Broken for all: the case for persistent identifiers for digital cultural heritage resources



- 1. John Kunze, California Digital Library
- 2. Deb Holmes-Wong & Zahid Rafique, USC Digital Library
- 3. Megan Lohnash, California Revealed & Allison Lund, Metropolitan New York Library Council
- 4. Adrian Turner, California Digital Library
- 5. Matthew McKinley, Omeka

June 2021





Why persistent URLs?



AWWW...DON'T

It's just a 404 Error!

What you're looking for may have been misplaced in Long Term Memory.

Because persistent, reliable, unbroken web links (URLs) are rare

- Average URL lifetime is 100 days
- But why not just do an internet search when you need one? Because:
 - Scholars, writers, and researchers compile reference (URL) lists with great care
 - URLs cannot be controlled once they're in the wild:
 books, articles, news, Wikipedia, theses, social media, DPLA, ...



Where can we go for stable URLs?



I know! We can trust URLs from *libraries, archives, and museums*, since they safeguard cultural and scientific heritage, right?

NOPE

Shockingly, few memory organizations and their vendors have software solutions that support persistent URLs.



How cultural heritage links break

GLAMs use DAMS,

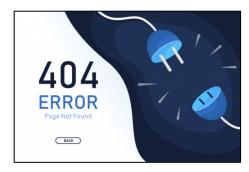
DAMS use databases,

databases re-generate new URLs,

frequently.

For example, widespread URL change can occur

- when the database is reloaded,
- when the software is updated,
- when the content migrates between DAMS vendors





Broken for all



- all citations in articles, books, blogs, wikipedia, theses, ...
- all links in twitter, Facebook, Linked Open Data, library MARC records, ...
- all aggregator links, such as Digital Public Library of America, Calisphere, ...



Permalinks – persistent URLs

Permalink: a persistent URL

- once published, it won't need to be changed
- as system changes, the permalink keeps working

Example: many WordPress URLs are permalinks



http://arks.org/blog/the-louvre-collection-goes-online-with-483000-arks

- if the database or version changes, software keeps them working
- if the domain changes, content managers "redirect" to keep them working



Persistent identifiers Permalinks

A *persistent identifier* is an internally labeled permalink.

- label tells you what kind of PID (persistent ID) it is
- well-known kinds: ARK, DOI, Handle, PURL, URN
- some kinds have fees; all make you redirect; none prevent error, fire, war, ...
- but one big advantage of PIDs: in the wild they self-declare *intent*





PID examples and labels

```
https://n2t.net/ark:/53355/cl010066723
https://doi.org/10.1038/sdata.2018.95
https://hdl.handle.net/4263537/4001
https://n2t.net/urn:nbn:fi:tkk-004781
```

404 Error
Don't panic, and make sure to watch your oxygen levels.

Return to safety

GLAMs must go to great lengths to support PIDs or Permalinks on their own.



DOIs for Discovery: A CONTENT dm Use Case

Open Repositories June 9, 2021

Deborah Holmes-Wong, Director, University of Southern California Digital Library dhwong@usc.edu Zahid Rafique, MIS Director, University of Southern California Digital Library zrafique@usc.edu >

Background

University of Southern California Libraries

- Digital Library
- Digital Repository

Information Sciences Institute

- FaceBase
- GUDMap
- RBK
- Synapse

Other units produce

- Technical reports
- Data sets
- Journals
- News reports

Issue: Migration & Citation

- Migrating can't keep/reuse URLs from previous systems
 - Migrated 3 times in 25 years
 - 3 times all links broke: citations to nothing
 - Reharvest (Calisphere, Worldcat, etc.)
- We now have
 - o 500,000 objects, 1.5 million files
 - 1.9 million unique users per year (Google Analytics)
 - o 1000 license/use requests/year
 - Classroom/Exhibit use w/Scalar.
- We are migrating to a new system.

This time must be different!



DOIs

Discovery: We need to provide a URL that doesn't change over time for data harvesters and end users alike because we are going to migrate our content into new systems and researchers need to be able to get back to our assets from other works.

CONTENTdm Issues

- No support for permalinks
- CONTENTdm(CDM) packages assets/digital objects simple items/compound objects
 - Metadata + files
 - Not individual files
- Packaging
 - Best Individual images, maps, books, ETDs.
 - Worst EAD folder level objects, multipart objects
- Difficult to get the DOIs into CDM
- DOIs available in OAI harvest only (We use our own data provider not CDM)



Implementation

Example

DOI:

https://doi.org/10.25549/whit-m12485

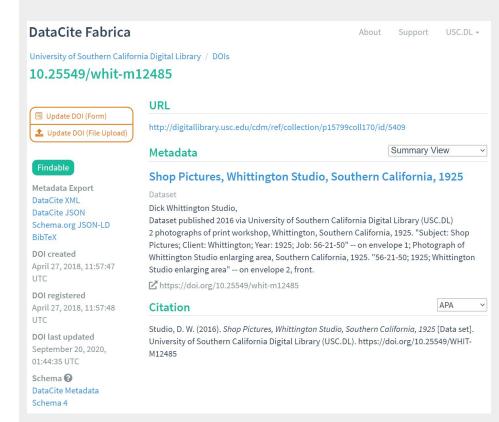
Base URL: https://doi.org

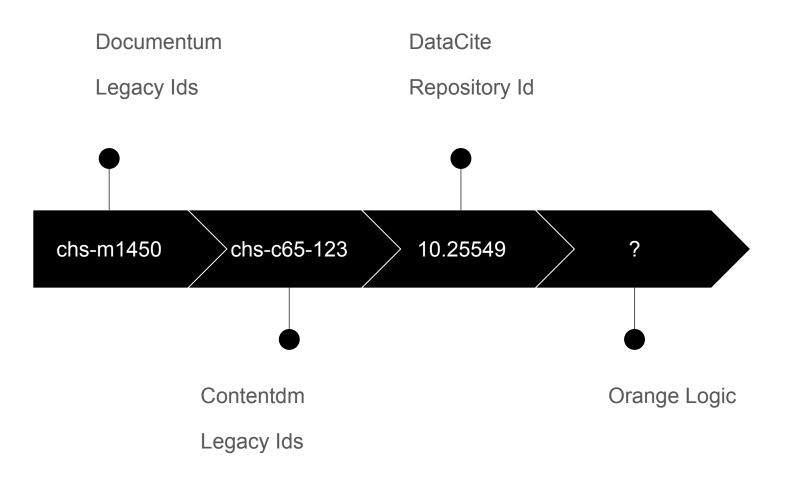
Repository Id: <u>10.25549</u>

Resource Id: whit-m12485

Resource:

http://digitallibrary.usc.edu/digital/collection/p15799coll170/id/5409





Components

- DataCite REST API Mint the DOIs
- DataCite MDS API Create Metadata in Fabrica
- 3. OCLC Contentdm API Pull resource metadata
- Local MYSQL database Processing queue
- Java based Application Glue all of it together Updates OAI Data Provider



Ongoing Challenges

Granularity

Migration

Maintenance

What unit gets the DOI?

- Files
- Digital Objects?

Decision: Digital Objects

Metadata + Files

Updating to include DOI

Decision: Only available through data provider.

Include DOI in new system as migrating and creating new content.

Managing DataCite

DOIs only work if maintain them.

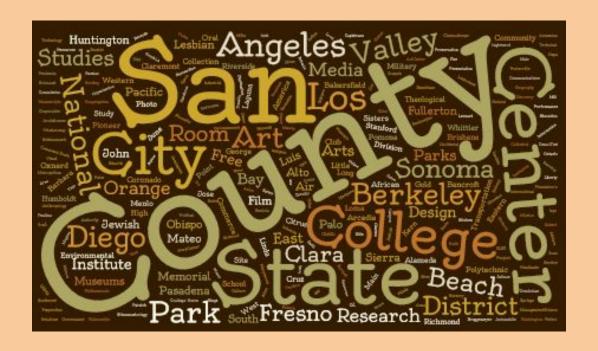
- Who's responsible for maintaining DOIs not managed by Digital Library or ISI?
- Will they know they need to manage them?



Islandora to Archipelago

June 9, 2021

Megan Lohnash - mlohnash@californiarevealed.org
Digital Repository Manager



Who is California Revealed?

California Revealed's History

- 2010 California Audiovisual Preservation Program founded
 - ContentDM
 - Internet Archive
 - Linear Tape Open (LTO) storage
 - OAI harvests of records by aggregators
- 2015/2016 migration to Islandora
- 2017 became California Revealed Program
- 2018 public facing website launched
- 2018 restart of aggregator harvesting
- 2020/2021 migration to Archipelago

Our Identifiers and URLs

- Object Identifier
- Internet Archive URL
- ARK Identifier
- Islandora 7 PID (or Node Id)

Why do we care about Persistent IDs and URLS? Access continuity

- Long term management of digital assets
- Aggregator harvest
- References to California Revealed
 - Exhibits and Publications
 - Researchers
 - Educator resources

Persistent URLs in Archipelago

Strategies for Migrations and Out-of-the-Box(es)

archipelago.nyc/
github.com/esmero/archipelago-deployment



Metropolitan New York Library Council

Allison Lund - <u>alund@metro.org</u>

Digital Projects & Metadata Librarian

Archipelago

- An open source repository software
- Development home at the Metropolitan New York Library Council (METRO)

- Slim, smart open source repository software built on top of Drupal 8/9
- Custom Drupal Modules
 - Strawberry Field modules
- Curated list of well maintainedCommunity Drupal Modules
- A custom **Docker** deployment strategy that delivers all required Services
 - o Local / Production /
 - o Or "Bare-metal" for rebel-hearts
- Getting started **Configurations**
- Basic Metadata Templates (IIIF, MODS, DC, LoD Display all via <u>Twig</u>)
- Matching **Webforms** for ingesting



Archipelago Links & Facts

for further reading!

Links:

archipelago.nyc
Archipelago Playground
Google Group
Technical roadmap
Active development repos
Archipelago 1.0.0-RC2 Specs and
features

Facts:

Current stable branch is 1.0.0-RC2
Archipelago is GNU V3
Archipelago has an Advisory Board
Archipelago has a lovely caring
Community
Archipelago extends and empowers
Drupal 8/9



Archipelago and Persistent URLs

- Out-of-the-box(es)Deployments
- Technical Strategy for Persistent URL support

Persistent URL for every Archipelago Digital Object or Collection:

- https://yourdomain/do/UUID
- Whether ingested via webform, remoteAPI (drush), or batch import (via the <u>AMI module</u>)

Persistent URLs for **IIIF manifests** too:

 http://yourdomain/do/UUID/metada ta/iiifmanifest/default.jsonld

More info on <u>Archipelago's **File Persistence</u> <u>Strategy</u>



- Migration from Islandora 7
 to Archipelago
- Background Context for METRO's involvement

- METRO's development team, led by Archipelago's lead architect & METRO's Assistant Director for Digital Strategy <u>Diego Pino Navarro</u>, has been consulting with the development team at <u>Born Digital</u>, who is handling the migration of California Revealed from Islandora 7 to Archipelago
- METRO is also in the process of migrating our shared repository, Digital Culture of Metropolitan New York (<u>dcmny.org</u>), from Islandora 7 to Archipelago



- Migration from Islandora 7
 to Archipelago
- Background Context for METRO's involvement

- Most (almost all) of the collections in DCMNY were harvested into DPLA in the mid-2010s timeframe, through the now-sunsetted Empire State Digital Network DPLA Service Hub (closed since 2019)
- Any DPLA-harvested/aggregated DCMNY collections and materials will need to have redirects once migrated to Archipelago
 - Some of our DCMNY partners also use the current (I7) URLs on their own websites
- As always, we want to best support consistent and reliable access to our partners trusted digital resources for researchers, educators, and users of all shapes and sizes



- Migration from Islandora 7
 to Archipelago
- Technical Strategy for Persistent URL support

- Current data structure in California Revealed:
 - Includes the PID for Islandora 7 (namespace:UUID)
 - ARK IDs (that currently
 point to the I7 ID ~more
 about the ARK IDs a little
 later)
 - O Both of these two primary IDs, along with all of the other important IDs currently found in California Revealed's I7 Repository, will be part of the data structure for all materials migrated to Archipelago



- Migration from Islandora 7
 to Archipelago
- Technical Strategy for Persistent URL support

- During the initial migration phase, the <u>Archipelago Multi</u>
 <u>Importer</u> (with <u>Islandora 7 SOLR</u>
 <u>plugin</u>) will be configured to use:
 - the UUID from the PID
 (namespace: UUID) for the
 I7 Digital Object for the
 new UUID for the
 Archipelago Digital Object



- Migration from Islandora 7
 to Archipelago
- Technical Strategy for Persistent URL support

For the next step in the process for persistent URL support (redirection):

- Using a specially-configured internal resolver in Archipelago following the pattern(s) established by a controller.
- Take the PID (namespace:uuid), and, if it is valid, redirect permanently to the Archipelago /do/uuid

**Also under consideration is using an existing path alias Drupal extension (<u>Redirect</u>; or <u>Pathauto</u>)



- Migration from Islandora 7 to Archipelago
- Technical Strategy for Persistent URL support

- For ARK IDs in California Revealed (~back to Megan for more context on ARKs in CAR)
 - development teams at METRO and Born Digital are currently exploring tapping into the ARKs API (via the Internet Archive for C-AR) to update existing URLs to point to Archipelago /do/uuid's
 - could take place after or during the migration
 - generate new ARK IDs for future ingests (if desired)





Thank you!

Contacts:

- Diego Pino Navarro (<u>dpino@metro.org</u>)
 Archipelago's Lead Architect and A.D. for Digital Strategy at METRO
- Allison Lund (<u>alund@metro.org</u>)

 Digital Projects & Metadata Librarian at METRO
- Nate Hill (nhill@metro.org)
 Director at METRO

Born Digital:

- Pat Dunlavey (pat@born-digital.com)
 Senior Digital Repository Developer
- **Derek Merleaux** (<u>derek@born-digital.com</u>)
 Senior Consultant and Project Manager

More Information & Greatest Hits J

- https://archipelago.nvc
- Deploying your very own Archipelago (20 minutes):
 - https://github.com/esmero/archipelago-deployment
- Archipelago Documentation:
 https://github.com/esmero/archipelago-documentat
 ion/tree/1.0.0-RC2
- o Technical Roadmap: https://github.com/esmero/archipelago-deployment /issues/103
- Specs & Features for RC2:
 https://docs.google.com/document/d/1k935L6grlUki
 45GjBuVCnl5kWsw6RiEVBqd00pBv3Xo/edit
- Google Community Group: https://groups.google.com/g/archipelago-commons















Broken for All: Aggregation Considerations

Open Repositories 2021 2021-06-09

Adrian Turner California Digital Library adrian.turner@ucop.edu



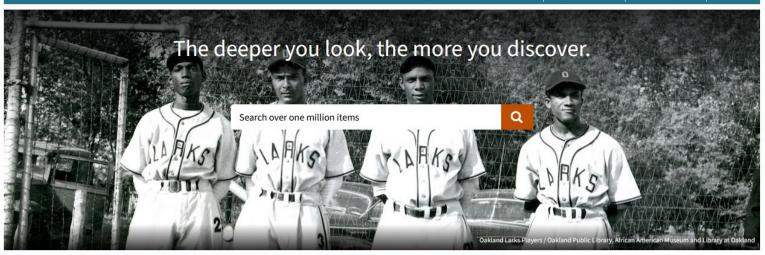


Contributing Institutions

Collections

Exhibitions

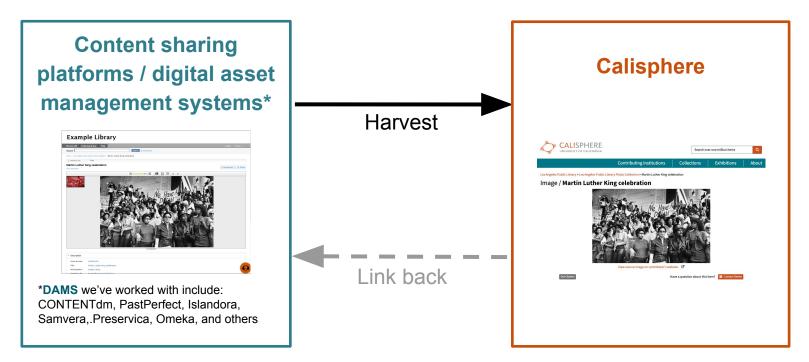
About



Calisphere is your gateway to digital collections from California's great libraries, archives, and museums. Discover over 1,975,000 images, texts, and recordings. Watch an <u>introductory video</u> to learn more.

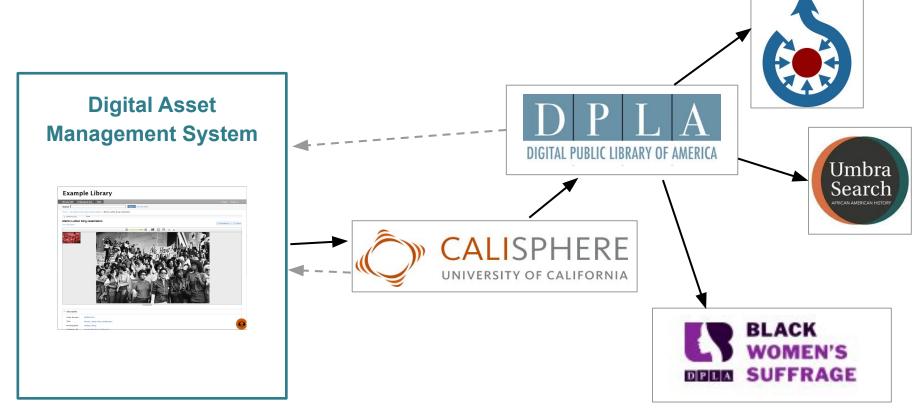




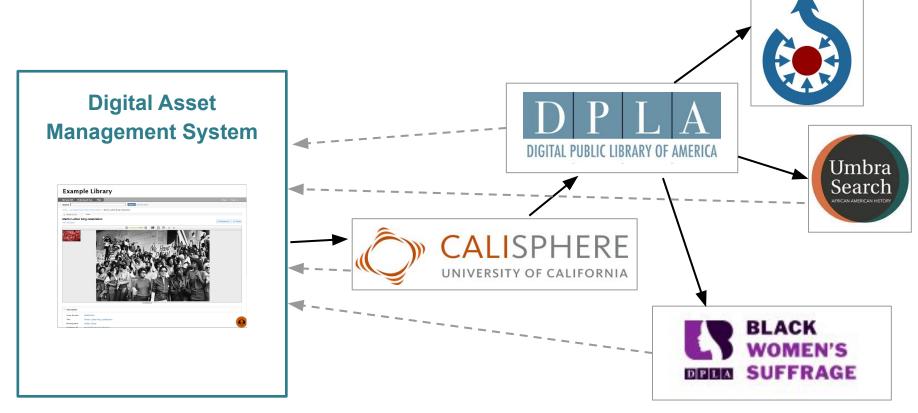


https://cdm12345.org/digital/item1

https://calisphere.org/item/85691e448b4d8f8884e6a/



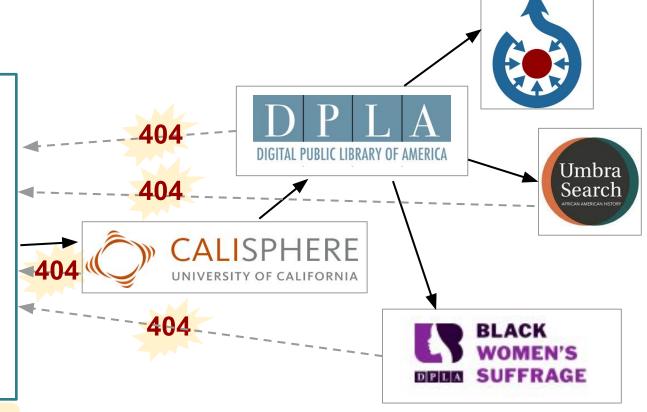
https://cdm12345.org/digital/item1



https://cdm12345.org/digital/item1







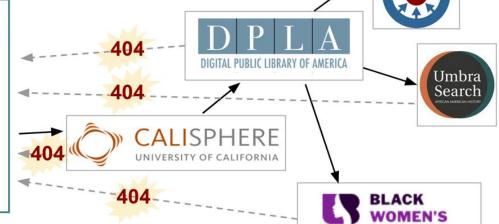
https://examplelibrary.org/item1





Digital Asset

Management System 1



DEUM SUFFRAGE

https://cdm12345.org/digital/item1

DAMS MIGRATION



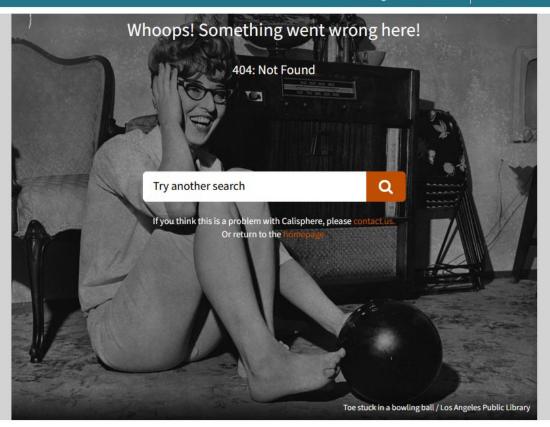
Contributing Institutions

Collections

Exhibitions

About

Q





As a **DAMS implementor**

- At minimum, redirect URLs for slight URL changes or migrations
- Ideally, use persistent URLs from a known PID scheme
- Include persistent URL support in your DAMS selection requirements
- Register this as a functional priority to DAMS developers, in your feedback and features wishlists

As a **DAMS** developer or service provider

- Please explore options for supporting persistent URLs as a core feature
- If facilitating migrations, support strategies to redirect URLs

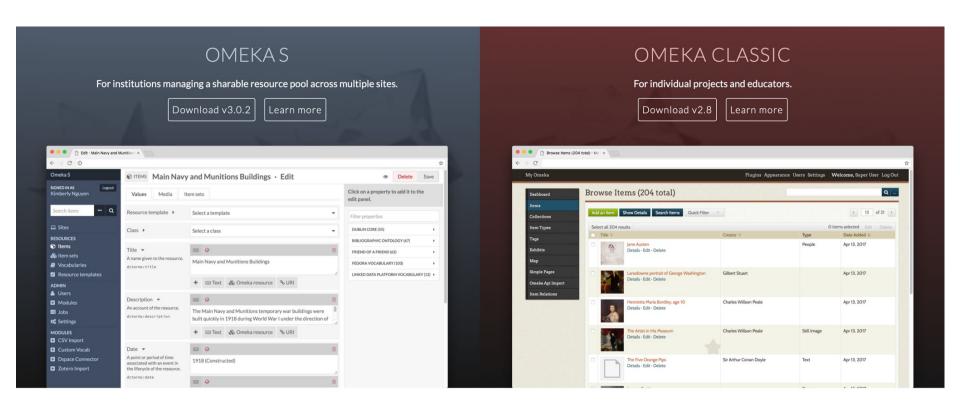
As a digital collections aggregator

- Proactively identify broken URLs; work with contributors on resolving the links
- Promote awareness of these issues and advocate for persistent URLs

Integrating Persistent Identifiers into Omeka S

Matthew McKinley
Programmer, Omeka.org
matthewjamesmckinley@gmail.com





Omeka provides open-source web publishing platforms for sharing digital collections and creating media-rich online exhibits.



Advanced Search Plus

By Daniel Berthereau

Add some fields to the advanced search form (before/after creation date, has media, media type, etc.).

Alt Text

By John Flatness

Customize alt text for media

Analytics Snippet

By Daniel Berthereau

Add a snippet, generally a javascript tracker, at the end of the public or admin pages, and allows to track json and xml requests.

Download

Latest Version: 3.3.5.3 Updated: October 27, 2020

Download

Latest Version: 1.2.1 Updated: February 25, 2021

opauted. 1 05. daily 20, 2021

Download

Latest Version: 3.3.3.1

Updated: February 15, 2021

PID Module Requirements

- Mint & assign new PIDs via API
- Discover & assign existing PIDs within metadata
- Resolve PID to user-friendly landing page
- Ability to (eventually) switch between multiple PID services



Existing ARK module

https://gitlab.com/Daniel-KM/ Omeka-S-module-Ark

Ark & Noid (module for Omeka S)

New versions of this module and support for Omeka S version 3.0 and above are available on GitLab, which seems to respect users and privacy better than the previous repository.

Ark & Noid is a module for Omeka S that creates and manages ark identifiers. They can be used in urls the admin and the public sides with the module Clean Url.

The ark identifiers can replace the default cool URIs of each record, that corresponds to the simple number of a row in a table of the database.

Arks are short, opaque, meaningless, universal, unique and persistent ids for any records. The acronym "ark" means "Archival Resource Key". This is a reference to the Noah's ark (digital documents will have a long life) and to the world of archives (Omeka can be an institutional archive) too. Optionally, the identifiers can be resolved via a service as N2T, the Name-to-Thing Resolver.

Arks may be built with the utility Noid, that creates nice opaque identifiers, that is integrated too via the library Noid4Php.

See a living example in the bibnum of PSL or in Bibliothèque patrimoniale of Mines ParisTech.

If you don't have an authority number, the plugin will create a non standard ark, but nevertheless a unique and opaque identifier that can be managed.



Why make another Module?

- Need for PID types other than ARKs
- More fine-grain control
 - Selective assignment, assign from existing, etc.
- Ownership by Omeka team



Issue: Shifting 'target'

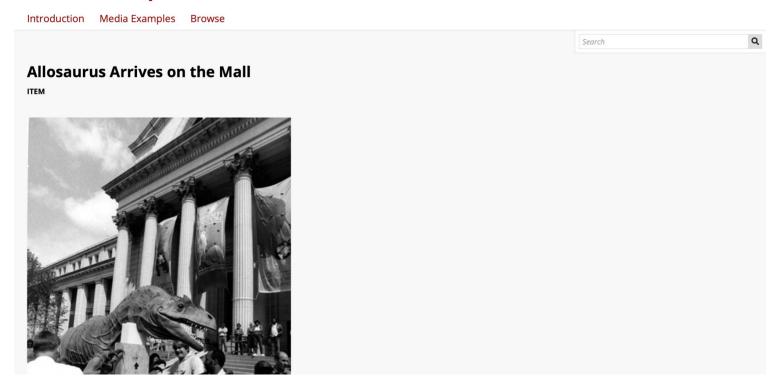
Mall History Content

About Monuments and Memorials ♥ Daily life on and around the Mall 🕶 Events See Also Q Search Allosaurus Arrives on the Mall ITEM Title Allosaurus Arrives on the Mall Description In April 1987 a one-ton fiberglass Allosaurus sculpture was moved into the National Museum of Natural History. A crowd formed on the Mall at the steps of the museum to watch the dinosaur lifted into the building with a winch. The 20-foot-long model was featured in the "Dinosaurs, Past and Present," a temporary exhibit displaying artistic representations of dinosaurs constructed in the 1800 and 1900s. After three months, the Allosaurus and other works traveled to museums across North America. Source Smithsonian Institution Archives. View Original. Date 04/28/1987 (Allosaurus Arrival) 06/04/1987 ("Dinosaurs, Past and Present" Opens) 08/31/1987 ("Dinosaurs, Past and Present" Closes)

http://dev.omeka.org/omeka-s-sandbox/s/mallhistory/item/1138

Issue: Shifting 'target'

Media Examples



http://dev.omeka.org/omeka-s-sandbox/s/media-examples/item/1138

Potential Solutions

- 'Update PIDs' per site (moving target)
- Assign new PID per site (multiple targets)
- Site-agnostic landing page (single target)



Single Target Approach

http://dev.omeka.org/omeka-s-sandbox/api/items/1138

```
"@context": "http://dev.omeka.org/omeka-s-sandbox/api-context",
"@id": "http://dev.omeka.org/omeka-s-sandbox/api/items/1138".
"@type": [
 "o:Item",
 "dctype:StillImage"
"o:id": 1138,
"o:is public": true.
"o:owner": {
 "@id": "http://dev.omeka.org/omeka-s-sandbox/api/users/1",
 "o:id": 1
"o:resource class": {
 "@id": "http://dev.omeka.org/omeka-s-sandbox/api/resource classes/33",
  "o:id": 33
"o:resource template": null.
"o:thumbnail": null,
"o:title": "Allosaurus Arrives on the Mall",
"thumbnail display urls": {
 "large": "http://dev.omeka.org/omeka-s-sandbox/files/large/50fcad8384a7b87868cb77eb656cb9d365f09d61.jpg",
 "medium": "http://dev.omeka.org/omeka-s-sandbox/files/medium/50fcad8384a7b87868cb77eb656cb9d365f09d61.jpg",
  "square": "http://dev.omeka.org/omeka-s-sandbox/files/square/50fcad8384a7b87868cb77eb656cb9d365f09d61.jpg"
"o:created": {
  "@value": "2017-04-25T18:39:39+00:00",
  "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
"o:modified": {
  "@value": "2017-06-05T20:05:00+00:00",
  "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
"o:media": [
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/media/3404".
    "o:id": 3404
"o:item_set": [
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/item_sets/275",
    "o:id": 275
"o:site": [
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/sites/4",
    "o:id": 4
```

Single Target Approach

http://dev.omeka.org/omeka-s-sandbox/api/items/1138

"o:id": 4

```
"@context": "http://dev.omeka.org/omeka-s-sandbox/api-context",
"@id": "http://dev.omeka.org/omeka-s-sandbox/api/items/1138"
  "o:Item",
 "dctype:StillImage"
"o:id": 1138,
"o:is_public": true,
"o:owner": {
  "@id": "http://dev.omeka.org/omeka-s-sandbox/api/users/1",
  "o:id": 1
"o:resource_class": {
  "@id": "http://dev.omeka.org/omeka-s-sandbox/api/resource_classes/33",
"o:resource template": null,
"o:thumbnail": null,
"o:title": "Allosaurus Arrives on the Mall",
"thumbnail_display_urls": {
 "large": "http://dev.omeka.org/omeka-s-sandbox/files/large/50fcad8384a7b87868cb77eb656cb9d365f09d61.jpg"
  "medium": "http://dev.omeka.org/omeka-s-sandbox/files/medium/50fcad8384a7b87868cb77eb656cb9d365f09d61.jpg"
  "square": "http://dev.omeka.org/omeka-s-sandbox/files/square/50fcad8384a7b87868cb77eb656cb9d365f09d61.ipg"
"o:created": {
  "@value": "2017-04-25T18:39:39+00:00",
  "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
                                                                                                       Content
"o:modified": {
  "@value": "2017-06-05T20:05:00+00:00".
  "@tvpe": "http://www.w3.org/2001/XMLSchema#dateTime"
                                                                                                       negotiation
"o:media":
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/media/3404",
    "o:id": 3404
"o:item_set": [
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/item sets/275",
   "o:id": 275
"o:site": [
    "@id": "http://dev.omeka.org/omeka-s-sandbox/api/sites/4".
```

http://dev.omeka.org/omeka-s-sandbox/perm/1138

Allosaurus Arrives on the Mall

ITEM



Planned Features

- Poll user community for initial PID schema
- Option to mint/assign PIDs for all new items
- Selectively mint/Assign PIDs on batch item edit screen
- Discover/assign PIDs from existing metadata field(s)



Thank you!

https://forum.omeka.org

Matthew McKinley
Programmer, Omeka.org
matthewjamesmckinley@gmail.com

