Online Collection Development:

New PURL Server, PURL Usage Report, & Web Archiving Update

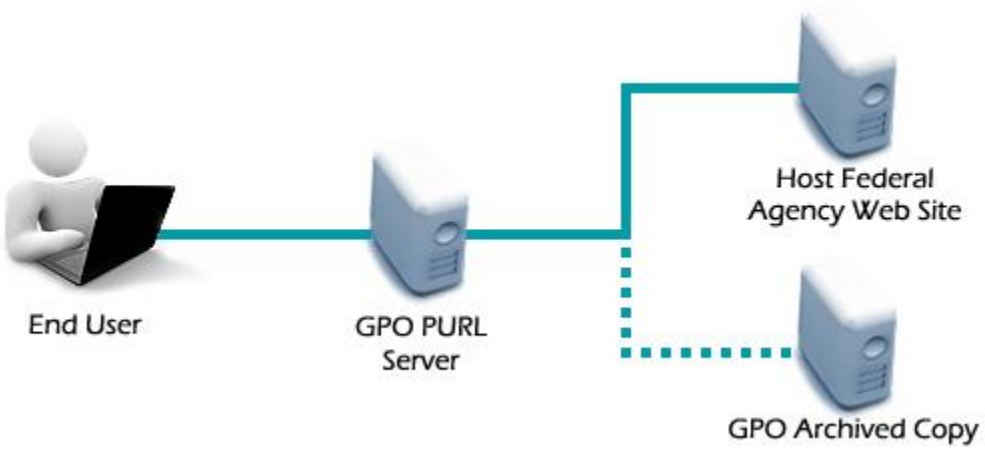
Ashley Dahlen
Dory Bower
October 18, 2016

Agenda

- New PURL Server & PURL Usage Report
- FDLP Web Archive

About PURLs

- Persistent Uniform Resource Locator = http://purl.fdlp.gov/GPO/gpo####
- Began use in 1998
- Currently have
  190,784 PURLs



- GPO routes people who click on the PURL to either agency website or GPO's archived copy on "GPO Permanent" server.

New PURLZ Resolver Software

Improved hosted solution that provides:

- Higher availability for users of PURLs
- Redundancy
- More functionality for the management of the PURLs

Why use PURLs?

Using agency URLs	Using PURLs
You do the maintenance.	GPO does the maintenance.
Can do link usage analysis using click through software	Can do link usage analysis using click through software or GPO's PURL Usage Report tool

PURL Usage Report tool

Analyze:

- What is being used
- When is it being used
- Where your researchers are discovering your PURLs

Document:

- The library's continued need for participation in the FDLP
- Topical areas of interest to your researchers
- Topical areas possibly needing development

Overview of tool – Saving a pattern profile

IP addresses	Domains or hosts
What are your institution's public IP address ranges? • External networks (including WiFi and Ethernet) • Virtual Private Network (VPN) tunnels • Any other proxy/intermediary servers	Library website LibGuides Catalog Database aggregator (like EBSCO) Link service providers (like SFX) WiFi network(s) Building network(s) _____.on.worldcat.org Proxy servers

* Truncation is available

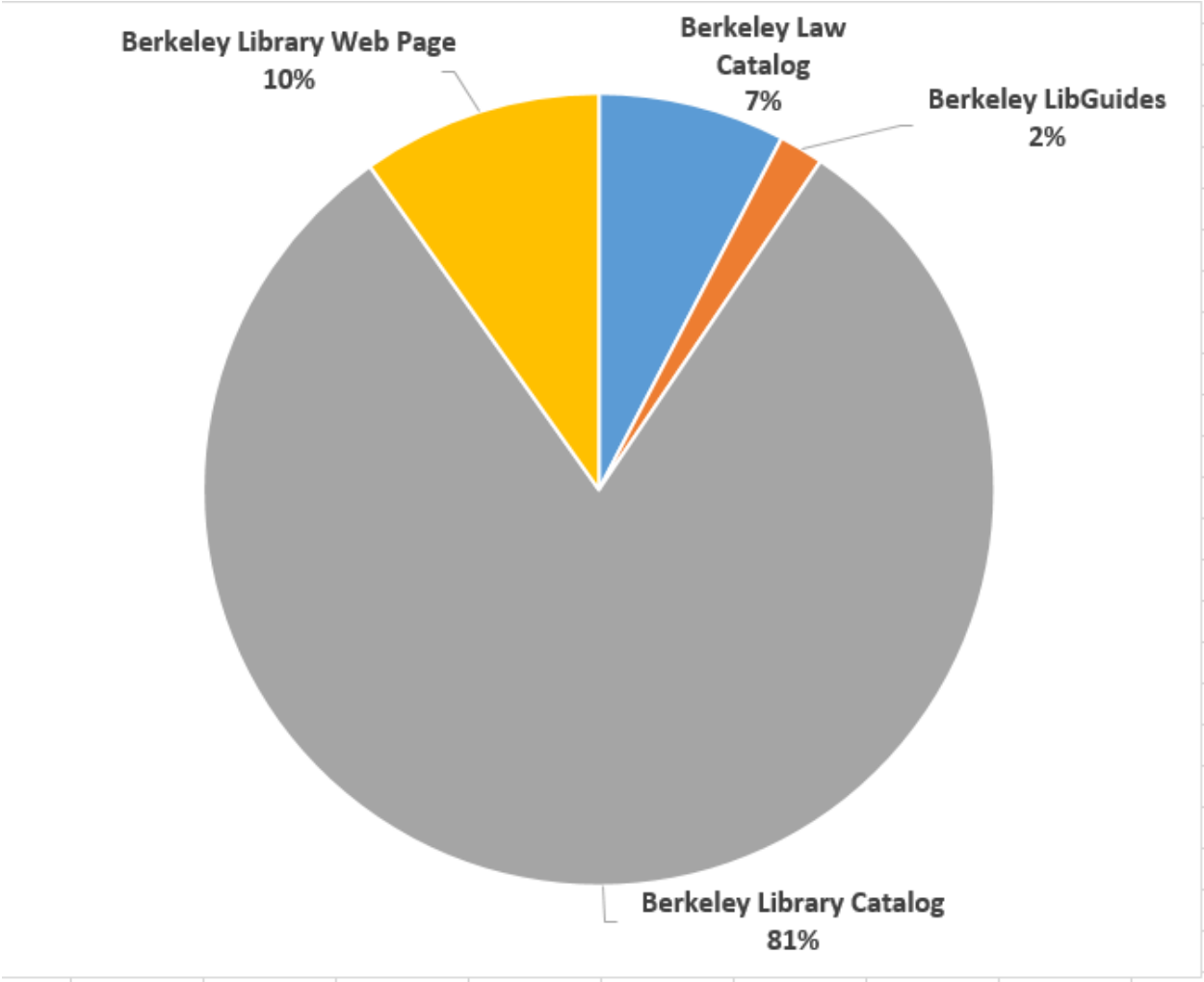
PURL Usage Report

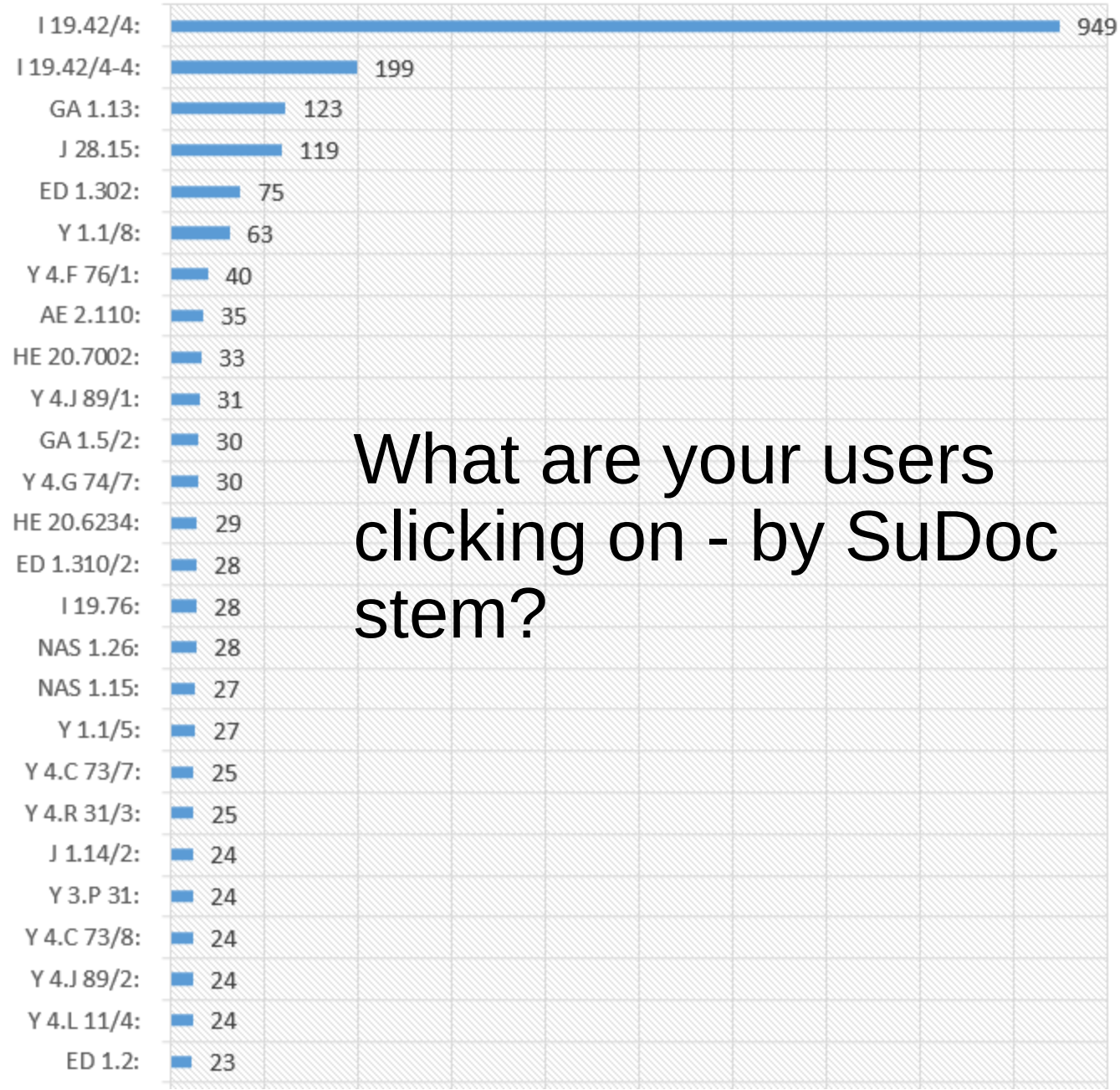
	A	B	C	D	E	F	G	H
1	Timestamp	Host	PURL	Target	SuDoc Number	Title	Author	Year
2	4/1/2016 0:04	c-67-162-1-174.hsd1.	/GPO/LPS16899	http://govinfo.library.unt.edu/r	GP 3.35:PERFREV	Free National Performa	National Performanc	1993
3	4/1/2016 0:05	c-76-28-159-127.hsd:	/GPO/gpo38977	http://permanent.access.gpo.gc	HE 20.6234:69	Nurse practitioners, cer	Park, Melissa,	2011
4	4/1/2016 0:06	c-67-162-1-174.hsd1.	/GPO/LPS387	http://govinfo.library.unt.edu/npr/library/rsreport.html				
5	4/1/2016 0:06	c-67-181-185-176.hs(	/GPO/gpo65226	http://permanent.access.gpo.gc	J 34.5:D 83	Impaired driving (United	States. Department	2009
6	4/1/2016 0:10	c-67-181-185-176.hs(	/GPO/gpo45020	http://permanent.access.gpo.gc	HE 20.7062:D 83/4	Drinking and driving : a	National Center for I	2011
7	4/1/2016 0:12	c-76-28-159-127.hsd:	/GPO/gpo14341	http://permanent.access.gpo.gc	Y 10.9:P 56	Physician extenders, th	United States.Congre	1979
8	4/1/2016 0:17	c-98-235-93-177.hsd:	/GPO/gpo58195	http://permanent.access.gpo.gc	J 36.2:C 86/7	Police programs to prev	Braga, Anthony Allan	2012
9	4/1/2016 0:17	c-98-235-93-177.hsd:	/GPO/gpo58195	http://permanent.access.gpo.gc	J 36.2:C 86/7	Police programs to prev	Braga, Anthony Allan	2012
10	4/1/2016 0:18	c-98-235-93-177.hsd:	/GPO/gpo58123	http://permanent.access.gpo.gc	J 36.2:C 86/3	Police enforcement str	Braga, Anthony Allan	2008
11	4/1/2016 0:37	c-73-41-74-65.hsd1.c	/GPO/gpo19193	http://www.gpo.gov/fdsys/pkg,	Y 4.G 74/9:S.HRG.112-1	Enhancing the Presider	United States.Congre	2012
12	4/1/2016 0:41	c-174-50-68-192.hsd:	/GPO/LPS53454	http://permanent.access.gpo.gc	HE 20.3861:F 73/2/200(	Foot care.		2000
13	4/1/2016 0:41	c-174-50-68-192.hsd:	/GPO/LPS53454	http://permanent.access.gpo.gc	HE 20.3861:F 73/2/200(	Foot care.		2000
14	4/1/2016 0:51	c-73-137-100-26.hsd:	/GPO/gpo48633	http://permanent.access.gpo.gc	A 110.8:D 36/2013	Guide to food defense in	slaughter and proc	2013
15	4/1/2016 0:51	c-73-137-100-26.hsd:	/GPO/gpo51925	http://permanent.access.gpo.gc	A 110.8:D 36/2	Food defense guidelines	for the transportati	2013
16	4/1/2016 0:57	oskicat.berkeley.edi	/GPO/gpo65631	https://www.gpo.gov/fdsys/pkg	Y 4.AP 6/1:AP 6/10/201	Consolidated Appropri	United States.Congre	2016
17	4/1/2016 1:18	c-98-215-68-72.hsd1.	/GPO/gpo36754	https://fraser.stlouisfed.org/tit	C 3.47/2:	Financial statistics of	cities having a populati	1932
18	4/1/2016 3:12	c-73-66-76-170.hsd1.	/GPO/gpo25199	http://www.gao.gov/assets/60(	GA 1.13:GAO-12-491	Homelessness [electroi	United States.Gover	2012
19	4/1/2016 4:34	c-67-175-49-220.hsd:	/GPO/gpo21991	http://permanent.access.gpo.gc	HE 20.7002:AU 8/2/201	Autism spectrum disorders [electronic resou		2010
20	4/1/2016 4:43	c-73-164-51-179.hsd:	/GPO/gpo45836	http://www.gao.gov/assets/66(	GA 1.13:GAO-14-46	Social security death da	United States.Gover	2013
21	4/1/2016 8:11	c-24-125-31-137.hsd:	/GPO/LPS71232	http://permanent.access.gpo.gc	C 60.11:06-439	Resolving interference	Sanders, Frank H.	2006
22	4/1/2016 9:38	c-73-148-65-91.hsd1.	/GPO/gpo21415	http://permanent.access.gpo.gc	HE 20.3852:D 63/2/301	Alzheimer's disease [electronic resource]		2011

Header	Example	Notes
Timestamp	2/1/2016 0:03	MM/DD/YYYY HH:MM:SS Military hours based on Eastern Standard Time
Host	c-73-52-15-242.hsd1.pa.comcast.net 132.194.3.169 vm136.lib.berkeley.edu	You see IP/host name/domain name (never reported out more than once)
PURL	/GPO/gpo15223	The PURL that the user clicked on
Target	http://www.nwtrb.gov/reports/reports.html	Where the PURL ultimately routed the user to
SuDoc Number	Y 4.EN 2:S.HRG.110-228 J 29.45:	SuDoc number, as pulled from CGP
Title	Report to the U.S. Congress and the U.S. Secretary of Energy /	Title, as pulled from the CGP
Author	United States.Congress.Senate.Committee on Veterans' Affairs.	Author, as pulled from the CGP
Year	2004 199u	Year, as pulled from the CGP

Analysis of report…

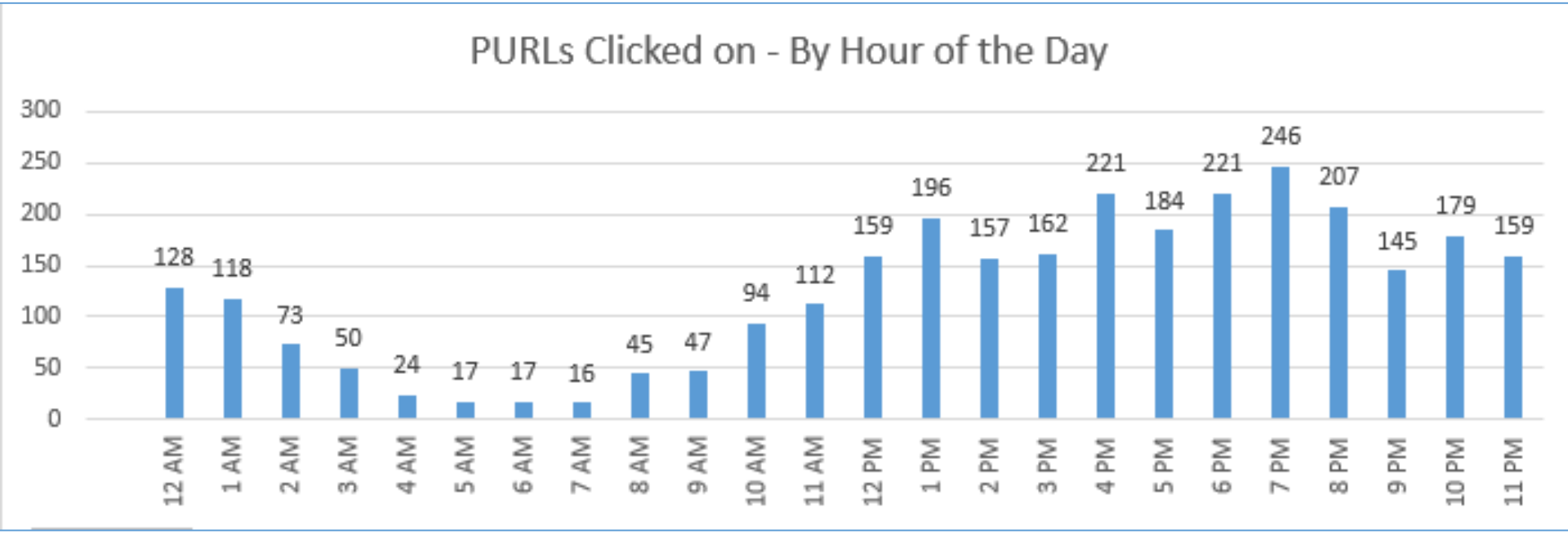
When domain information is provided, where are researchers finding your PURLs?





What are your users clicking on - by SuDoc stem?

When were users clicking on your PURLs?



???

Look at usage over a 3 month period

	A	B	C	D	E
1	SuDoc Stem	12/1/2015	1/1/2016	2/1/2016	Total
2	I 19.42/4:	533	570	949	2052
3	Y 4.L 11/4:	279	17	24	320
4	GA 1.13:	74	58	123	255
5	I 19.42/4-4:	34	10	199	243
6	Y 4.G 74/7:	19	94	30	143
7	Y 1.1/8:	46	33	63	142
8	NAS 1.15:	80	21	27	128
9	Y 4.F 76/1:	44	44	40	128
10	J 28.15:			119	119
11	AE 2.110:	40	35	35	110
12	ED 1.319:	105		3	108
13	Y 4.J 89/1:	29	31	31	91
14	Y 1.1/5:	25	30	27	82
15	ED 1.302:	3	3	75	81
16	EP 1.23/2:	32	40	8	80
17	I 19.76:	23	26	28	77
18	Y 4.R 31/3:	35	14	25	74
19	J 1.14/2:	25	23	24	72
20	Y 3.P 31:	29	18	24	71

Want more information? Detailed instructions?

- Persistent Uniform Resource Locator (PURL): Explanation, Purpose, and Tracking Usage at Your Library (FDLP.gov article)

- Understanding PURL Usage at Your Library...(webinar and webcast about the tool)

GPO's work with online content

FDsys / govinfo

GPO's Permanent Server

FDLP Web Archive

FDSYS FEDERAL DIGITAL SYSTEM
AMERICA'S AUTHENTIC GOVERNMENT INFORMATION

govinfo

FDsys / govinfo

- Content management system
- Users can search the full text of the content, or they can catalog the content.

GPO's Permanent Server

- 'Documents' saved on local server and backed up
- Resources are PURLed and cataloged (CGP)

FDLP Web Archive

- Captures website content in native interface
- Point-In-Time captures or 'snap shots'
- Websites regularly crawled, indexed and searchable on Archive-It, cataloged in CGP

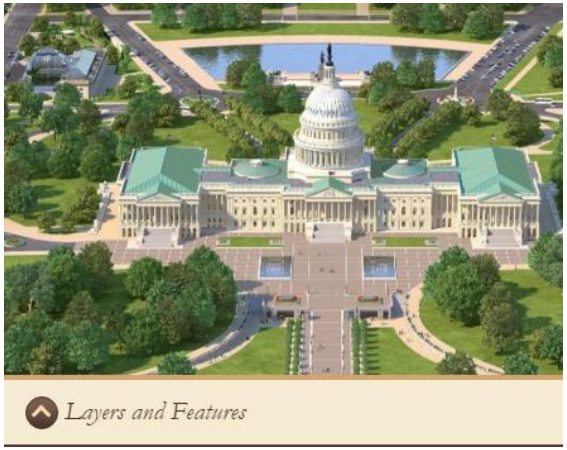
Decision process for online content

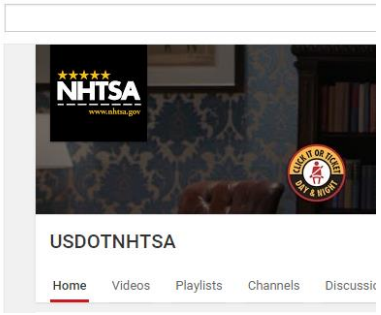
- FDsys/**govinfo**:
  - Used for content with agreement by authoring agency to ingest and preserve
- Permanent Server:
  - Used for individual monographs and serials
- FDLP Web Archive:
  - Used for full websites using Archive-It service
- Partnership:
  - Used for hard-to-harvest sites, databases, or sites with ‘real time’ dynamic information

FDLP Web Archive – why we archive websites

- The web is dynamic and constantly changing.
- New pages appear without notice.
- Changing content on existing pages
- Link Rot = Complete removal of websites without prior notice or changes in URLs that make bookmarked content inaccessible

Types of file formats and sites captured





How GPO captures Federal websites

Take a look at the LSCM Workflow in a 2014 webinar:

Web Archiving for the FDLP

What to capture and what to skip:

- Must be within scope of FDLP
- Avoid duplication of effort with other institutions
- Not distributing through print
- Content less likely to be cataloged
- Avoid duplication of what is in FDsys or archived by other agencies
- SuDoc Y3 sites: commissions, committees, independent agencies
- Non standard government sites or jointly managed sites like www.benefits.gov

GPO does not capture Congressional Committee websites because that is captured in NARA's Congressional and Federal Government Web Harvests that is done at the end of every congressional term

Avoiding Duplication - Collaboration

NATIONAL LIBRARY OF MEDICINE

LIBRARY OF CONGRESS

THE NATIONAL ARCHIVES
ARCHIVES.GOV

Smithsonian

ARCHIVE-IT

SAA Society of American Archivists

IIPC netpreserve.org
INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

Nominate websites for the FDLP Web Archive

Nominations:

- Document Discovery
- AskGPO
- fdlpwebarchiving@gpo.gov

FDLP Web Archive

- 128 collections on Archive-It
- 8.7 TB of data
- Over 64,300,000 urls
- 144 Records in CGP

Access and Discoverability

Catalog of U.S. Government Publications (CGP)
https://catalog.gpo.gov/F

- Basic search page, type **webarch** in search box
  - Can also type specific search terms, example: **webarch and Holocaust**
- Advanced search page, select "**FDLP Web Archive**" catalog
- Identify web-archived collections through "INTERNET" in the SuDoc number

Access and Discoverability

Catalog of U.S. Government Publications (CGP)

- PURL to the "calendar page" of the website's main URL(showing the harvest dates)
- FDLP Web Archive = website level collections harvested and managed by GPO
- PURL to other agency web archive pages
  - Example: National Institute of Nursing Research
    - » PURL: http://purl.fdlp.gov/GPO/gpo37375
    - » CGP system number: 901666

Access and Discoverability

Archive-It Interface

https://archive-it.org/home/FDLPwebarchive

- Search for "**GPO**" or "**FDLP**" from main search box
- A 'collection' is associated with an agency or specific website resource, like Architect of the Capitol or Benefits.gov.
- Search through FDLP Web archive or within a specific collection
  - You are searching all the website text that has been indexed and can search URLs.

Goal = Accessible Government Info

Getting patrons to the Government information they need requires that the librarian uses their knowledge of Government information.

- Identify what Government agency was likely to produce the information that is needed for the correct time period.
- Identify what type of publication the information is likely to be found in.
- Figure out how the library staff can get the publication or resource in the hands of the patron.

Times have changed…

Image sources: https://blogs.loc.gov/picturethis/files/2015/11/jewel-8d02860v.jpg
https://www.loc.gov/rr/program/bib/libsci/readers.jpg

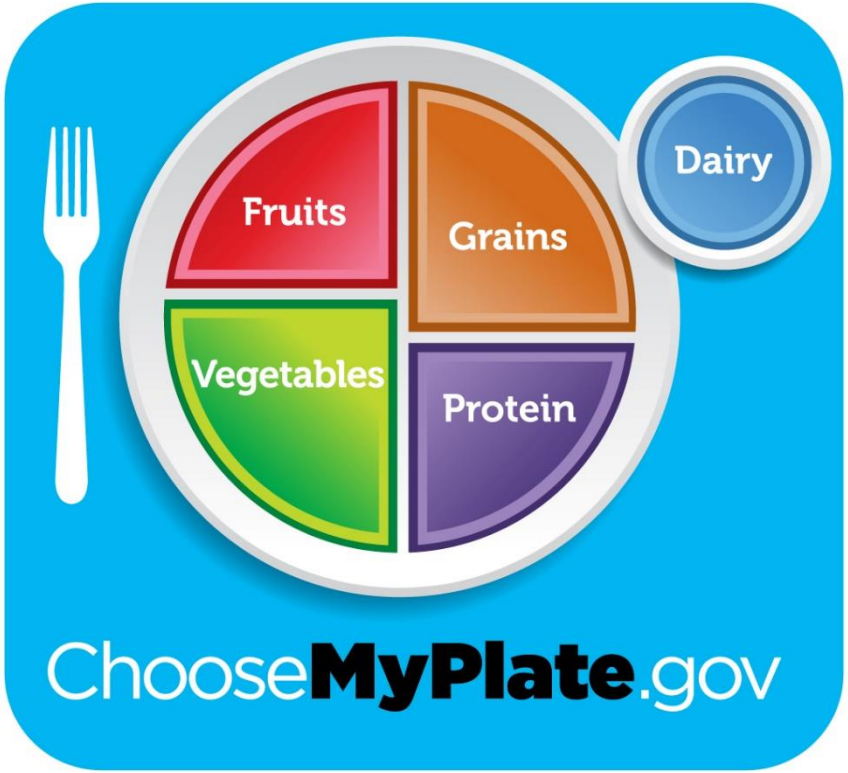
How can libraries use this content?

- When you cannot find what you know used to be on an agency website
- When the Government shuts down or when a website goes down
- When an agency or commission ceases to exist
- When a researcher is studying a point-in-time – See what has been captured - like in the End-Of-Term Harvest
- When a researcher finds a reference to something likely to be on a website, but only for a short period of time (e.g. something time specific like the Olympics)

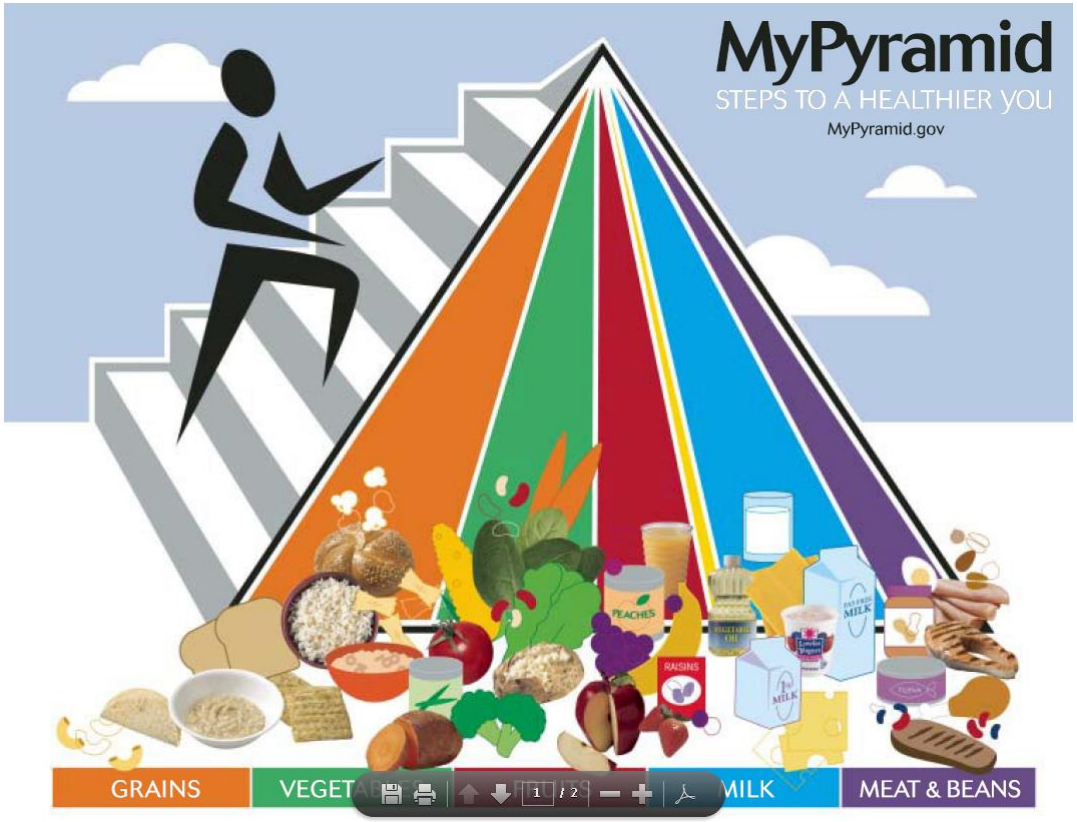
How can libraries use this content?

- When a researcher wants to find all iterations of something or track the evolution of something over time
- When a researcher is trying to gauge how an agency collaborated with the public or engaged in eServices
- Researching how Federal agencies link to or reference the work of other Federal agencies (Tip – Do a search on ‘climate change’)
- When searching for tutorials, quick guides

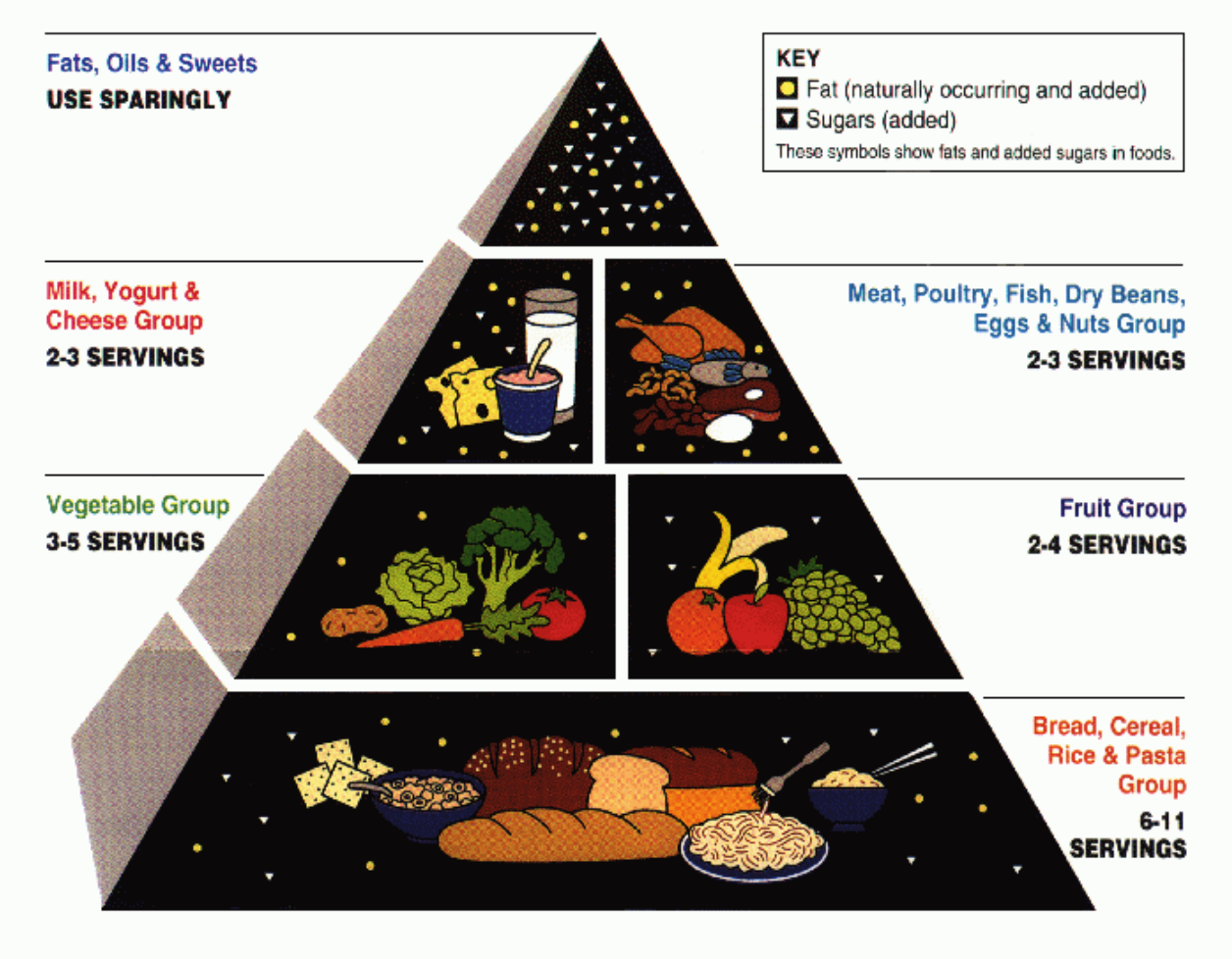
Example one – seen this?



Remember this?



How about this?



And…


For Health...eat some food from each group...every day!
GROUP ONE
GREEN AND YELLOW VEGETABLES...
some raw—some cooked, frozen or canned
GROUP TWO
ORANGES, TOMATOES, GRAPEFRUIT...
or raw cabbage or salad greens
GROUP THREE
POTATOES AND OTHER VEGETABLES AND FRUITS
raw, dried, cooked, frozen or canned
GROUP FOUR
MILK AND MILK PRODUCTS...
fluid, evaporated, dried milk, or cheese
GROUP FIVE
MEAT, POULTRY, FISH, OR EGGS...
or dried beans, peas, nuts, or peanut butter
GROUP SIX
BREAD, FLOUR, AND CEREALS...
Natural whole grain—or enriched or restored
GROUP SEVEN
BUTTER AND FORTIFIED MARGARINE
(with added Vitamin A)
U.S. NEEDS US STRONG
EAT THE BASIC 7 EVERY DAY
U.S. GOVERNMENT CHART
IN ADDITION TO THE BASIC 7...
EAT ANY OTHER FOODS YOU WANT
U.S. DEPARTMENT OF AGRICULTURE

Example two – reference found to traveling exhibit

AUGUST 1936 | GERMANY | SPORTS | BOYCOTT | TO BERLIN | THE OLYMPICS | THE AFTERMATH

Languages: English | 简体中文 | Español | عربي

Home | Site Menu

August 1936

Previous | Next

▲ The last of 3,000 runners who carried the Olympic torch from Olympia, Greece, arrives in the Lustgarten in Berlin to light the Olympic Flame and start the 11th Summer Olympic Games.
—USHMM #21674/Bettmann/CORBIS

For two weeks in August 1936, Adolf Hitler's Nazi dictatorship camouflaged its racist, militaristic character while hosting the Summer Olympics. Minimizing its antisemitic agenda and plans for territorial expansion, the regime exploited the Games to impress many foreign spectators and journalists with an image of a peaceful, tolerant Germany. Having rejected a proposed boycott of the 1936 Olympics, the United States and other western democracies missed the opportunity to take a stand that contemporary observers claimed might have restrained Hitler and bolstered international resistance to Nazi tyranny. After the Olympics, Germany's expansionism and the persecution of Jews and other "enemies of the state" accelerated, culminating in World War II and the Holocaust.

Other interesting examples…

- "Electoral College" yields lots of blog entries on Federal agency websites on the highly anticipated 2012 election.
- "Section 508 tutorial" yields insight into what it takes to make a website accessible. (How will the guidance change over time?)

The User eXperience (UX)

- Archive-It is very responsive to feedback about the tool.
  - They are focused on the content management aspect.
  - User interfaces can be done by the institution.
  - Research methodology is evolving:
    - Hackathons – using indexed web archive content to extract information from a large corpus of content
    - Pulling in content from sources like Facebook

Examples of alternate portals/sites

Columbia University - Human Rights Web Archive:

- Locally-hosted: https://hrwa.cul.columbia.edu/
- Archive-It hosted: https://archive-it.org/collections/1068

Library of Congress:

- Locally-hosted: https://www.loc.gov/websites/collections
- Archive-It hosted: https://archive-it.org/explore?q=library+of+congress

Reference Tip - Explore Other Web Archive Projects

Example: End of Term Harvest

Goal is to capture:

- Agency websites (legislative, executive, judicial)
- Congressional member websites (including social media)

…before the new administration comes in

2008, 2012, & 2016 captures are valuable information about the Federal Government at those time periods.

EOT Harvest

- 2004 – NARA
  - Captured Federal Government websites
- 2008 – LOC, Internet Archive, UNT, California Digital Library, GPO
  - Expanded the scope to include legislative, executive, and judicial branch websites
- 2012 – LOC, Internet Archive, UNT, California Digital Library, GPO, Pratt Institute School of Information and Library Science
  - Even more content to capture given evolving social media
- 2016 – LOC, Internet Archive, UNT, California Digital Library, GPO, Stanford University Libraries, George Washington University Libraries

Check it out: **http://eofarchive.cdlib.org/**

Nominate sites for 2016 crawl:
http://digital2.library.unt.edu/nomination/eth2016/

Questions?