Digital Preservation of Federal Information Summit

Introduction

On April 3-4, 2016, information management thought leaders from across the United States met in San Antonio, Texas for a summit to foster cross-sector action to preserve and provide access to digital government records and information.

Hosted by the University of North Texas (UNT), the Digital Preservation of Federal Information Summit gathered stakeholders from a variety of public and private organizations, including archivists, librarians, technologists, program officers, executive directors, and other interested parties.

The meeting sought to engage national leaders in a structured, facilitated dialogue on at-risk digital government records and information. It also aimed to explore the development of a national agenda to address the preservation and access of priority content in this area.

The Summit's facilitated sessions were structured to produce several outcomes, including determining priorities for digital government records and information preservation action, and practical next steps to address these priorities.

Authored by the event's facilitators and edited by participants, this report offers a synopsis of the event's structure, work sessions, and outcomes. It summarizes the Summit's discussions and recommendations around how best to ensure the longevity, accuracy, and accessibility of U.S. digital government records and information for an informed citizenry. For the full list of participants and facilitators, please see p. 17 of this report.

Digital Government Records and Information

In the pre-digital production era, a clear workflow accounted for the preservation of most government records and information. Federal agencies created content, and when that content was ready to be disseminated or archived, they usually transferred it to the Government Printing Office (GPO, now the Government Publishing Office) and/or the National Archives and Records Administration (NARA). It was necessary for all federal agencies to engage with these specific federal entities in order to circulate their work in print forms.

The resulting workflow ensured that agencies that bore clear regulatory responsibility for the selection and preservation of government records and information for long-term access by U.S. citizens were involved as soon as a record or report/publication was completed. The GPO, in turn, partnered with a swath of public-, academic-, and other libraries, designated as "Federal Depository Libraries", who received, archived, and disseminated many of the resulting physical reports and publications that were selected for long-term preservation and access. Government records appraised as having permanent value were accessioned into the collection of the National Archives according to the timeline

established in the agency's records retention schedule. The National Archives made these records available to the public.

For printed records, reports, and publications, this workflow continues to function well today.

However, today, most government information is produced and disseminated *digitally*. This workflow is not nearly as predictable or smooth in a digital environment, and the number of publications has exploded.[1] Individual federal agencies now have the ability to quickly and easily publish their work themselves, without involving the GPO or NARA -- as per Office of Management and Budget (OMB) A-130 guidelines. So long as the work is not categorized officially as a "record" or a "report" or a "publication," regulatory authority does not require agencies to maintain content themselves, nor to provide it to the GPO or NARA for ongoing care.

Information produced by agencies in digital form should be scheduled as a record like all other government-produced records. The Federal Records Act covers most information created or received by agencies in the course of conducting government business, regardless of format. Rcords that are appraised as permanent should be transferred to NARA, usually after 15 years. Whether that is consistently happening is another question, whether the 15 year waiting period when agencies are responsible for providing access to their own information meets the needs of access and preservation is another question, and whether other institutions might want to keep some information that NARA appraises as temporary is yet another.

The resulting quick circulation of information that falls into the gap between official "records" (NARA) and official "reports and publications" (GPO) arguably serves the short-term interests of U.S. citizens. However, it breaks the chain of custody that has long ensured that government records and information are assessed, selected, and preserved.

In other words, this workflow shift has undercut the GPO/NARA-based central pathway to selection and long-term retention and preservation for much of our government's output as necessary for the "National Bibliography."

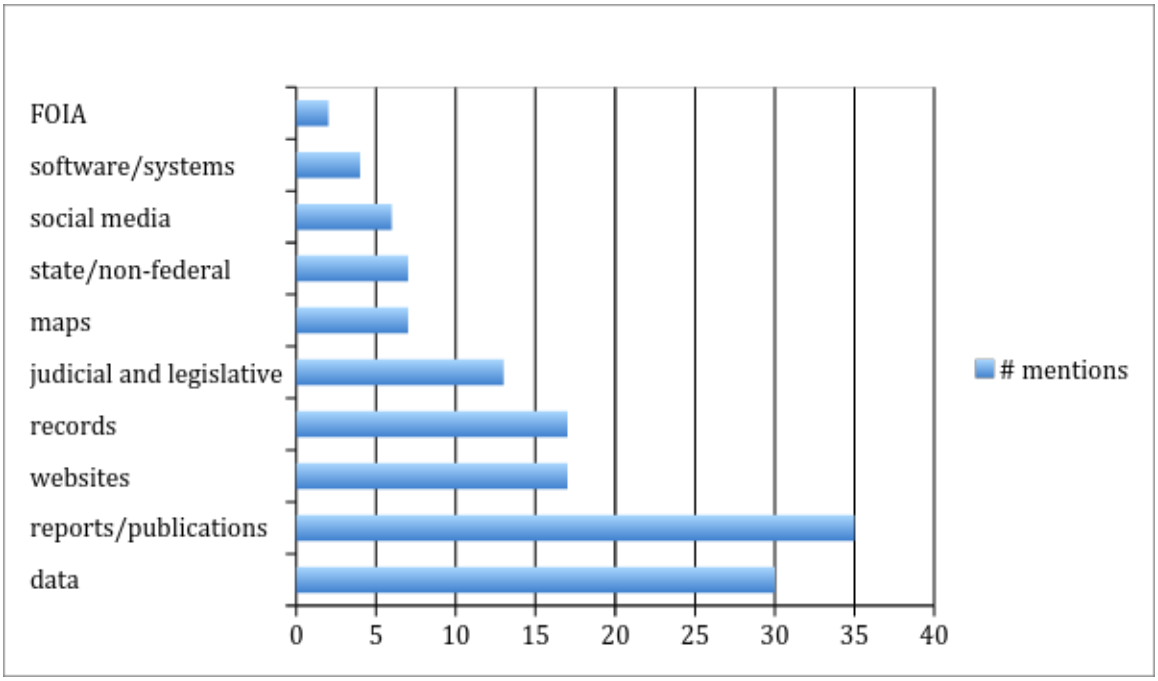
The goal of the Summit was to gather the stakeholder groups involved in government information preservation to explore how best to ensure the longevity, accuracy, and accessibility of *digital* U.S. government information for an informed citizenry. The Summit's participants discussed this challenge from many angles over the two-day meeting.

[1] For more, see "Born-Digital U.S. Federal Government Information: Preservation and Access," March 2014. Prepared by James A. Jacobs for "Leviathan, the Center for Research Libraries Global Resources Collections Forum." http://bit.ly/Jacobs-born-digital-leviathan-report

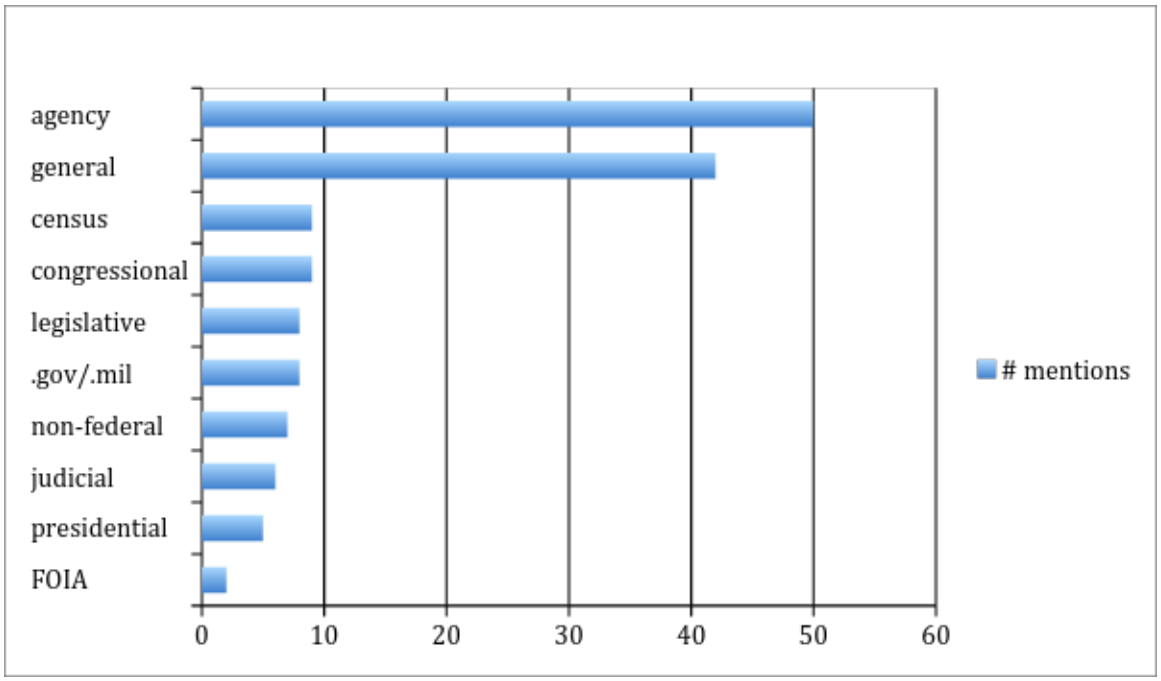
Identifying At-Risk Digital Government Information

In the first session, participants were asked to use sticky notes to individually identify what digital government information should be preserved, and to note the preservation status of each content type/category they identified. See Appendix A for the full list.

As shown in Chart 1, participants identified 10 main categories or types of content, and participants were also asked to mark categories as "preserved", "not preserved" or "somewhat preserved". The most popular types identified were data, reports/publications, websites, and records of various types.



Viewed through a different lens, the same information can be broken down by topical types, as displayed in Chart 2 below.



Clear trends emerged in this exercise across participants, with an emphasis on particular categories of content, including reports and publications, data, and agency information. Looking across all categories, 41 notes were marked preserved, 37 were marked not preserved, 14 were marked as mixed, and an additional 54 had no marking.

For phase two of this session, individuals worked in three groups according to randomized assigned tables. Each built upon its individual responses by now charting out the following information as a group: 1) What government information resource types/resources need preservation attention, 2) What is the current preservation status for each, and 3) What are the barriers to preservation for each. See Appendix B for full charts transcribed from the event.

Looking across the three groups, three of the same categories appear, signaling broad consensus on these data types: “Data,” “Websites,” and “Reports.” Additional categories appear in two of the three groups: “Records,” “Maps,” and “Legal.” Notably, the majority of categories appeared in only one of the three groups: “Judicial,” “Presidential,” “Code and Computer Programs,” “Images/media,” “Digitized,” “Non-crawlable Content,” “State Department documents,” “Social media,” and “Special.” From this data, we see both convergence and divergence in awareness of and/or concerns about many digital government information types among this multi-sector group.

The groups then established selection/prioritization criteria that could be used to determine what government information resource types and resources should be collected, archived, and made accessible. Each group came up with similar categories, as seen in the comparison chart below:

Group A	Group B	Group C
Importance to the public	Potential use/value by public	Public interest
Framework for risk evaluation	The longtail	Vulnerability/risk/coverage
Scoping and funding analysis	Size/scope; Format	Feasibility; Funding; Format
Scientific and historical value	Historical significance	
National experience, Rights, Government accountability		Official status, Institutional mandate

Finally, the groups used their selection criteria to rank their resources/resource types in order of priority. "Data" was a clearly shared concern, as two of the three identified "Data" as the top priority, and the third included it as its fourth priority. All three groups also mentioned "Reports" as one of their top three priority content types. Two of the three groups also identified "Websites" as one of their top concerns; interestingly, one group did not include websites at all, instead focusing on "Non-crawlable content." Similarly, two of the three included "Software" as a priority.

Their concerns diverged in additional priorities, with one group each including the following: Social Media, Legal, State Department content, Maps/GiS.

To wrap up this first work session, participants were asked to shift to new tables, again according to randomized assignments. At each table, participants were asked to close their eyes and then one at a time, going clockwise around the circle, each participant said what digital government information content type or category s/he believed was most at risk. After going around the tables once and sharing a variety of perspectives, participants were asked to close their eyes again and continue going around the table until all participants reached consensus (i.e. until all participants said the same content category/type).

This "Consensus Circle" exercise quickly yielded agreement at two of the three tables, both selecting "data" by the third time around the circle. The third table came to consensus after a few more rounds, selecting "Website Data" as a compromise between two strongly held opinions at this table.

This consensus on "data" should not be taken out of context. However, coming to quick consensus—and a (mostly) shared consensus room-wide—helped us to identify and mark a topical area for the remaining work the facilitated group would accomplish on site in San Antonio.

Building Preservation and Access Opportunities

Pecha-Kucha presentations (20 slides, 20 seconds each) were given by Jefferson Bailey and by Michael Nelson/Herbert Van de Sompel on infrastructure and workflow issues. These set the tone for the facilitated work sessions that followed.

The room stayed in its second configuration, with randomized groups that had already reached consensus on a content type. Each group was asked to now work on the content type it had identified. Using a combination of real and imagined tools and methods, each table was challenged to design a collection, archiving, and access pathway(s) for that content type.

The resulting workflows provide an initial mapping of some of the elements needed for progress. Each illustrates a set of opportunities for collaborative building of tools and standards around government digital data preservation.

Group 1 focused on replicating a traditional library life-cycle model, focusing on the following steps for government digital data:

1. Identification/inventory
2. Selection
3. Description
4. Access
5. Preservation

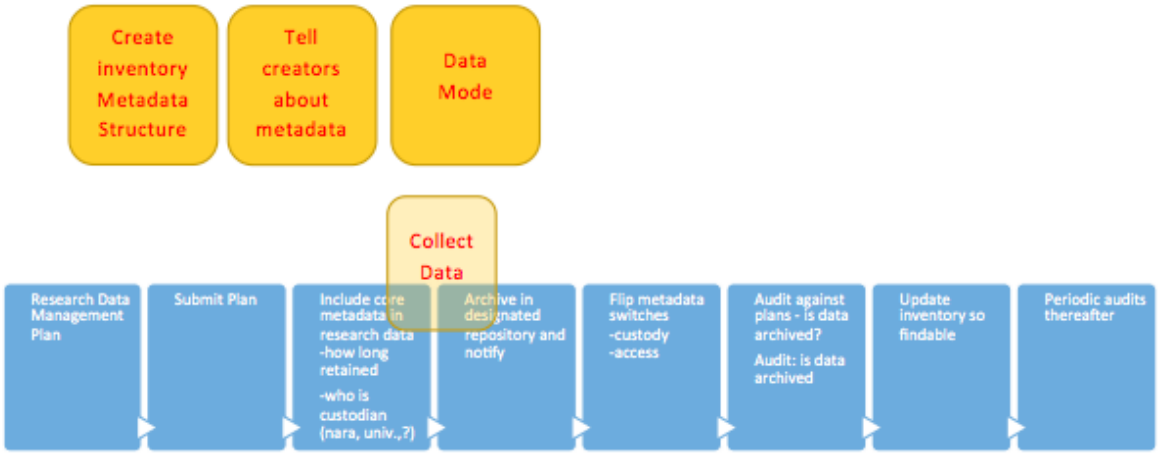


Image 1: Group 1 workflow draft for Government Digital Data

Group 1 identified the following elements/tools/standards as currently missing and in need of being established to make this access and preservation activity possible:

1. Resolution service (DOI)
2. Formatting
3. Incentives
4. Selection tools

Group 2 focused first on defining data as structured information, products of research efforts. The group marked two categories of concern: data created by federal agencies, and data created in federally funded work. The group discussed the need to build on the work of a set of stakeholders that have worked on two core issues—data specifications and metadata vocabulary—including Horizon 2020 and other European Union efforts.

Group 2 marked the following elements/tools/standards as currently missing and in need of being established:

1. Business models
2. Custodian designation
    a. An important conversation here focused on the need to build on the capacities offered by cloud-based storage to make possible a shift in *designation* rather than a shift in *environment* for content. In other words, if a collection is stored in a cloud-based environment, it should be possible to change custodians and preservation actions via a simple rule-based action rather than having to physically transfer the data to a new location.
3. Metadata specification
4. Verification authenticity and auditing tools (in essence, creating a kind of parole officer for data)
5. Inventory

The overall workflow Group 2 began to chart on site included the following:

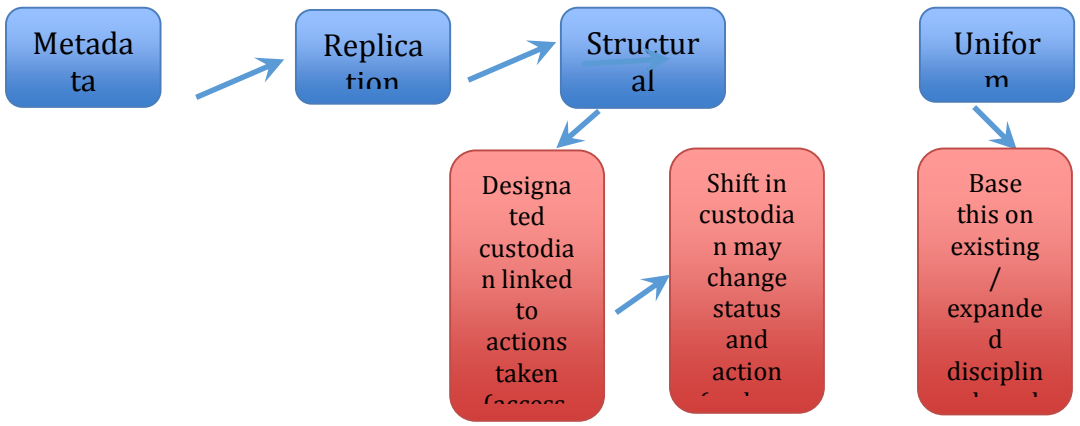


Image 2: Group 2 workflow draft for Government Digital Data

Group 3 identified a set of top-level steps and deeper sub-steps as necessary elements in government digital data access and preservation. As depicted in Image 3 below, the top-level elements included the following:

1. Discovery
2. Gathering
    a. Website crawling
    b. Deep web harvesting
    c. Non-web data
3. Public Access

Image 3: Group 3 workflow draft for Government Digital Data

Consistent across these three workflow imaginings, the groups talked about how to build on existing specifications (e.g., data management plans from federal granting agencies) to facilitate the identification and preparation of content for inclusion. They also all discussed how to use metadata to trigger actions over time, including shifts in custody and shifts in various curatorial actions. All three groups also discussed the need for analyzing use cases and undertaking awareness-raising efforts with the federal agencies. These consistencies point to areas of needed attention; they also move in the direction of infrastructure creation and away from some of the "ad hoc" activities currently underway.

Individual group "hot topics" included Group 3's emphasis on "deep web harvesting" and the lack of tools to perform this type of ingest; Group 2's work to identify ways to simplify the hand-off from agency to preservation entity via the uses of metadata tags and cloud technologies; and Group 1's focus on establishing a strong metadata specification as the foundation for life-cycle management of digital government information.

Transition Talks

At the end of the first day, we took 30 minutes to synthesize what we had done over the course of the day and to spotlight key issues that had arisen for participants. The first major issue discussed was the costs incurred (staffing, money, time) when a federal contract expires and the agency determines it needs to move to a new storage entity. The movement of content between two storage entities is expensive because the data shifts from its "at rest" state to a state of transfer. The cost of transferring content between clouds could be lessened if, instead of requiring a physical relocation of the content from one set of servers to another, we built a national cloud infrastructure that enabled an agency to shift the designated custodian (and possibly, the archiving and preservation actions it undertakes) via the content's metadata *without* moving the content from one server farm to another.

Additional issues that arose at the end of the first day included the need for incentive structures to help agencies willingly share their content for preservation (as one person noted, "There is a gap between the content that is considered 'records,' which NARA collects, and 'reports and publications,' which the GPO collects, and most agencies want to land squarely in that gap. Then, they can have full control over their content and don't have to take any actions to share it with another federal entity.") Participants talked about the "data management plan" enacted by federal funding agencies as a possible model that might be helpful to instate with federal agencies.

The group also discussed the frustration of not having enough data to understand the behaviors of agencies in digital publishing today. One research division at the Library of Congress may be looking at this issue later this year.

Finally, the group discussed the difficulty in advocating for changes due to mandate constraints. The constitutional separation of powers draws a deliberate line between executive branch and legislative branch agencies. There is no penalty for noncompliance with content handoffs (records to NARA, publications to GPO), and there is no way for NARA/GPO to do more than provide guidance on how to enact a handoff. The Federal Records Act (for "records") and OMB Circular A-130[2] (for "information resources"), in other words, are limited in impact because preservation compliance is not currently audited. Research institutions and other longstanding partners in the preservation of printed government records may be able to play a helpful role in identifying and crawling content that falls into the "gap" between "records" and "publications."

Two additional issues raised at the close of the first day were NSA phone data (is it preserved? Can it be preserved?) and FOIA requests (what happens to 1) the requests, 2) the response, and 3) where relevant, the documents released due to a FOIA request?).

On day two of the event, the conversation and work shifted from the "what" to the "how and who." We opened with a 30-minute "What's on your mind" session, where everyone

[2] Office of Management and Budget Circular A-130
https://www.whitehouse.gov/omb/circulars_a130_a130trans4/

was invited to share what they thought about and talked about overnight that had not yet made it into the workshop documentation. Participants were especially encouraged to bring up issues they believed needed attention via a cross-sector stakeholder group, whether now or later.

The resulting "Parking Lot" included a range of items, as documented below:

- We need to leverage work that is already done
- We need to create/articulate incentives for others to do this work
- We need a standard way to identify and make available content: print = books, catalog; digital = ??
- We need to rely upon existing networks and associated activities where possible; maybe enlarge those networks instead of creating new ones
- Key stakeholders
    - Missing voices today (here) include – vendors, scholars
- We have a dire need for an environmental scan
    - Who and where--difficult to fund or empower
- Articulating (and creating?) incentives for agency participation are core to this work's success
- We need to think through how to move people from institution-level views to system-level views. What's difficult to justify at the institution level is hard to ignore if you look at it from a system level.
- What to do with outsourced security certificates (David Rosenthal talked with us about the possibility of using https and security certificates as a mechanism for identifying content to ingest; outsourcing of certificates makes this impossible[3])
- We need solid definitions – publication, record
  create picture of what & where
- Preservation and access – we need to determine, are they the same thing or separate?
- What is the role of…
    - Government
    - Nonprofits
    - Private sector
- Criteria for prioritization (remember our biases)
- What exists
    - What do we want to collect
    - How long is each piece valuable
    - To whom can it be made accessible (privacy, national security restrictions, etc.)
    - Is that happening

Mobilizing Collaborative Efforts

The second day of the event focused on creating a vision statement and mobilizing around that vision.

[3] For more, see David Rosenthal, "The Curious Case of Outsourced CA", http://blog.dshr.org/2016/04/the-curious-case-of-outsourced-ca.html

Participants sat in new table configurations that were established prior to the event through random assignment. We started the first session with two pecha kucha presentations, by Katherine Skinner and Mary Molinaro, both focused on coalition-building and multi-partner initiatives.

We then moved to a 15-minute plenary activity in which we sought to articulate the change we wish to see in digital preservation of federal government information. The activity began with the presentation of a placeholder agenda statement, as follows:

"U.S. digital government information is selected, collected, archived, and made accessible to the public."

All participants were asked to add to, edit, and revise the statement until it worked for everyone. We edited the document "live" on screen, finalizing it as follows:

"U.S. digital government information is openly selected, collected, consistently described, registered, preserved, and made freely accessible to the public."

This statement was then used as our focal point for a "World Cafe" exercise. In this exercise, each table was assigned a task. Participants worked together on one of three tasks at one of three tables for 15 minutes, taking notes as a group (no single designated scribe) on a flipchart in the middle of the table.

At the end of 15 minutes, the participants were instructed to choose a different table, leaving one participant behind at each of the three tables to explain what the groups had accomplished and documented thus far.

The new groups then built upon the work of the previous groups for 15 minutes, again documenting their work on the flipchart, and then they were asked to switch tables again, choosing a new table, and leaving a different participant behind at each table to explain what the groups had accomplished and documented.

This loop was repeated one more time, and in the last shift, participants who had visited all three tables were asked to return to their initial table, and participants who had stayed behind once were asked to move to whatever table they had not yet visited.

After these final groups assembled, we had each report out about the evolution of the response and to evaluate where there were points of synthesis and convergence in the extended work of each table over time. The three questions that the three tables grappled with were as follows:

1. What resources and opportunities can we leverage toward this change?
2. Who needs to be involved? What are their roles?
3. How will we know if we are successful?

Below, we include images of the flipcharts and brief distillations of the top themes and consensus points that arose in each table's work.

Table A: What resources and opportunities can we leverage toward this change?
Synthesis points:

- The range of resources available include existing laws and schedules (IP, record schedules, executive order for open machine readable data~~.ison~~); standards (existing/emerging); existing tools (Fedora, distributed digital preservation networks, Memento, REST, ResourceSync, etc); existing content (portable web documents, existing web archives of gov info); and existing partners (federal depository libraries, stakeholder/user feedback to ID what's useful and needed, Michael and the Memento team's ongoing work, existing groups).
- The range of opportunities includes the potential for NGO/watchdog groups to pair up and advocate for change, the general push in gov toward open gov and transparency, existing federal agency conversations and groups, and the changing role of the depository libraries.

Flipchart image:

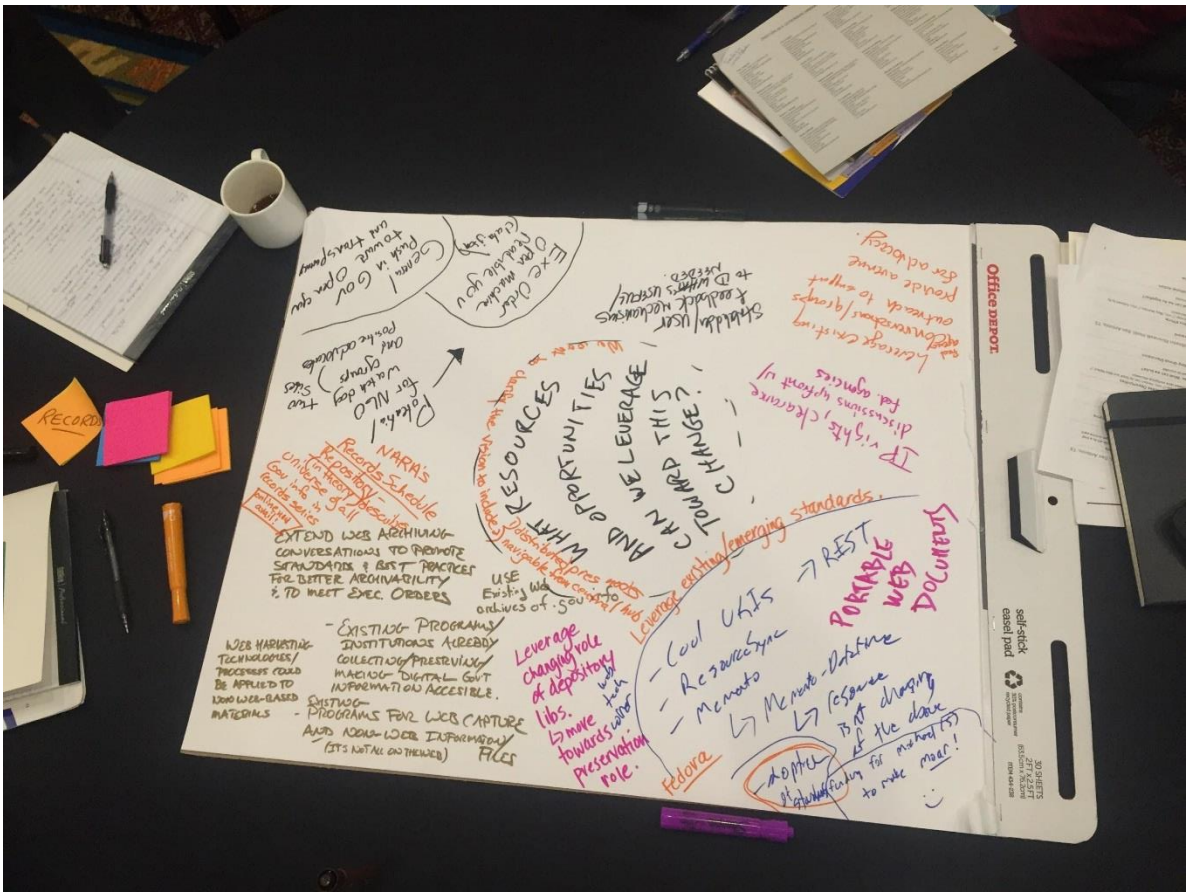
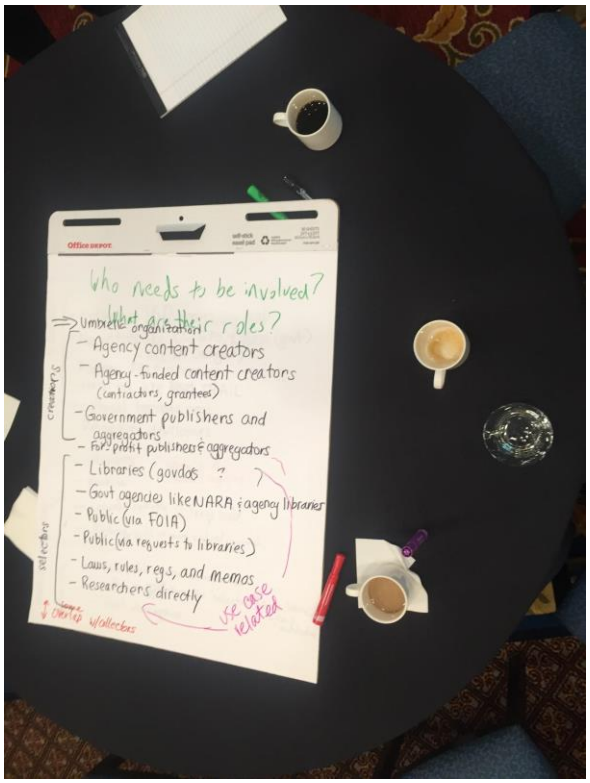


Table B: Who needs to be involved? What are their roles?
Synthesis points:
The groups that need to be involved and the roles played by each were documented by Table B as follows:

Umbrella Organization	○ ?
Creators	○ Agency content creators ○ Agency-funded content creators (contractors, grantees) ○ Government publishers and aggregators ○ For-profit publishers and aggregators
Selectors	○ Libraries (gov docs, ?) ○ Government agencies like NARA and agency libraries ○ Public (via FOIA) ○ Public (via requests to libraries) ○ Laws, rules, regulations, and memos ○ Researchers directly
Collectors	○ Collectors TBD – missing gap ○ Harvesters like LC, GPO, NARA (gov't); IA (non-gov't), UNT, Stanford ○ Academic FDLP libraries ○ Consortia like TRAIL ○ Cultural heritage organizations ○ Researchers / developers
Describers/Registers (central registry?)	○ Catalogers ○ Machine ○ Crowdsourcing/ large scale ○ Finding aids – more archival approach
Preservation	○ Non-profit preservation agencies ○ Different people than collectors // same basic group institutions ○ NARA, other agencies

Table B Images:



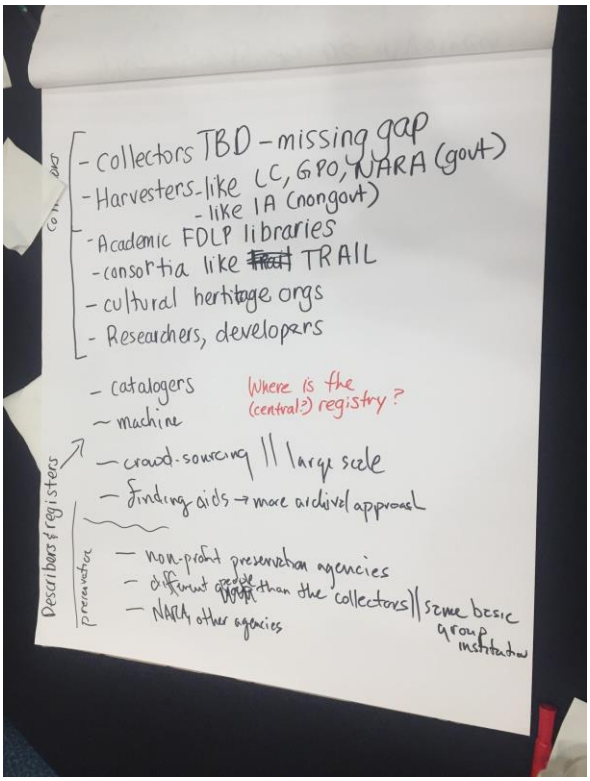


Table C: How will we know if we are successful?

Table C's activities sought to define outcomes against which the success of a group toward the stated vision above could be measured.

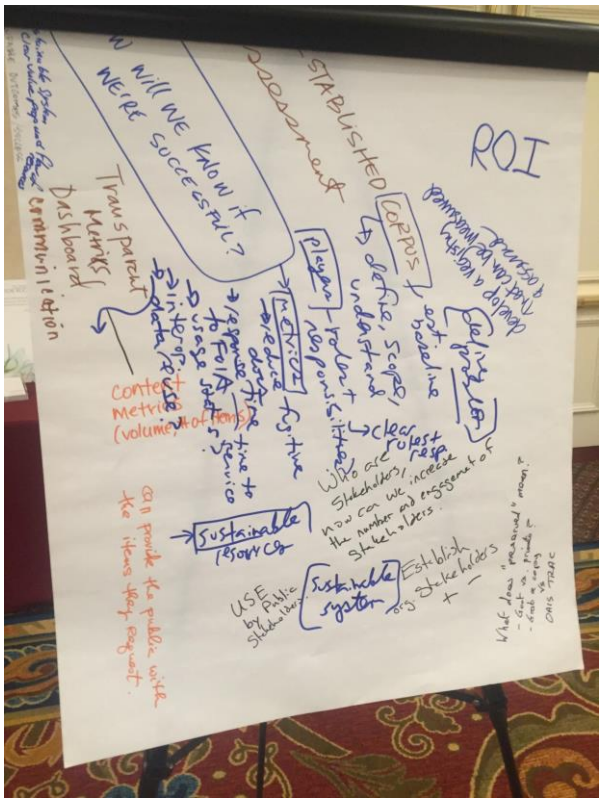
Success metrics this group identified included the following:

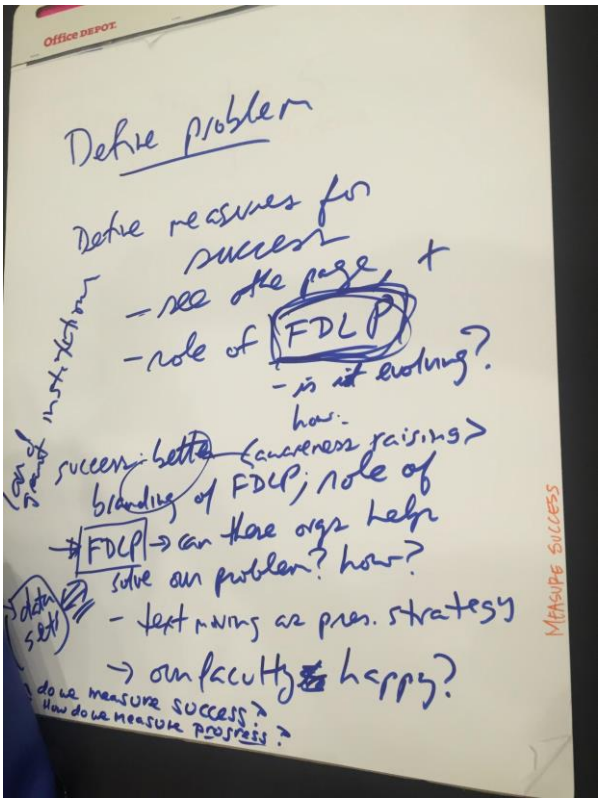
1. an established corpus, as measured against a baseline
    a. # of items, volume
2. a registry (similar to the "Keepers Registry" http://thekeepers.org/registry.asp for journals) against which compliance can be measured and assessed
3. an increase in the number and engagement of stakeholders
    a. Usage stats
    b. Data reuse
    c. healthy ROI
4. Raised awareness of need (perhaps measured by funds donated or citizen advocacy or some other metric?)
5. Response to FOIA (does the response time decrease as more content is in preservation status? Is more content made available at FOIA request because it is archived?)

Qualities of the outcomes measurement system that this group identified include:

1. Transparency
2. Dashboard views
3. Ability to assess according to a baseline
4. Demonstration of ROI
5. Show preservation levels (e.g., differentiate between grabbing a copy and preserving it in a TRAC-audited OAIS)
6. Usability

Table C Images:





Next steps

The final session returned to the initial goals of the event and invited participants to identify key next steps, as based on their conversations and work sessions. They identified the following:

1. Conduct baseline/environmental scans
2. Complete and release a report from the event, edited by the participants
3. Compile and release a position paper from the event, authored by the participants
4. Establish a Listserv, opt in by participants
5. Identify possible facilitators who can serve the "umbrella" function for a longer-term effort or coalition
6. Create a definition list for terminology
7. Work towards public programming/awareness raising

8. Develop a proof of concept

Conclusion

Several key recommendations were introduced by this cross-sector stakeholder group to improve the longevity, accuracy, and accessibility of U.S. digital government records and information for an informed citizenry. Specifically, the group returned many times to the degree of ignorance we all have about the scope and scale of the problem at hand. What is actually collected and preserved today (and in what manner)? What is *not* collected and preserved today? How many entities are actively collecting and preserving content, and what schedule or intent do they express for this content? The need for ***environmental scans*** and for a ***registry*** were perhaps the most immediate calls for action.

Although the group arrived with many feeling trepidatious about forming any type of cross-sector workgroup or alliance, by the end of day two, the group enthusiastically approached the idea of engaging facilitators to play an "umbrella function" to glue everyone together for longer-term action. ***Forming an active coalition of interested institutions across the public and private sectors*** was discussed. Existing coalitions were mentioned as possible models, including Technical Report Archive and Image Library (TRAIL), which is administered and hosted by CRL, and the GPO's Federal Information Preservation Network (FIPNet).

Finally, the group returned many times to the need for definitions to ensure that what one group says is understood correctly by other groups in this cross-sector landscape. Distinctions between terms like "record" and "information," and "preservation" and "distributed preservation," were especially highlighted for their potential to sow misunderstandings. By developing a shared vocabulary, a diverse coalition of interested institutions/people could begin to define the framework within which they will operate.

This event accomplished a great deal; participants agreed that additional conversation and work in this area would be fruitful to pursue.

DPFIS Attendees			
#	First name	Last name	Organization
1	Jefferson	Bailey	Internet Archive
2	Greg	Eow	MIT
3	Tara	Das	Columbia University
4	Ammie	Feijoo	IMLS
5	Declan	Fleming	UCSD
6	James	Jacobs	Stanford University
7	Leslie	Johnston	National Archives
8	Heather	Joseph	SPARC
9	Ron	Larson	University of Pittsburgh
10	Joan	Lippincott	CNI
11	Mary	Molinaro	DPN
12	Michael	Nelson	Old Dominion University
13	Michael	Neubert	Library of Congress
14	Trevor	Owens	IMLS
15	Maliaca	Oxnam	University of Arizona
16	Kristi	Park	Texas Digital Library
17	Mark	Phillips	UNT
18	Meg	Phillips	NARA
19	David	Rosenthal	Stanford University
20	Anthony	Smith	Government Publishing Office
21	Herbert	Van de Sompel	Los Alamos National Labs
22	David	Walls	Government Publishing Office
23	Kate	Wittenberg	Portico
24	Robert	Wolven	Columbia
25	Sinai	Wood	Baylor University
Facilitators			
1	Martin	Halbert	University of North Texas
2	Roberta	Sittel	University of North Texas
3	Katherine	Skinner	Educopia Institute