th>
<th>Birthplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas P Creighton</td>
<td></td>
<td>San Bernardino, California</td>
<td>Male</td>
<td>46</td>
<td>Married</td>
<td>White</td>
<td>1894</td>
<td>Texas</td>
</tr>
<tr>
<td>Maudie M Creighton</td>
<td></td>
<td>San Bernardino, California</td>
<td>Female</td>
<td>35</td>
<td>Married</td>
<td>White</td>
<td>1905</td>
<td>Texas</td>
</tr>
<tr>
<td>David M Creighton</td>
<td></td>
<td>San Bernardino, California</td>
<td>Male</td>
<td>10</td>
<td>Single</td>
<td>White</td>
<td>1930</td>
<td>California</td>
</tr>
<tr>
<td>Robert T Creighton</td>
<td></td>
<td>San Bernardino, California</td>
<td>Male</td>
<td>6</td>
<td>Single</td>
<td>White</td>
<td>1934</td>
<td>California</td>
</tr>
<tr>
<td>Item Number</td>
<td>Film/Digital Note</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>California (San Bernardino County)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Catalog Record**

California, 1940 population census : population schedules

**Citation**

https://www.familysearch.org/ark:/61903/1:1:K98H-2G2 (Thomas Percy Creighton) as json
## Organization and Volume of Minting

<table>
<thead>
<tr>
<th>Namespace or Name Assigning Authority</th>
<th>Description</th>
<th>Approximate Count In Millions</th>
<th>Annual Increase In Millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>Historical record persona</td>
<td>8800</td>
<td>1,511</td>
</tr>
<tr>
<td>1:2</td>
<td>Historical record</td>
<td>5300</td>
<td>452.27</td>
</tr>
<tr>
<td>2 (1–3)</td>
<td>Pedigree data</td>
<td>1500</td>
<td>73.05</td>
</tr>
<tr>
<td>3 (1–4)</td>
<td>Digital images of documents</td>
<td>4300</td>
<td>344.64</td>
</tr>
<tr>
<td>4</td>
<td>FamilyTree person records</td>
<td>1400</td>
<td>43.2</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>21300</strong></td>
<td><strong>2,424</strong></td>
</tr>
</tbody>
</table>
Managing Access and Routing

- Routing of ARKs is basically the same as all other resources managed at familysearch.org.
- https://www.familysearch.org/ark:/61903/1:1:K98H-2G2 is seen by DTM. If Accept header specifies html, forward to appropriate application (typically in Heroku); if json, forward to the 1:1 resolver.
- https://www.familysearch.org/ark:/61903/3:1:3QSQ-G9MT-N9ZF is seen by DTM. Based on Accept header it forwards to the image viewer app (Heroku) or the 3:1 resolver. The 3:1 resolver will authorize and redirect to a temporary signed S3 URL.