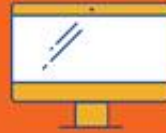# DOIs, PURLS, URIs Oh My! An Introduction to Persistent Digital Identifiers

Presented by DLC PURL Working Group

James R. Jacobs, Stanford University

Will Stringfellow, Vanderbilt University

Ashley Dahlen, GPO

Working Group Members

Renée Bosman, DLC Member

Alicia Kubas, DLC Member

Rick Mikulski, DLC Member

Laura Sare, DLC Member

Robbie Sittel, DLC Member

Will Stringfellow, DLC Member

James R. Jacobs, Community Member

Shari Laster, Community Member

Alec Bradley, GPO Liaison

Ashley Dahlen, GPO Liaison

Cynthia Etkin, Designated Federal Official

Electronic Document Retrieval

- Electronic objects, such as files and documents are stored on file servers inside of a computer network

- File server is a computer or computers within a network which store information, files, etc. and respond to requests for information

- When accessing a document or object, your computer first accesses a network via an IP address (ex. 104.16.165.242 = IP address of **govinfo** network)

- Then sends a request to the file server in the network to search for a specific object

- File server responds to the request by searching its inventory of stored objects and either provides object, if available or returns an error message
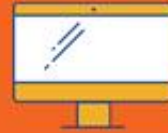
Uniform Resource name (URN)

- URNs are URIs which are globally unique identifiers defined within URN namespaces

- URN namespaces are a collection of URNs that are:
  1) Unique
  2) Assigned in a consistent and managed way
  3) Assigned to a common definition

- Always starts with prefix **urn:**

- A URN representing a document can be translated into a URL by a "resolver."

- URN namespaces are required to be registered with the Internet Assigned Numbers Authority (IANA)

Uniform Resource Identifiers (URI)s

- URIs are a string of characters used to refer to a specific resource
  - Helpful for finding resources without needing to know what network to access
- Follow predefined syntax rules, but are
- The 1:1 relationship between a resource and its URI is what makes URIs an efficient way to unambiguously identify something
- URI vs URL. A URL is a type of URI

Persistent Identifiers (PI)

What are persistent identifiers?

- Long-lasting references to a document, file, web page, or other object over the internet.

- Render the traditional identifiers (ex. ISBN or URL) resolvable and retrieval on the web.

- Persistent identifiers are used for digital objects if available long enough that its default URL will likely change — over several hardware and software generations.

- Persistence is both a matter of good policy and practice AND technology.

Persistent Identifier Schemes:

- Permalink

- Archival Resource Key (ARK)

- Handles

- Digital Object Identifier (DOI)

- Persistent Uniform Resource Locator (PURL)

Different schemes all act as redirections, from the identifier to the current URL of the object.

Permalink

- A persistent identifier (PI) and sometimes used synonymously with PI

- Most often used within a content management system

- Help with search engine optimization

- Characteristics of a good Permalink:
  - It is secure (http**s**)
  - Describes the content of the webpage
  - Words separated by hyphens

Permalinks and PURLs are both are PIDs and redirect to a web resource

- A URL's domain is not changed with a Permalink and links are intended for years.

  Permalink example: https://www.fdlp.gov/superintendent-of-documents-public-policies

- PURLs have an independent domain and links are intended for decades.

  PURL example: https://**purl**.fdlp.gov/GPO/gpo137179

  Redirects to: https://www.**govinfo**.gov/content/pkg/GOVPUB-GP3-fdd7fc362c00c359f1c307d35747e7fb/pdf/GOVPUB-GP3-fdd7fc362c00c359f1c307d35747e7fb.pdf

Archival Resource Key (ARK)

- Developed by the California Digital Library, an **Archival Resource Key (ARK),** is a **persistent, actionable naming scheme**.
- Links the user to three things:
  1) the object
  2) its metadata, and
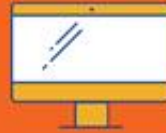 3) the commitment statement of the provider.

Statement of commitment for objects hosted by the UNT Libraries –
"The UNT Libraries' assigns identifiers within the ARK domain under the NAAN 67531 and according to the following principles:
* No ARK shall be re-assigned; that is, once an ARK-to-object association has been made public, that association shall be considered unique into the indefinite future."
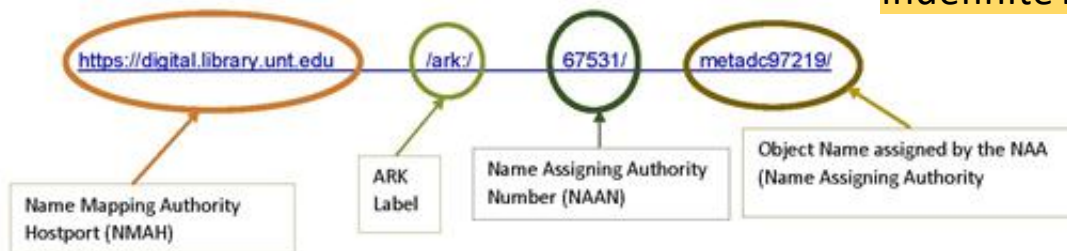
The ARK scheme produces a special kind of URL that formalizes the roles of the Name Assigning Authority (NAA) and the Name Mapping Authority (NMA).

https://digital.library.unt.edu   /ark:/   67531/   metadc97219/

Name Mapping Authority Hostport (NMAH)

ARK Label

Name Assigning Authority Number (NAAN)

Object Name assigned by the NAA (Name Assigning Authority

Handles

- The Handle System is a registry for persistent identifier strings, and is the basis for the DOI system.

- Handles consist of a prefix (for example, 20.1000/100) which indicates the namespace and naming authority, and a suffix, which indicates the local name for the digital resource

Digital Object Identifier (DOI)

- DOI is a "digital identifier of an object" not "identifier of a digital object"

- A DOI is a series of numbers and punctuation that provides a unique identification for objects of any type.

- DOI is a framework for persistent identifiers that provides actionable, unique identification for digital resources. Can be assigned to any entity including physical, digital or abstract.

- DOIs contain a prefix and a suffix separated by a slash (prefix/suffix) 10.NNNN/NNN

DOI (continued)

Example doi 10.1000/182   *note many are displayed and resolved as URLs

https://doi.org/10.1000/182

Prefix 10.1000

Suffix 182

doi:10.   1000   /   182

Designates handle as   Identifies   Item ID
part of DOI namespace   registrant
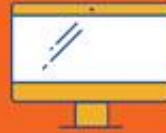
Persistent Uniform Resource Locator (PURL)

- Permalink intended to last for years.

- GPO has been using PURLs since 1998.

Finding the PURL

PURLs are found in catalog records.

Example: Catalog of U.S. Government Publications

Finding the PURL

PURLs are found wherever links are used

Example: LibGuides

Source:
https://libguides.daltonstate.edu/c.php?g=24569&p=288066

How PURLs work

- User clicks on PURL

- Their computer/device hits GPO's PURL server and are redirected

- Redirect takes the user to where the content described in the record resides

- GPO manages redirects so libraries do not have to modify their links

- Redirect may be changed over time, but PURL stays the same

Content may reside:

govinfo

Agency website

GPO server

Digital Access Partner

Web Archive Collection

End user   GPO PURL server
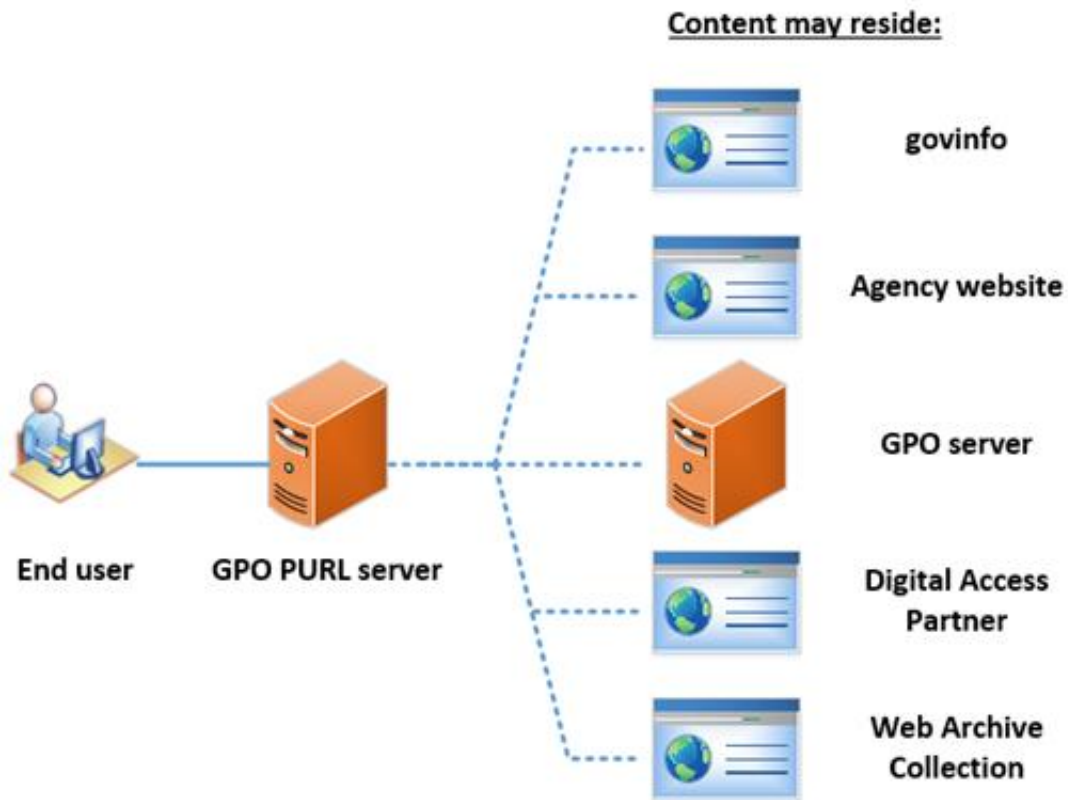
September 2020 *estimates*

Where PURL redirects to	Number of PURLs
GPO server	99,702
govinfo	87,442
Agency website	30,437
Digital Access Partner	26,797
Web Archive	18,179
Total	**262,557**

Pie chart showing: GPO server 38%, govinfo 33%, Agency website 12%, Digital Access Partner 10%, Web Archive 7%
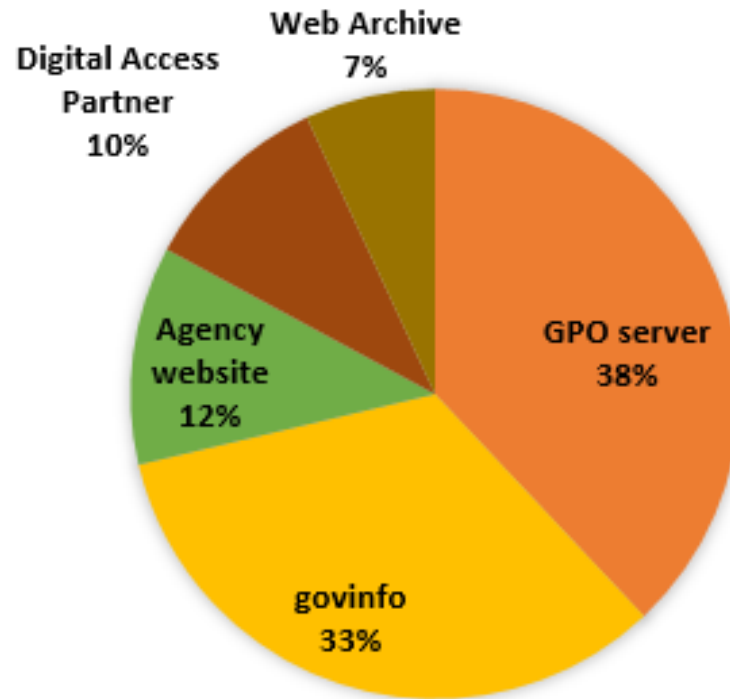
Redirects – GPO cataloger has to pick how to direct user to content described in the record

govinfo	Easy and stable
Agency website	How stable is the site? Update frequency? Does the agency remove content over time?
Harvest locally for GPO server	Monograph? Serial? Multipart mono? Chapters broken out into separate files? Special features that cannot be captured when copied?
Digital Access Partner	Any agreement in place where content is already archived? (ex. NASA NTRS)
Web Archive	Successful capture of all content and features?

A record may have multiple PURLs to ensure stability and latest editions are available, or to ensure stable content is available along with additional features (ex. accompanying spreadsheet)

GPO PURL Statistics

- Count of PURLs created between FY96 through FY20: **257,798**

- Count of PURLs modified or redirected: **520,295**

- The above means that for every PURL created since 1998, it has been updated an average of **2 times**.

Challenges to PURLs

Remember the 2010 PURL server crash?

The PURL server is now hosted, secure, and backed up.

We've done test rollovers with no problems (you all didn't even notice)

When PURLs don't work…

- Agency website goes down (ex. Federal agency shutdown)

- Changing Federal agency web domains (ex. NOAA)

- Agencies redesign their websites (usually for the better) – however, it forces GPO to identify how to best re-route PURLs to the content's new location. Hopefully globally.

- The Internet Archive does routine maintenance on their files during normal workdays. When this happens, their content is not accessible. So if you click on a PURL that routes you to the Internet Archive and you get an error message, wait a day and see if the PURLs are back up and running.

- Agencies may not be following the http protocol proper 404 error code. So a link checker sees a page and thinks everything is okay even though there is no content there.

Interestingly, most problems with PURLs today are not related to the stability of the content, but rather to network security...

- Increasing **network security** on Federal agency websites makes it hard to get to content (ex. DOD)

- If a PURL routes you to an **http** web address, you should be redirected to **https**. The Federal government was directed to migrate to **https** which is a secure connection.

- The content is there, but your library may need to install **security certificates**. If a page says, "this page is unsafe," you can view the content if you know the browser. Otherwise, install the security certificates on your computers so your computers recognize that the site is safe.

Recent issue - DNSSEC protocol

- **Domain Name System Security Extensions** (DNSSEC) is a new protocol that your IT may be adapting.

- Internet Service or Network Providers that have adopted the DNSSEC will not resolve links whose websites are not configured for DNSSEC.
  - * This includes retired domains that are unable to validate and confirm what the security system is requiring.

PURL prefixes that read:	Need to be amended to read:
http://permanent.access.gpo.gov	https://permanent.fdlp.gov
http://FRwebgate.access.gpo.gov	https://govinfo.gov
http://purl.access.gpo.gov	https://purl.fdlp.gov
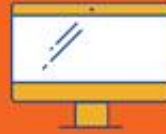
DNSSEC issue

- GPO changed records in CGP, however local library catalog records need to be updated as well.

- CGP on GitHub has a zip file of 170,501 records that were changed.

- News alerts at:
  - https://www.fdlp.gov/news-and-events/4589-purl-access-issues-update
  - https://www.fdlp.gov/news-and-events/4673-cgp-records-with-updated-purls

Dead or lost content? Practices evolve over time

PURLs are designed to route people to the content described in the catalog record. The concept of archiving developed later on.

Until 2005, the 'permanent copy' was the print copy that resided in depository libraries. Beginning in 2005, GPO began archiving a digital backup copy.

Fast forward a couple of decades – if the content is no longer available where it once was, we have to try to find another digital version of it because we do not have it on GPO's server. If we cannot find another version, you get redirected to a PURL History page.

Dead or lost content? Other reasons

- Changing scope determinations per [SOD-PPS-2019-1](#) (Example: some Federal Reserve content)

- Withdrawn at the request of the agency

- Withdrawn at the request of Congress
  - Example: Legislative branch appropriations – removed from **govinfo** as individual annual files and replaced with one Appropriations (omnibus) file – the resource as described in the record, no longer exists.

- Tool has ceased
  - Example: Catalog record for FDsys is 'dead' now because the resource described in the record has been replaced with **govinfo** (which has its own catalog record)
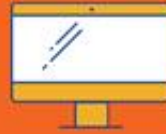
Fun fact:
In Sept., 2020 there were 230,107 records with PURLs in them. (Record count differs than count of PURLs)

561 had content that was no longer available.

Dead content = .0024% of records for digital content

Heads up!

The DLC PURL Working Group will be gathering feedback in the future about the challenges your library has with making digital content available.

Questions?