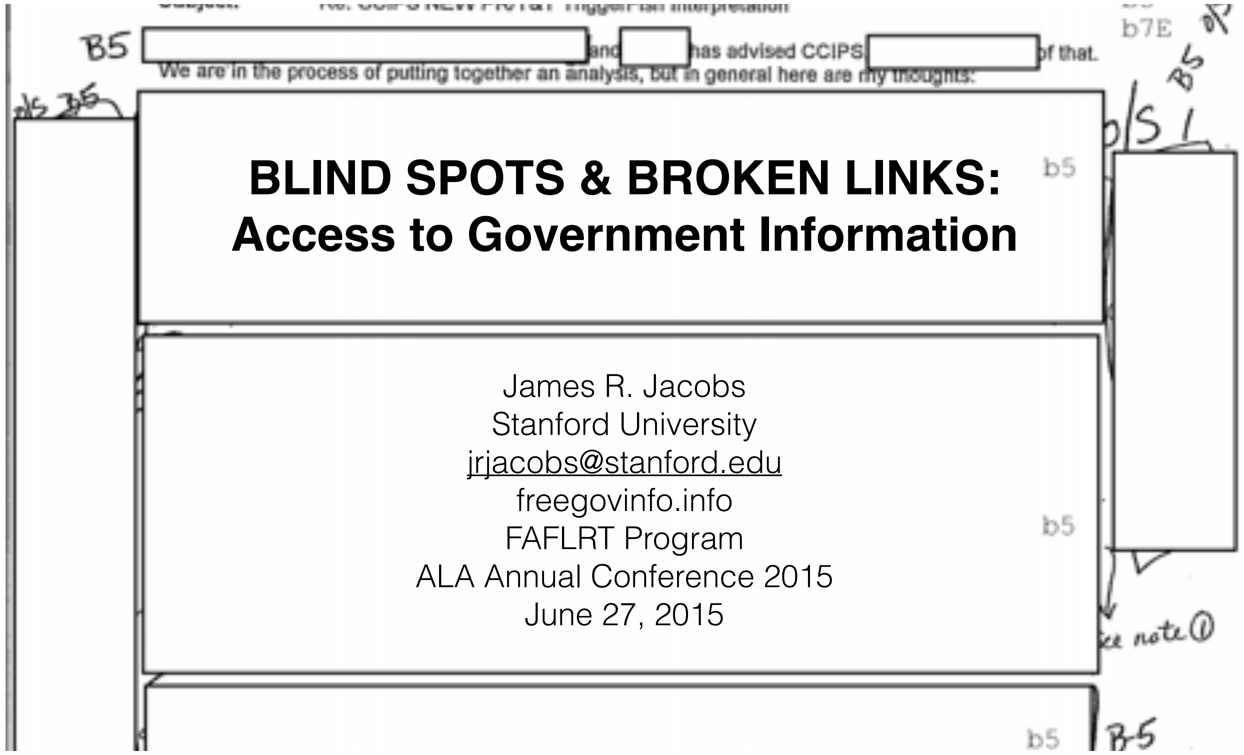


B5
b7E
B5
b5 B5
b5
b5
b5
b5
B5
see note ①

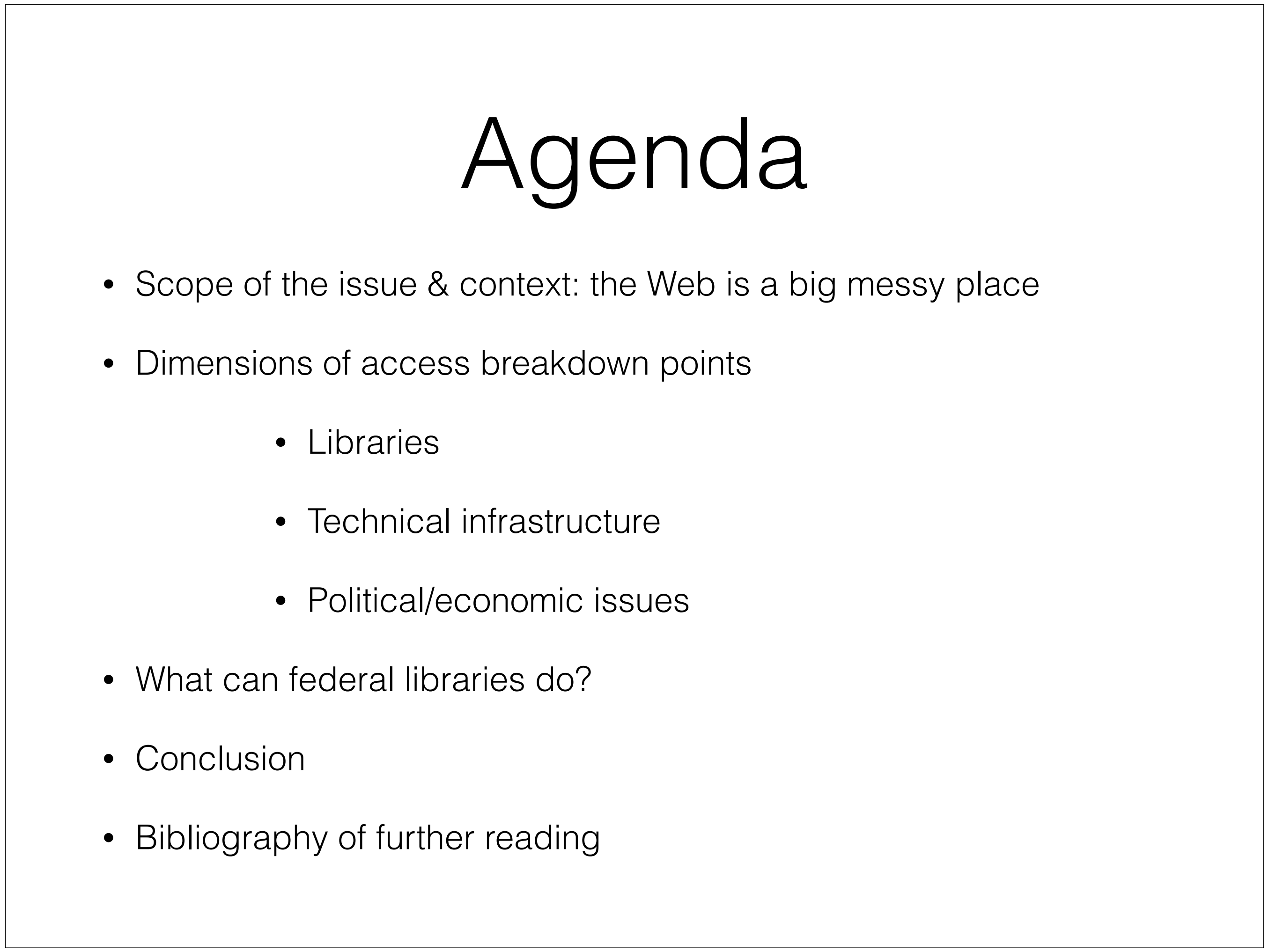
[redacted] and [redacted] has advised CCIPS [redacted] of that. We are in the process of putting together an analysis, but in general here are my thoughts:

BLIND SPOTS & BROKEN LINKS: Access to Government Information

James R. Jacobs
Stanford University
jrjacobs@stanford.edu
freegovinfo.info

FAFLRT Program
ALA Annual Conference 2015
June 27, 2015

*This background is a redacted page of a FOIA'd FBI document re Stingray. MuckRock received 5000 pages of FOIA'd documents which were almost completely redacted!



Agenda

- Scope of the issue & context: the Web is a big messy place
- Dimensions of access breakdown points
    - Libraries
    - Technical infrastructure
    - Political/economic issues
- What can federal libraries do?
- Conclusion
- Bibliography of further reading

Good morning. Thank you for coming. And Thanks Anne for inviting me to talk to you today.

I'm hear to talk about open access to govt information. But really I'll talk about "access" since by definition, federal govt info is in the public domain and therefore "open." IMHO Access is not equal to availability on the web or transparency. Access doesn't just happen serendipitously. Access demands:

systematic preservation
thoughtful curation
human expertise
Understanding political economic context

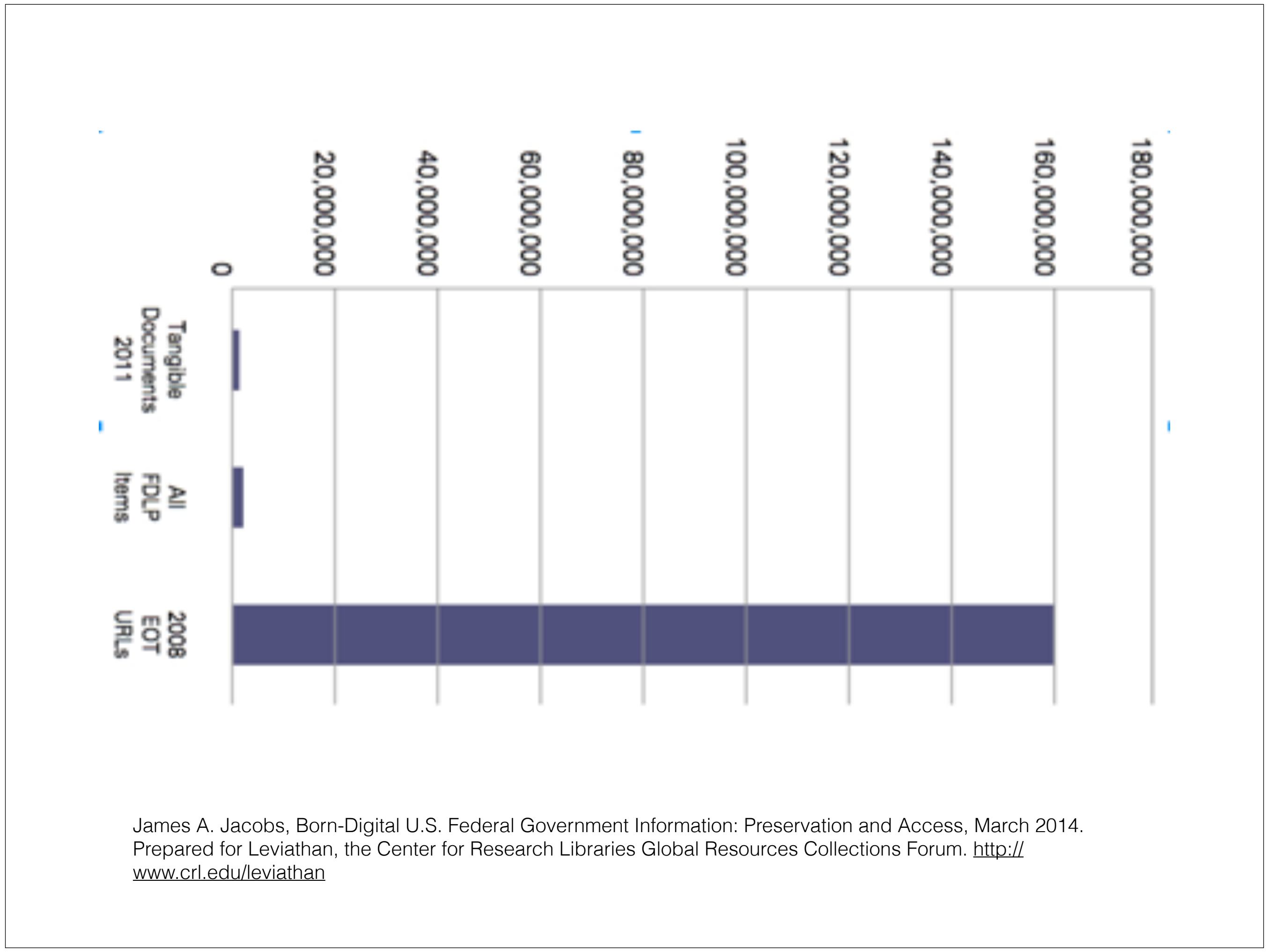
2) in order to more fully understand the complexities inherent in access, I'll talk about some Access Breakdown points as I see them in order to see access more critically and comprehensively. I'm assuming that most if not all of you have a solid working knowledge of FOIA from the trenches of your jobs, so if you don't mind, I'm going to weave FOIA in with access as I see it as another function of access to a particular swath of the govt information sphere. My apologies in advance if you were hoping for a 101 level overview of FOIA.

I'll broadly organize my talk about these breakdown points into libraries as the traditional centers of information access, technical infrastructure, and political/economic issues, but please realize that these 3 are not mutually exclusive but are intertwined and inseparable.

1. Libraries
2. Technical issues
3. Political/economic issues

I'll end with some ideas for what federal libraries can do to help alleviate these access breakdown points.

Scope of the Issue & Context



James A. Jacobs, Born-Digital U.S. Federal Government Information: Preservation and Access, March 2014. Prepared for Leviathan, the Center for Research Libraries Global Resources Collections Forum. http://www.crl.edu/leviathan

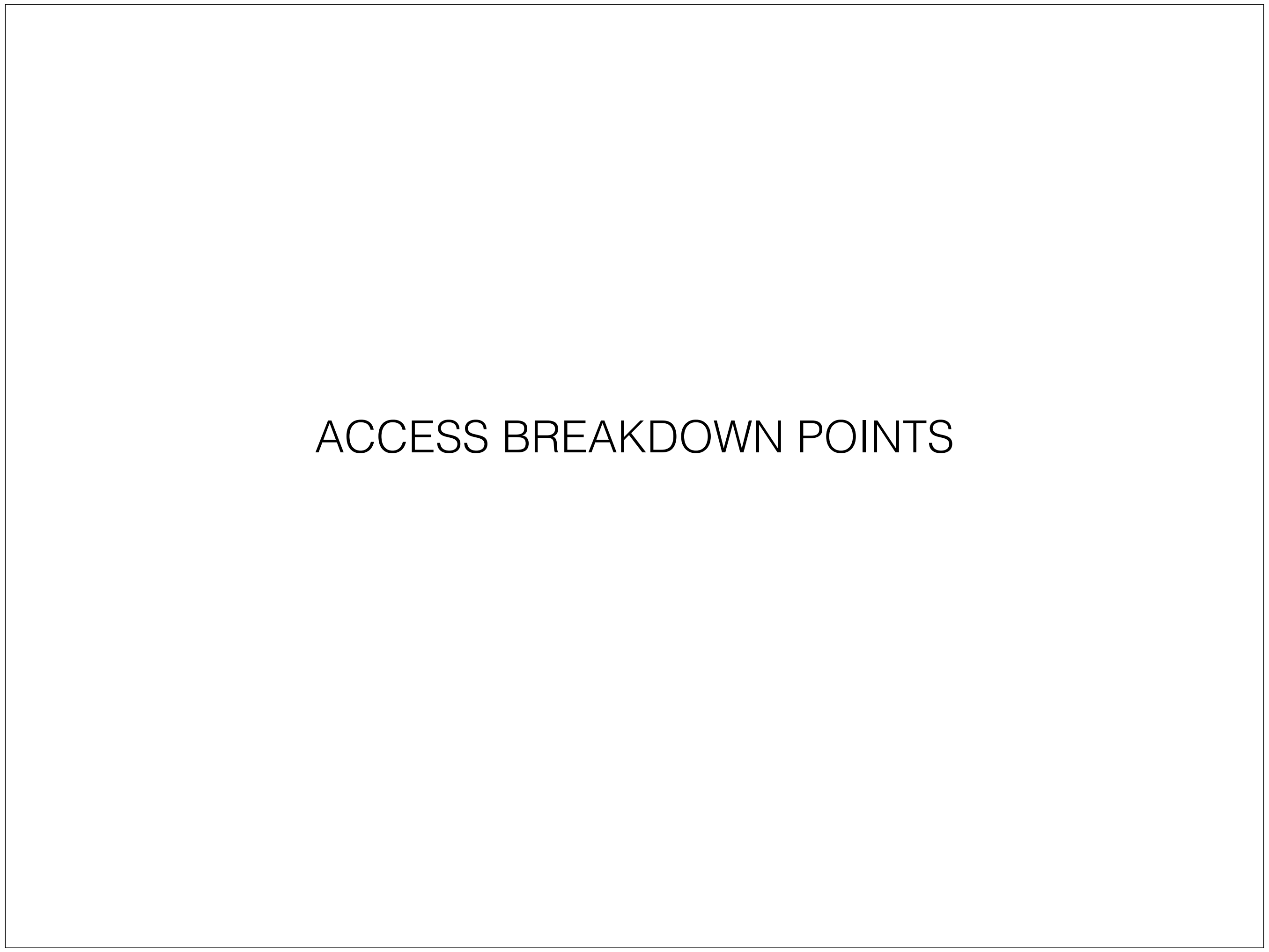
[SLIDE 3: BORN-DIGITAL GRAPH]

II. Scope of the problem

With that said, there has been massive growth in the amount of government information on the Web over the last 25 years. Today, GPO quotes 97% of FDLP materials as being online only, and virtually every agency has an online presence of some sort.

But what does that 97% figure really mean? This is a chart taken from a report my doppleganger JIM Jacobs wrote in March 2014 for CRL entitled "Born-Digital US Federal Govt Information: Preservation and Access." Apologies for having to turn this chart sideways, but it more clearly shows what this 97% figure actually means. It compares the amount of documents distributed by GPO in 2011 (@10,000 items) to the estimated 2.3-3 million total items distributed throughout the FDLP since 1813, to the number of URLs crawled by the 2008 .gov/.mil End of Term Web Crawl (160 million).

even if only 1/1000 of those 160 million URLs are actual documents, that’s still 160,000 documents, data sets and other published content across the .gov/.mil domain, 16 times more than the number of docs distributed to libraries in 2011. Only a very small portion of these ever make their way into the national bibliography or are preserved in any substantative way. Add on top of that govt records that have been released via the FOIA process (either officially hosted in agency FOIA reading rooms, posted by NGOs, or via unofficial leaks from Edward Snowden, Chelsea Manning, Wikileaks etc.) that by and large do not get cataloged. That's a lotta pasta as they say!

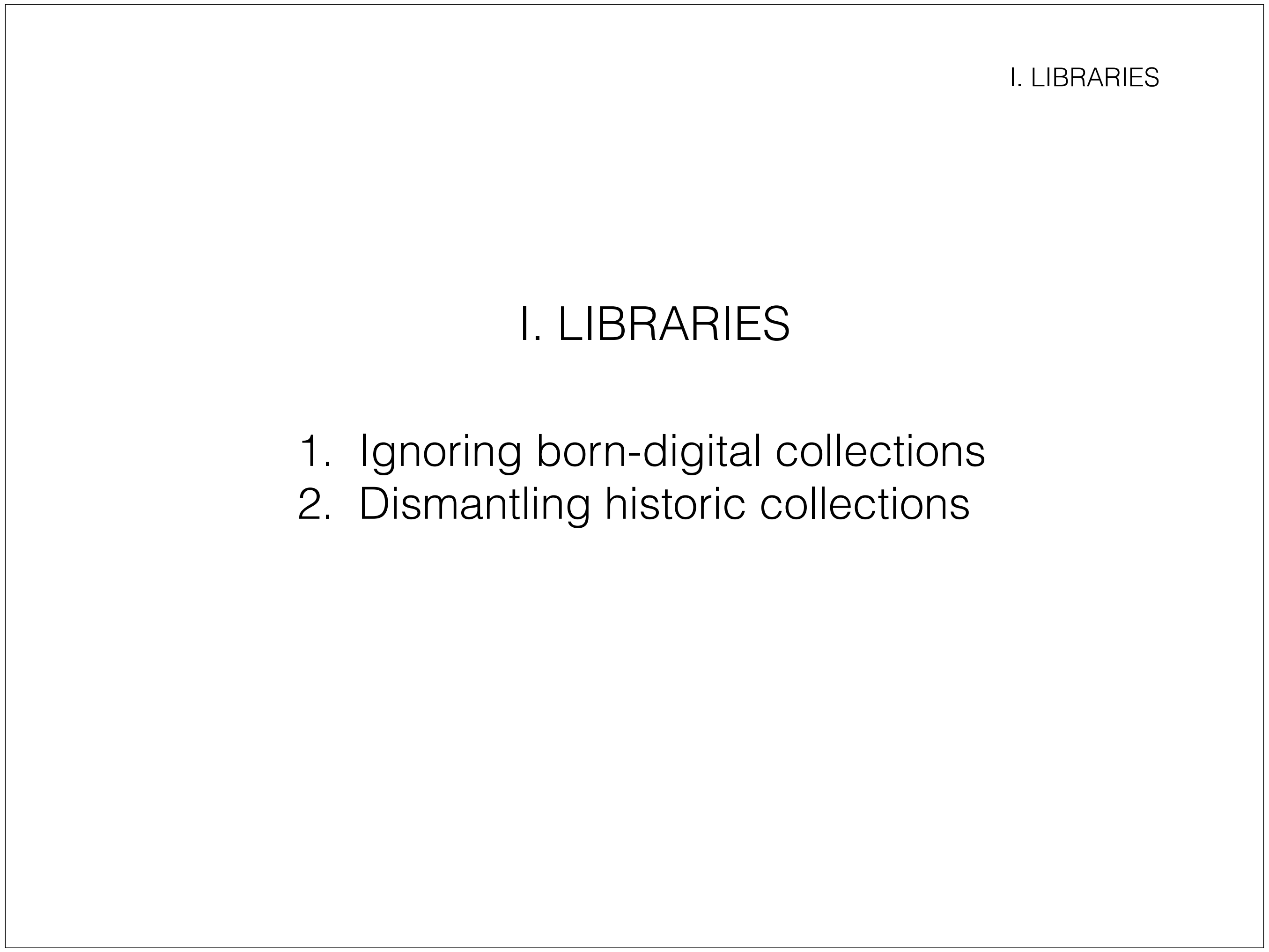


ACCESS BREAKDOWN POINTS

Now that we have an idea of the scope of born-digital govt information, let's delve more closely into the access breakdown points. These points can be technical as well as social or political. As I said, these points are intertwined and can be quite intractable. We're often blinded by the abundance of online information, but it's important to look at the access breakdown points in order to understand and forward the cause of access in a wholistic way.

In 2008, GAO sold exclusive rights to its 20,597 legislative histories of most public laws from 1915-1995 to Thomson West in order to have them digitized. In august, 2009, GPO's PURL server crashed and was not available for several weeks, rendering links to thousands of Federal publications in thousands of library catalogs 404 file not found. In early 2011, the whole country was abuzz with the release of a large cache of mostly un- or non-classified US State Department cables by Wikileaks. In march 2013, NASA took its technical reports server (http://ntrs.nasa.gov/) offline -- and cut access to hundreds of thousands of technical reports -- based on the off-hand comment of a Congressman. In august, 2014, FDLP.gov was hacked (http://freegovinfo.info/node/9014).

Each of these incidents have implications for ongoing and longterm access of both government publications and government records, FDLP materials and FOIA'd information.



I. LIBRARIES

I. LIBRARIES

1. Ignoring born-digital collections
2. Dismantling historic collections

With these examples in mind, let's delve in to the issues a little deeper.

[SLIDE 6 LIBRARIES]

I. LIBRARIES

 * Ignoring Born digital collections
   * many FDLP libraries are no longer building collections, but merely pointing to GPO content, and licensing content from information publishers/vendors (pointing is not collecting). Agency libraries are good but obscure sources of information. This is doubly problematic because FDLP materials are not getting into the national bibliography, and library users are less likely to find FDLP materials. The fugitive document problem is accelerated in the born-digital era.

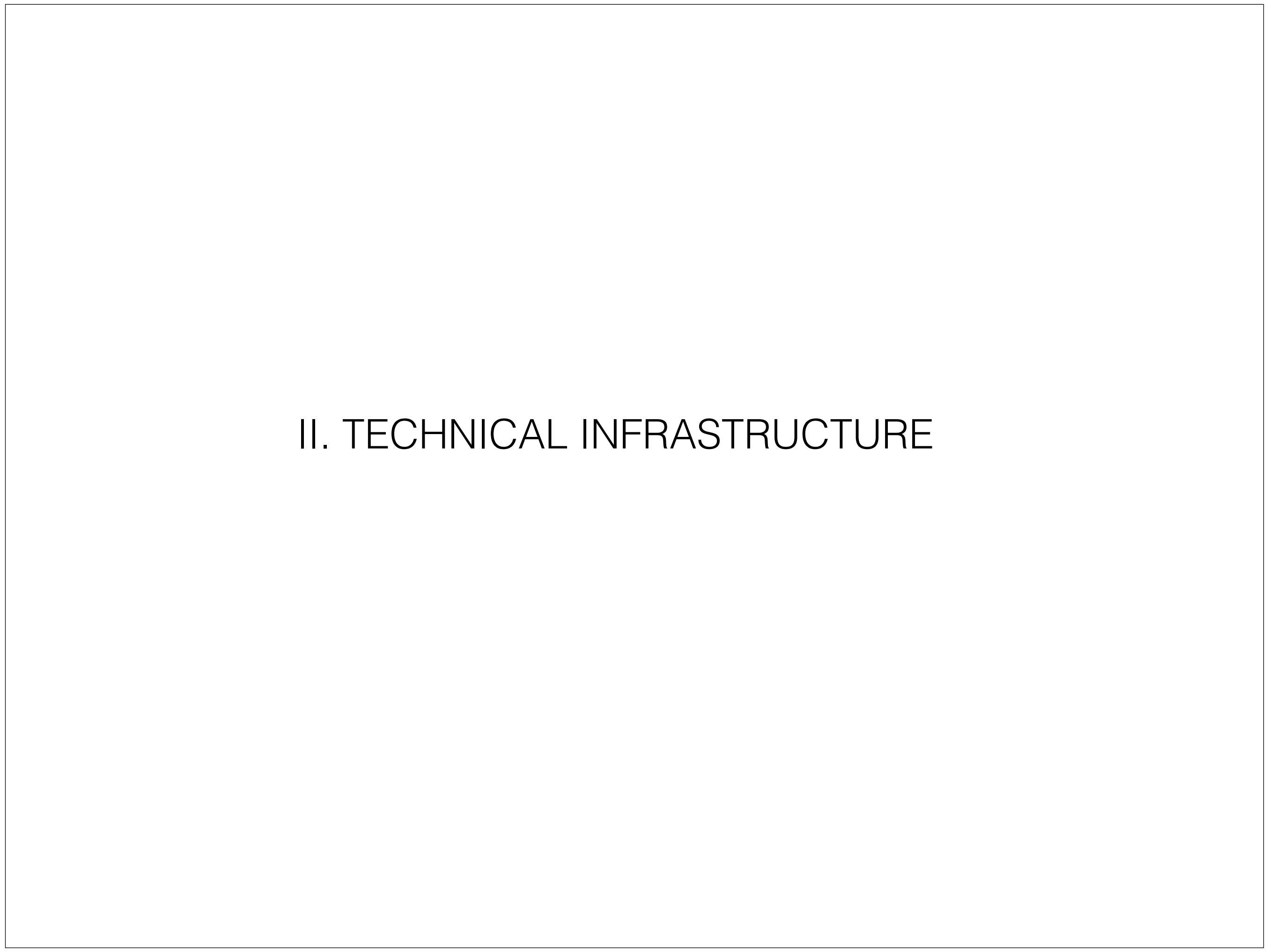
   * Some agencies and their libraries are building digital archives (EPA, DTIC, NOAA, LoC, and those of the national libraries for example) but these databases, aka the dark web, are neither findable nor collectable, and are basically silos of their own materials only, not curated collections across agencies or disciplines. So there's no easy way for a user to find information on a topic that spans agencies.
   *
   * Fugitives: When we depend on pointing instead of collecting http://freegovinfo.info/node/3900

 * At the same time, libraries are Dismantling their historical documents collections. As I've said over and over in many different venues, digitization is not preservation. Many FDLP libraries are heavily weeding their historic collections under the misplaced assumption that most historic documents are or soon will be digitized -- and more importantly under the false assumption that users will only want online access to digitized publications. We know this is a false dichotomy because there is much anecdotal evidence of users requesting paper documents after finding digitized versions online. Reasons range from missing and/or poorly scanned pages (lots of them!), large documents not easily read online, redigitization for special purposes (scanning and keying of tabular data, corpus analysis of documents scanned but not OCR'd etc).

We have to remember:

pointing is not collecting
access is not preservation
Federal and FDLP libraries still need to work at collection development collaboratively in order to provide *long-term* access to govt information.

next let's talk about



II. TECHNICAL INFRASTRUCTURE

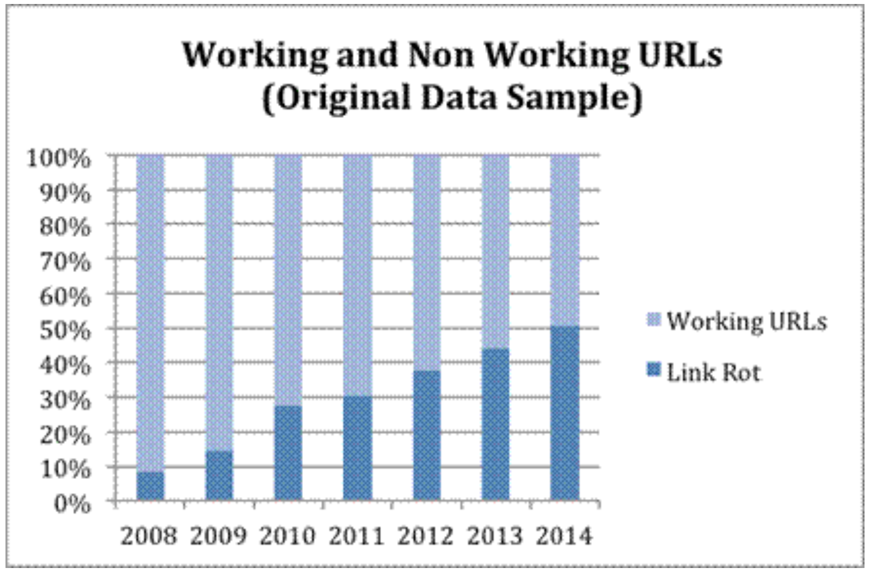
[SLIDE 7: TECHNICAL INFRASTRUCTURE]

II. TECHNICAL INFRASTRUCTURE

When we search Google, we always get results. But people generally don't think about what's missing from their search results. There's a dark side of the Web called "link rot" and it's parallel but lesser known twin "content drift." Access today does not equal permanent public access. There is largely an absence of a long-term preservation system for born-digital govt information.

II. TECHNICAL INFRASTRUCTURE

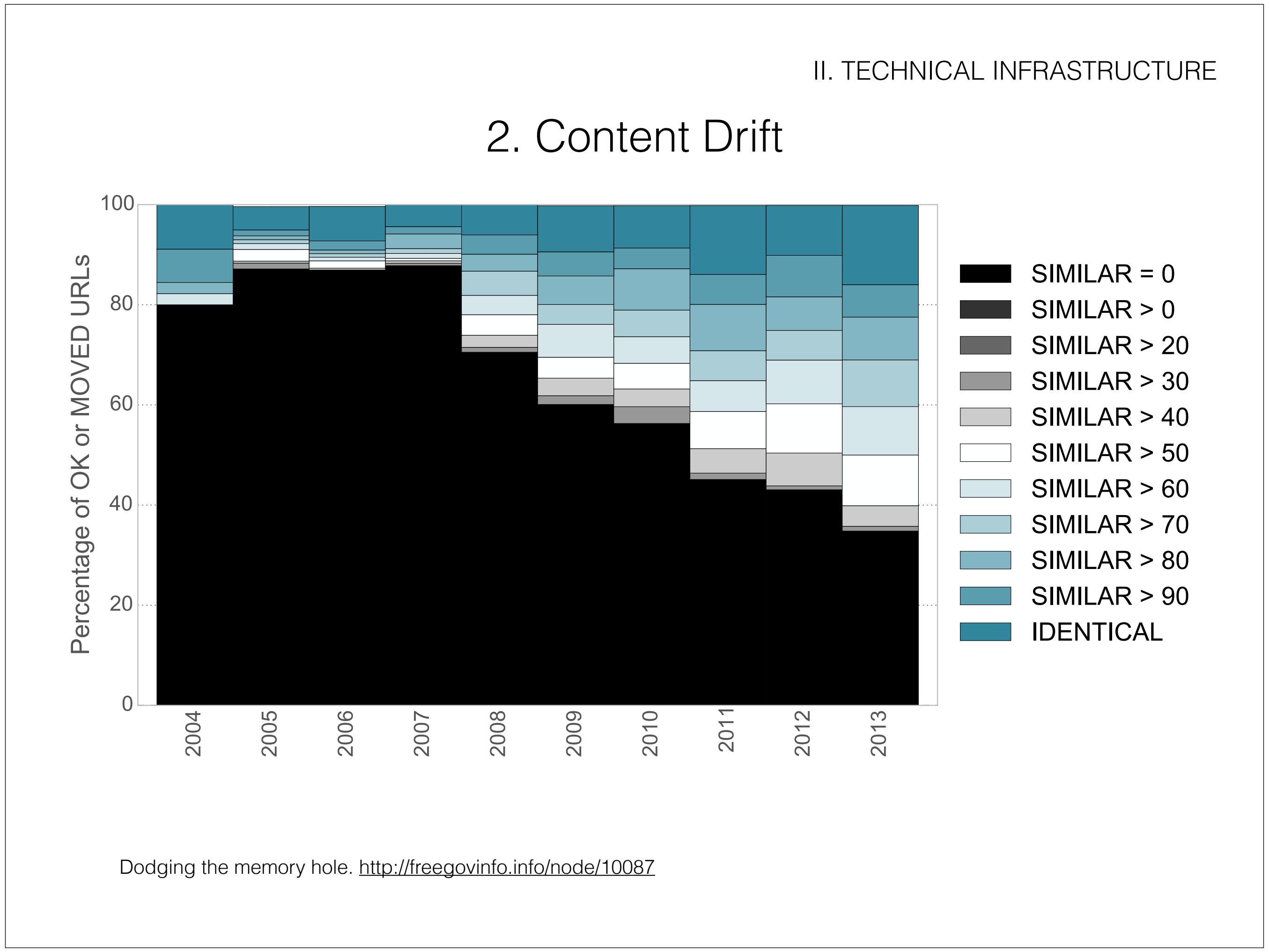
1. Link Rot



2014 Link Rot Report, Chesapeake Digital Preservation Group. http://bit.ly/2014-link-rot-report

[SLIDE ON LINK ROT]

While the average lifespan of a physical government document is 50 years, according to the Internet Archive, the estimated lifespan of a URL is 44 - 75 days. While .gov Web sites are slightly more stable than some Web sites, “link rot,” the process by which Internet hyperlinks disappear, is a real and growing concern. According to data collected by the Chesapeake Digital Preservation Group, which has been studying link rot of the .gov domain since 2008, 51% of the urls from their original 2008 data set were 404! Linking does not equal collecting.



II. TECHNICAL INFRASTRUCTURE

2. Content Drift

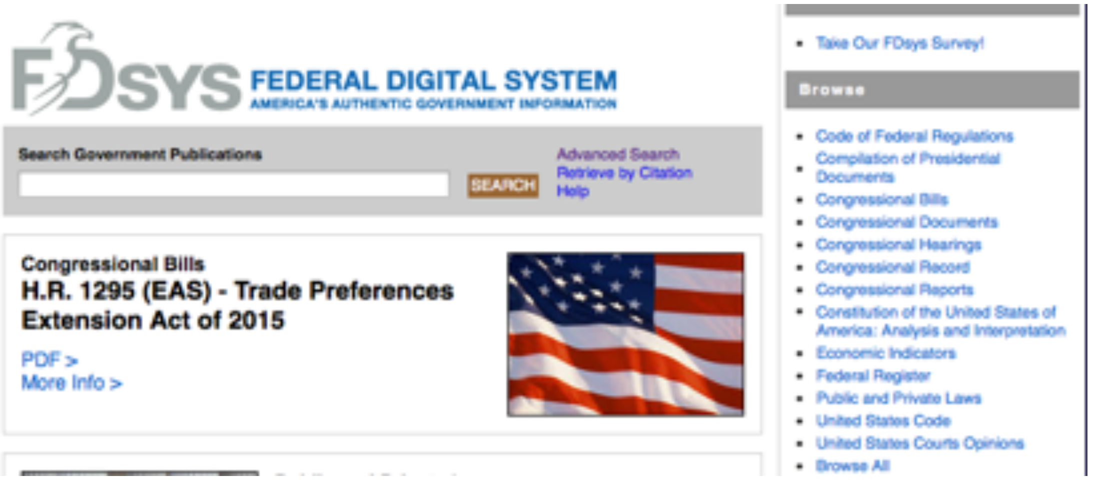
Dodging the memory hole. http://freegovinfo.info/node/10087

[SLIDE ON CONTENT DRIFT]

"content drift" is a term used in the Web archiving community. It is the process whereby an archived file has changed since being archived. Andy Jackson from the UK Web Archive at the British Library gave a presentation in April, 2015 at the International Internet Preservation Consortium, General Assembly 2015. In trying to answer the question “How much of the content of the UK Web Archive collection is still on the live web?” his research showed that 50 percent of content had gone, moved, or changed so as to be unrecognizable in only one year. After three years the figure rose to 65 percent.

In short order, born-digital content on the Web either changes or disappears in days or months, not years or decades.

3. Lack of Preservation Infrastructure



FDSYS FEDERAL DIGITAL SYSTEM
AMERICA'S AUTHENTIC GOVERNMENT INFORMATION

Search Government Publications

SEARCH

Advanced Search
Retrieve by Citation
Help

Congressional Bills
H.R. 1295 (EAS) - Trade Preferences Extension Act of 2015

PDF >
More Info >

- Take Our FDsys Survey!

Browse

- Code of Federal Regulations
- Compilation of Presidential Documents
- Congressional Bills
- Congressional Documents
- Congressional Hearings
- Congressional Record
- Congressional Reports
- Constitution of the United States of America: Analysis and Interpretation
- Economic Indicators
- Federal Register
- Public and Private Laws
- United States Code
- United States Courts Opinions
- Browse All

Federal Digital System (FDsys) http://fdsys.gov

[SLIDE: LACK OF PRESERVATION INFRASTRUCTURE]

Another infrastructure access breakdown point is centered around GPO and executive agencies themselves as the producers of information in all its guises.

* GPO is currently working on an internal TDR audit for FDsys, but GPO does not have an adequate preservation program in place for FDsys which includes a succession plan -- a requirement of the OAIS standard. GODORT and the FDLP community has been asking since at least the early 2000s for GPO to create a mirror of their content, but nothing is in place currently. LOCKSS-USDOCS program has stepped into that breach and is harvesting and collaboratively preserving all FDsys content, but there is no official MOU in place between GPO and LOCKSS -- that's partially on me since I'm the program lead for LOCKSS-USDOCS.

* Meanwhile permanent.access.gpo.gov -- the GPO junk drawer where they store many of their digital documents not in FDsys -- is running on rickety old servers and blocking crawlers via robots.txt, preventing the Internet Archive and others in the Web archiving community from harvesting and preserving those materials.

* Exec agencies are by and large not doing anything in the digital preservation space that I'm aware of (though I'd be happy to be proven otherwise!). Most are going it alone on the Web, creating 100s of single points of failure, and only a few are working w GPO to host their documents on FDsys (hat tip goes out to GAO, Treasury, NIST I recently learned is working toward an agreement w GPO).

III. POLITICAL/ECONOMIC ISSUES

III. POLITICAL/ECONOMIC ISSUES

SECRECY
WE NEED MORE SHOVELS!
FOIA
FOIA
WE NEED MORE SUNSHINE!
CLASSIFIED
M. WUERKER
POLITICO

Cartoon by Matt Wuerker. From Better Government Assn. Fair use claimed.

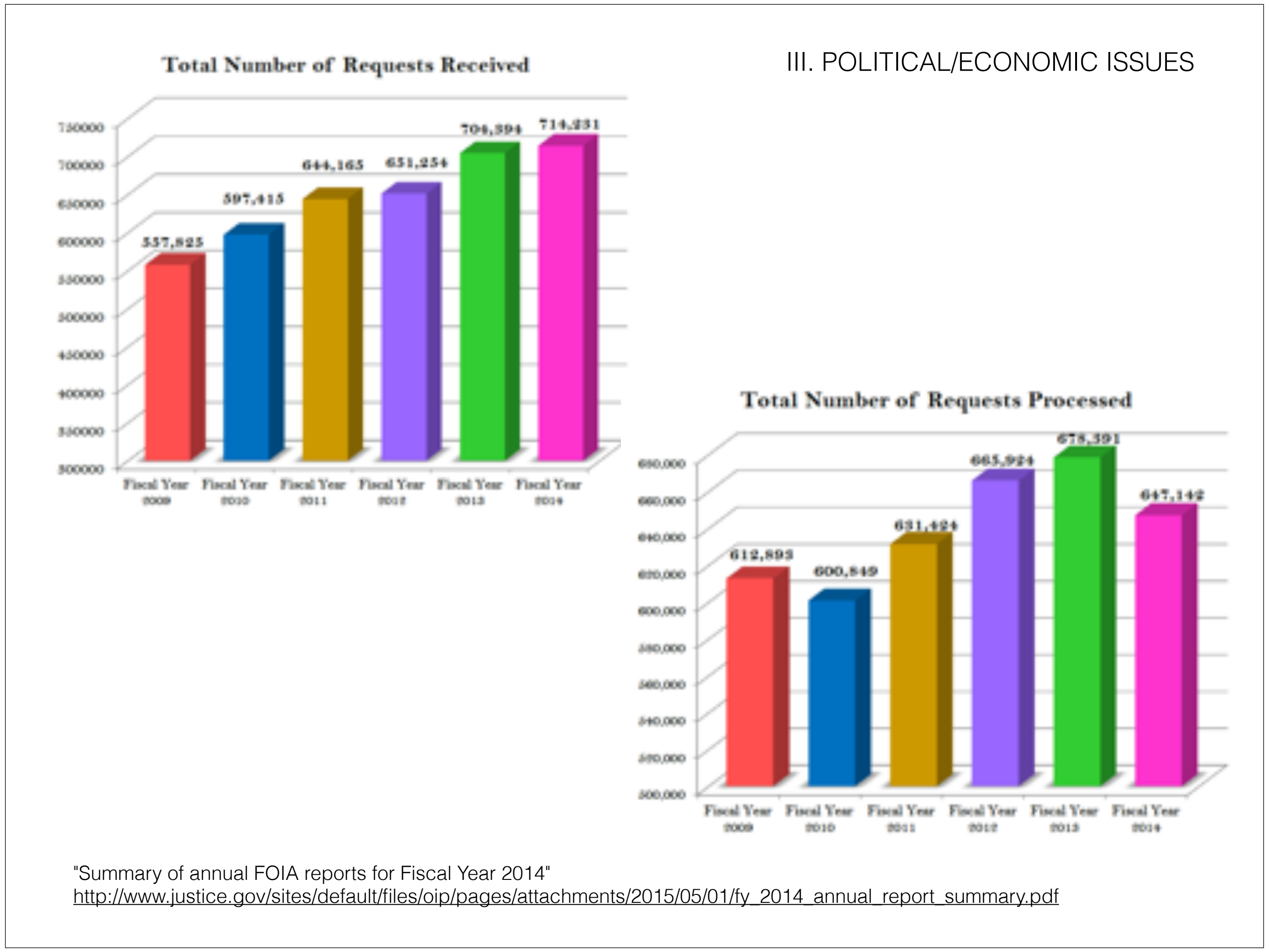
[SLIDE: SECRECY]

III. POLITICAL/ECONOMIC ISSUES

Government information is political by nature. At the same time, given the current political/economic context, govt information has turned increasingly into a valuable commodity. I don't have to tell you that federal libraries are under enormous budgetary contraints, while commercialization and privatization of govt information has accelerated.

Let's turn our attention briefly to FOIA. The first FOIA law came into effect in July 1967, and was designed to give better public access to the internal workings of executive agencies. However, this did not completely solve the access to govt records problem. There's long been documentation about public policy breakdowns in access.

FOIA'd information, once it's been released(!), is now more widely available through agency FOIA reading rooms and NGOs, and a growing number of organizations like muckrock and NSA Archives are helping to educate and assist the public in making FOIA requests.



III. POLITICAL/ECONOMIC ISSUES

"Summary of annual FOIA reports for Fiscal Year 2014"
http://www.justice.gov/sites/default/files/oip/pages/attachments/2015/05/01/fy_2014_annual_report_summary.pdf

[SLIDE: FOIA REQUESTS VS PROCESSED]

According to the dept of justice's "Summary of annual FOIA reports for Fiscal Year 2014" http://www.justice.gov/sites/default/files/oip/pages/attachments/2015/05/01/fy_2014_annual_report_summary.pdf The number of FOIA requests received over the last 5 years has grown exponentially to 714,000 requests for FY2014. While it looks from the statistics that fulfilment of requests is generally on par with # of requests, and though most agencies process requests in a timely fashion, the FOIA backlog has almost doubled in the last year, and there are certain agencies (I'm looking at you FBI!) that are not complying w President Obama's open govt initiative or with FOIA regulations, routinely have backlogs of years and force scholars like prolific FOIA submitter Ryan Shapiro from MIT -- who's thesis examines how the FBI monitors and investigates protesters -- to sue the FBI multiple times for non-compliance. There are many stories out there like Shapiro's which point to the politically problematic nature of FOIA. There's also the twin issues of overclassification and "controlled unclassified information" that needlessly obfuscate FOIA access to the point that only the most dedicated and glacially patient FOIA requesters are able to understand and successfully navigate the byzantine agency cultures in order to use the FOIA system.

* # of FOIA requests received vs # of requests processed (charts at http://www.justice.gov/sites/default/files/oip/pages/attachments/2015/05/01/fy_2014_annual_report_summary.pdf)
* overclassification and "controlled unclassified information"
* http://fas.org/blogs/secrecy/2015/05/cui-is-coming/
* eg of FBI Stingray FOIA request where FBI released 5000 blank pages http://freegovinfo.info/node/9968 (Images of redacted documents :http://www.frugal-cafe.com/public_html/frugal-blog/frugal-cafe-blogzone/wp-content/uploads/2014/03/redact-obama-cartoon.jpg)

> What was first seen as an emerging trend in April 1981 when the American Library Association Washington Office first started this chronology of items which came to our attention, had by December 1987 become a continuing pattern of federal government to restrict government publications and information dissemination activities. A policy has emerged which is less than sympathetic to the principles of freedom of access to information as librarians advocate them. A combination of specific policy decisions, the Reagan Administration's interpretations and implementation of the 1980 Paperwork Reduction Act (PL 96-511, as amended by PL 99-500), implementation of the Grace Commission recommendations, and agency budget cuts have significantly limited access to public documents and statistics.
>
> –*Less Access to Less information by and about the US government*
> http://freegovinfo.info/less_access

[SLIDE: LESS ACCESS QUOTE]

Acceleration of privatization of government functions and commercialization of government information in the 1980s and 1990s through to today has unintended consequences for access to govt information.

From 1981 to 1998, Anne Heanue and the fine folks at the ALA Washington Office published an amazing series called "Less Access to Less Information by and about the U.S. Government", a chronology of efforts to restrict and privatize government information.

While that publication has ceased (but is accessible on the Freegovinfo website!) There are 3 growing concerns over the last 20 years or so:

1: a growing number of agencies, in an attempt to cut staff -- but ironically NOT to save money! -- are outsourcing parts of the FOIA process. Since 2009, the government has awarded at least 250 FOIA related contracts, and in most cases contractors now outnumber government employees three to one. There's been a 40 percent jump in FOIA contracts since President Obama came into office (https://nsarchive.wordpress.com/2012/11/06/foia-for-profit/)

2. The public todo about Hilary Clinton's private email server while serving as Secretary of State brought to light another issue. More and more political officials are doing the government's business via non-governmental email and other communication networks, thus skirting FOIA laws since FOIA does not apply to private entities.

III. POLITICAL/ECONOMIC ISSUES

Image from trackaid.wordpress.com. Fair use claimed.

[SLIDE: PROFITABILITY]

3. Lastly, Commercialization of govt information is running rampant. Instead of promoting free access, libraries are increasingly relying on commercial vendors -- at least the ones that can afford to pay for access to commercial databases like Lexis Nexis, West Law, Proquest and neophites like Voxgov. Further, many private companies are offering agencies "no cost" contracts to digitize and privatize govt information under the guise of "efficiency." The Thomson-West GAO project is but one particularly egregious example of this pernicious problem.

-
- Big data: Weather, climate data, satellite data etc.

ACCESS REQUIRES:

- Systematic preservation
- Thoughtful curation
- Human expertise
- Understanding political economic context

[SLIDE: ACCESS REQUIRES]

So it seems on the surface that the public has amazing and unprecedented access to FDLP materials. But I hope my brief examination of some access breakdown points has convinced you that that's not necessarily true and that access requires:

Systematic preservation
Thoughtful curation
Human expertise
Understanding political economic context

WHAT CAN/SHOULD LIBRARIES DO?

- Archive-ready sites [http://archiveready.com]
- Well-structured, accessible, harvestable sites
- Collaborate with GPO for preservation
- Share metadata with all and sundry
- Work with your agency CIO's so that they understand that access demands action!

[SLIDE: WHAT CAN/SHOULD LD LIBRARIES DO?]

Given this technical, political-economic and social environment, what can federal libraries do? what should libraries do? I believe that Federal libraries and librarians are critical to the continued access to federal govt information in all its guises and can positively affect both technical infrastructures and information policy within their agencies. Here are some ways to do that:

1. is your site archive-ready? Is it well-structured and standards-based? Has your site been crawled by the end-of-term crawls? is your site in the WayBack Machine in its entirety? (hint: if you've got content only accessible via search engine or in the dark Web, then it's most likely NOT in the WayBack machine and not well preserved).
2. Do your library and agency sites have site maps with ../publications and ../data directories prominently displayed? (and how about /foia? /xml or /bulk? /feed? /social? /video or /A-V? (or maybe these are sub-dirs under /pubs and /data?) This is the best way to assure that Web archivists can harvest and preserve your data.

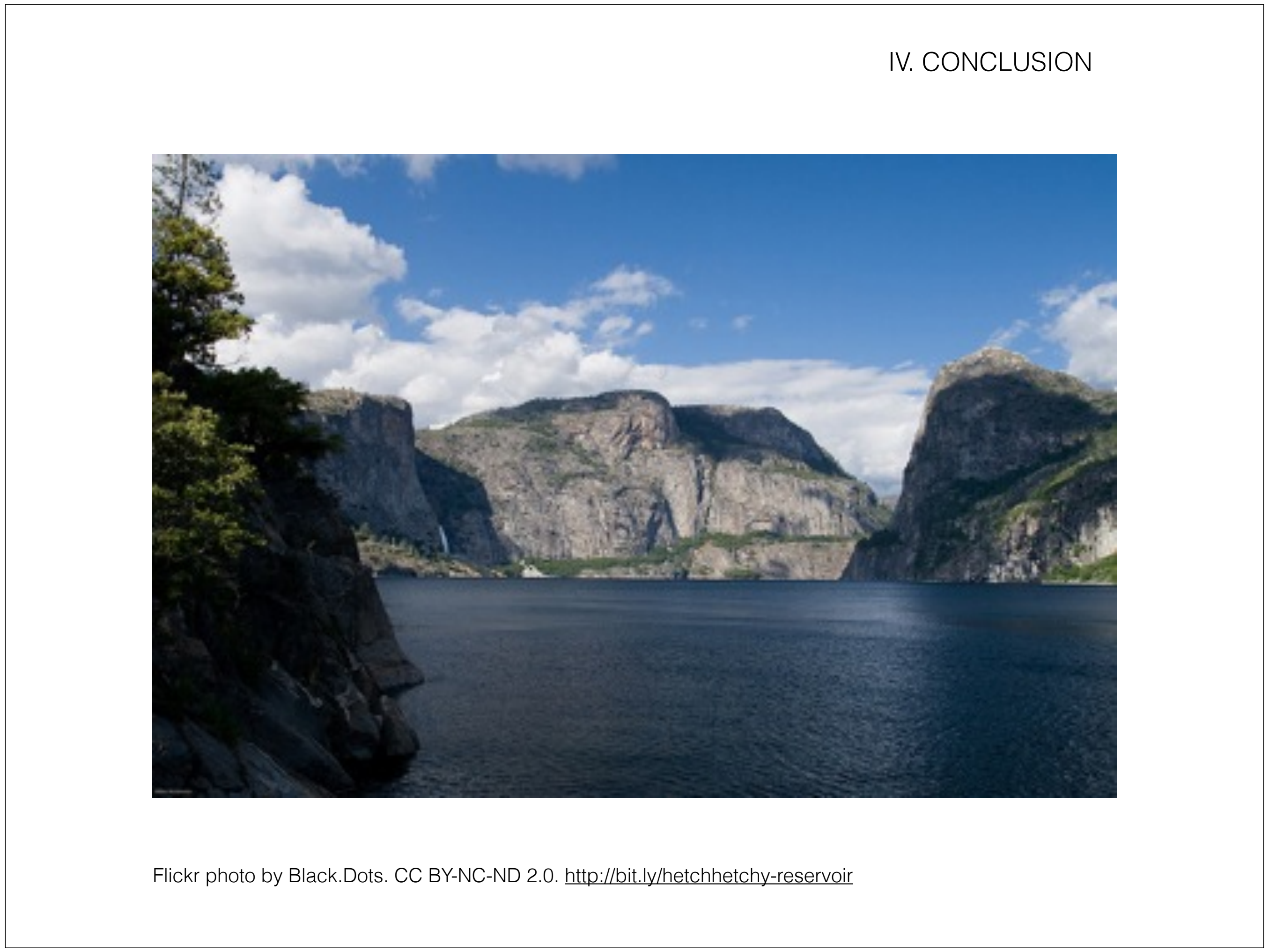
3. work with GPO to submit digitized and born-digital agency publications to GPO for inclusion into the National Bibliography and FDsys -- which will assure preservation via LOCKSS-USDOCS! Bottom line is that all agencies need to assure redundant access and preservation off of .gov servers. This will help to alleviate the fugitive problem.

4. Share metadata and content with libraries via OAI-PMH or other methods (OSTI's batch MARC record downloads!). The more catalogs contain records for agency publications, the more accessible those publications will be.

5. Make contacts with FDLP libraries and work with them on collections, preservation and public access.

6. Work with your agency CIO's on agency information policy so that they understand that access demands curation!

Federal libraries and the agencies they serve can better assure access -- after all, preservation is simply access tomorrow! -- if their sites are well-designed and standards-based, if they manage their content under the assumption that it will be crawled, collected, and preserved and facilitate those processes, and if they work with GPO to make sure their publications, data and statistics are hosted in FDsys and described in the national bibliography.



IV. CONCLUSION

Flickr photo by Black.Dots. CC BY-NC-ND 2.0. http://bit.ly/hetchhetchy-reservoir

CONCLUSION:

[SLIDE: RESERVOIRS]

What makes govinfo important (regardless of source or status or currency or even accuracy) is that they are records of government. They tell what government knew, or thought it knew, or wanted us to know, or wanted us to think we knew, etc. at any given point in time. They are *inherently* important because they record actions and workings and "knowledge" of govt. We cannot be self-governing without a full understanding of what those in govt do, think, say.

Each library must act as part of an ecosystem, not as a stand-alone entity with no responsibilities to others. We all need to act locally, think globally AND think locally, act globally. We should each be asking ourselves what part can I play to make the ecosystem stronger, more complete, more sustainable, etc.? That means every choice either reinforces a bad ecosystem (privatized, insecure, unsustainable, incomplete, etc.) or it reinforces a good ecosystem (public, free, sustainable, complete, with succession plans, rich metadata, interoperability, re-usability, etc.). The cost of inaction is too high. We cannot afford NOT to do the right thing.

Thanks!

Further Reading

- When we depend on pointing instead of collecting. [http://freegovinfo.info/node/3900]
- Less Access to Less information by and about the US government. [http://freegovinfo.info/less_access]
- A librarian reacts to “A librarian reacts to wikileaks.” [http://freegovinfo.info/node/3178]
- Meet the Punk Rocker Who Can Liberate Your FBI File. [http://www.motherjones.com/politics/2013/11/foia-ryan-shapiro-fbi-files-lawsuit]
- Dodging the memory hole. [http://freegovinfo.info/node/10087]
- FACA Surveys, ICE’s “Egregious” FOIA Violations, the FBI’s Convoluted FOIA Search Process, and Much More. [https://nsarchive.wordpress.com/2015/06/25/faca-surveys-ices-egregious-foia-violations-the-fbis-convoluted-foia-search-process-and-much-more-frinformsum-6252015/]
- Document Friday: 54,651,765 US Documents Classified in FY 2009. [https://nsarchive.wordpress.com/2010/04/16/document-friday-the-information-security-oversight-offices-report-to-the-president/]
- FOIA for Profit. [https://nsarchive.wordpress.com/2012/11/06/foia-for-profit/]
- OAIS / TDR presentation at FDLP. [http://freegovinfo.info/oais_tdr]
- Libraries, public access to information, and commerce. Herbert I. Schiller and Anita R. Schiller. In *The Political Economy of Information*. edited by Vincent Mosco, Janet Wasko.