The Challenge of Hidden Big Data Collections:
Making Digital Congressional Papers Available for Scholarly Research

Emily Boss / Head of Metadata and Cataloging / University of Nevada, Reno Libraries
Nathan Gerth / Head of Digital Services / University of Nevada, Reno Libraries
Jessica McMillen / Head, Digital and Web Services / West Virginia University Libraries

CNI 2018 Fall Meeting

6,271,611 objects

Our thesis: “Digital congressional collections present unique challenges that go beyond standard born digital acquisitions. More akin to big data, they include content that eclipses the capabilities of standard computation platforms and forces institutions to transform their systems accordingly.”

Roadmap

1. Nathan will explore the factors that can that can “hide” these collections from users
2. Emily will outline how they become transformative for institutions using the University of Nevada, Reno as a case study
3. Jessica will discuss the tool being developed at West Virginia Libraries to make constituent correspondence data searchable by researchers.

Hidden Big Data: The Challenges and Opportunities of Digital Congressional Collections

Nathan Gerth / Head of Digital Services /
University of Nevada, Reno Libraries

What is a Congressional Collection?

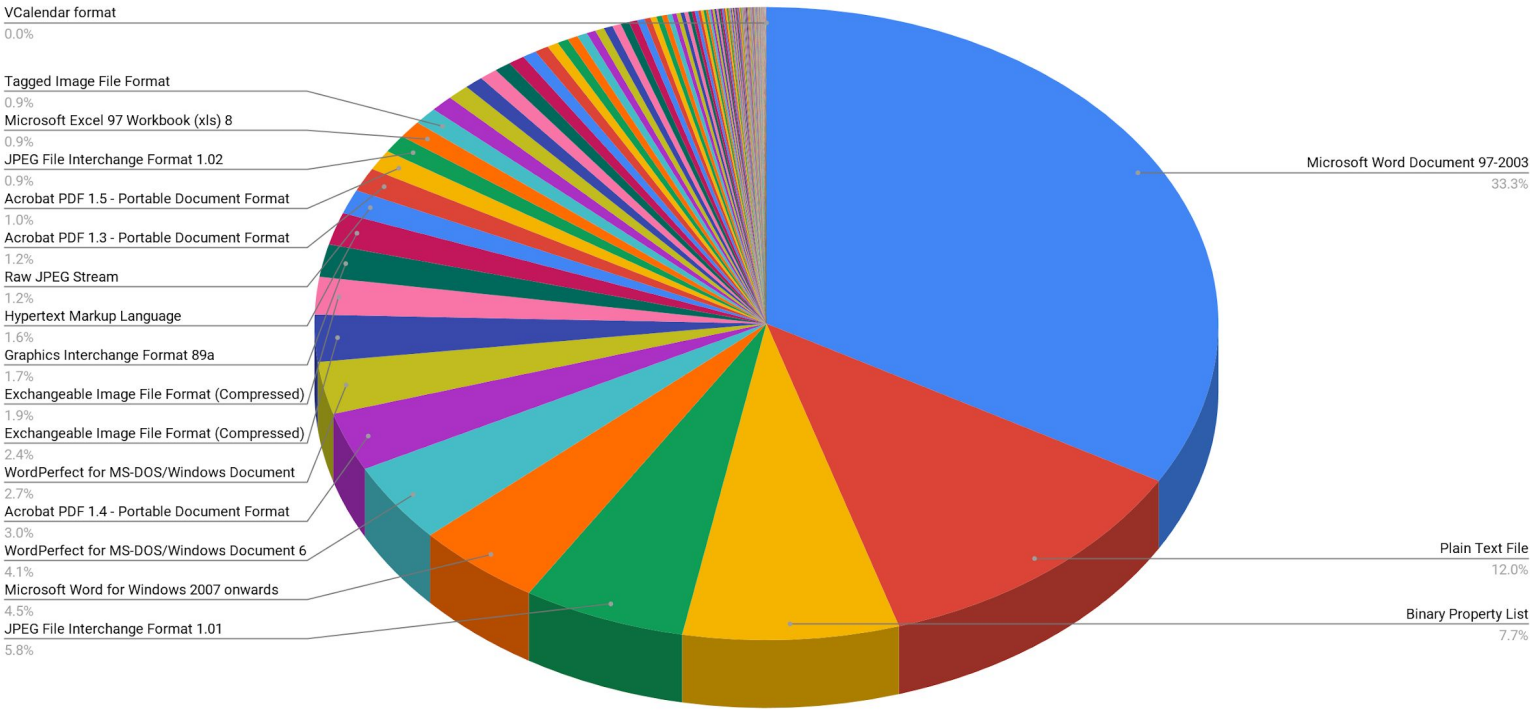
- Collections include all non-committee or non-classified records generated by an office
- These materials belong personally to the member
- Electronic records in these offices often include full information systems: email, CSS

What questions can be engaged with these collections?

- How do the American people engage with their representative institutions?
- How does technology evolve in government?
- How has the evolution of technology transformed the nature of politics?

Why does the digital content in these high-profile collections remain in the shadows?

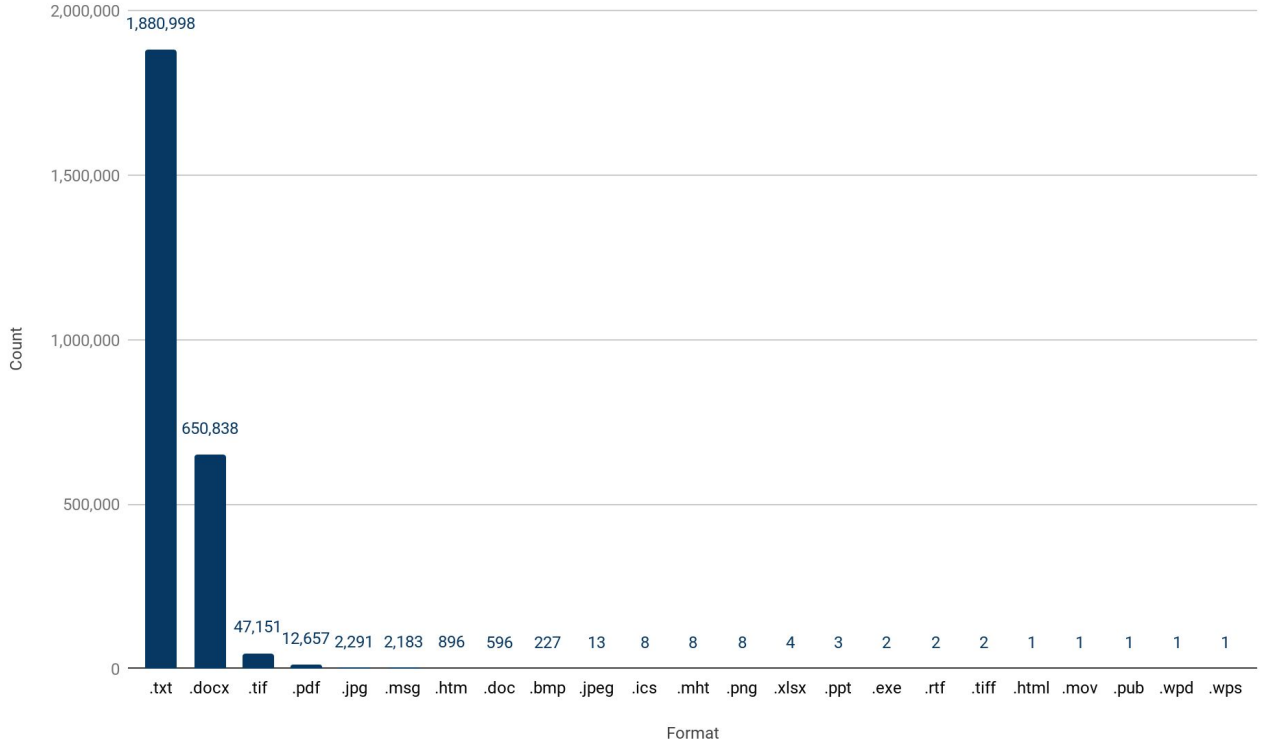
- The diverse and voluminous content in the collection
  - Reid Collection to date includes: 6,271,611 objects (12 TB)
- This poses challenges for storage, security, and carrying out effective workflows



File Formats, Senator Harry Reid Papers, DC Office (302 File Formats)

Why does the digital content in these high-profile collections remain in the shadows?

- The proprietary nature of the systems used by offices also create challenges
- CSS Data
  - Raw export consisting of millions of files and their data tables, but no computational layer to facilitate access to the data in its original form



Hidden Big Data: Senator Harry Reid Collection Constituent Services System Data

Congressional Collections and Transforming Library Systems at UNR

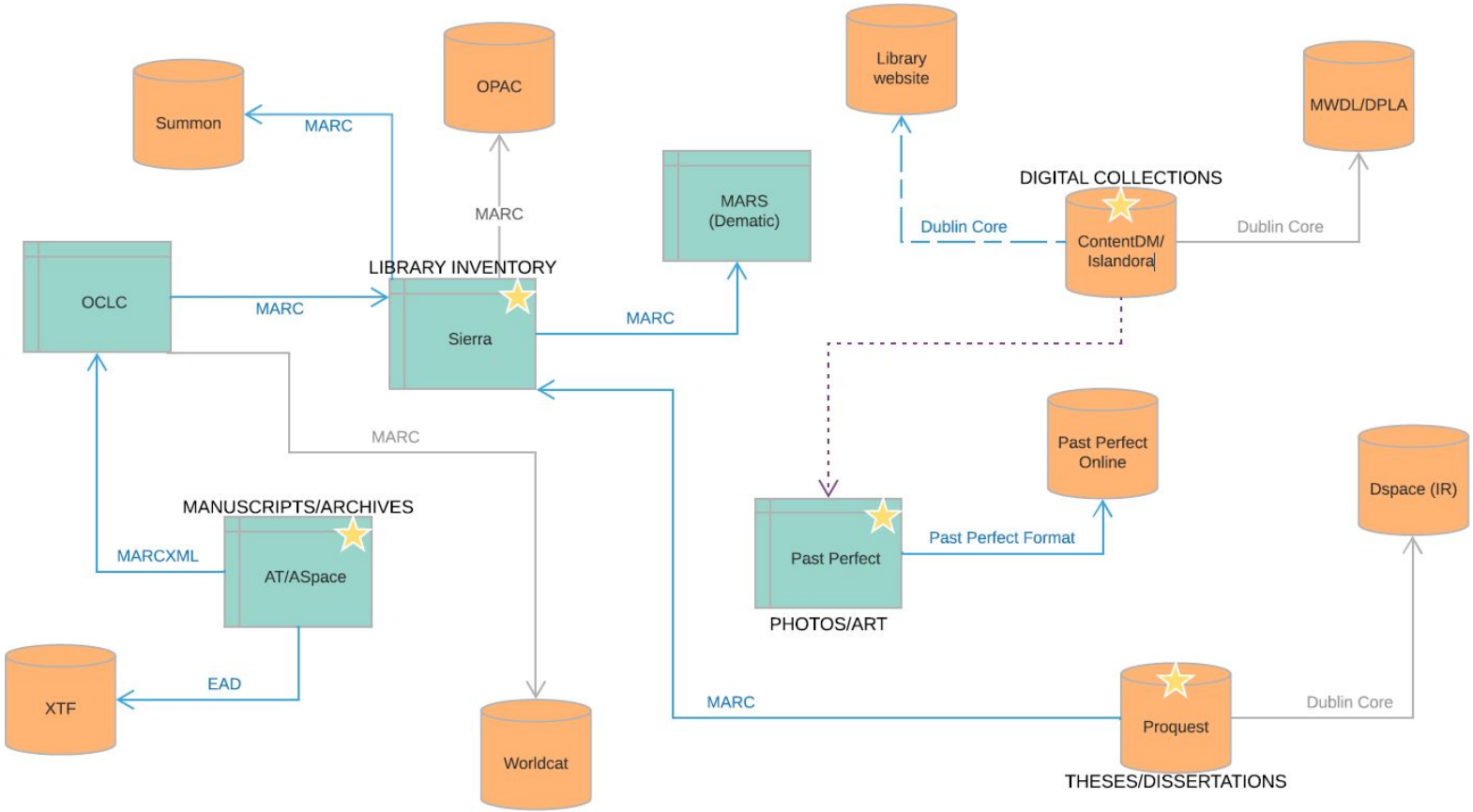
Emily Boss / Head of Metadata and Cataloging /
University of Nevada, Reno Libraries

Conceptually

- Research data collection vs. Collection of data for research
  - Storage
  - Data integrity
  - Data security
  - Rights management
- Multiplied exponentially

Infrastructure

- Tools to meet data demand
  - System mapping
  - “Reduce the toil”
  - Migrate, implement, sunset



System Mapping - April 2018

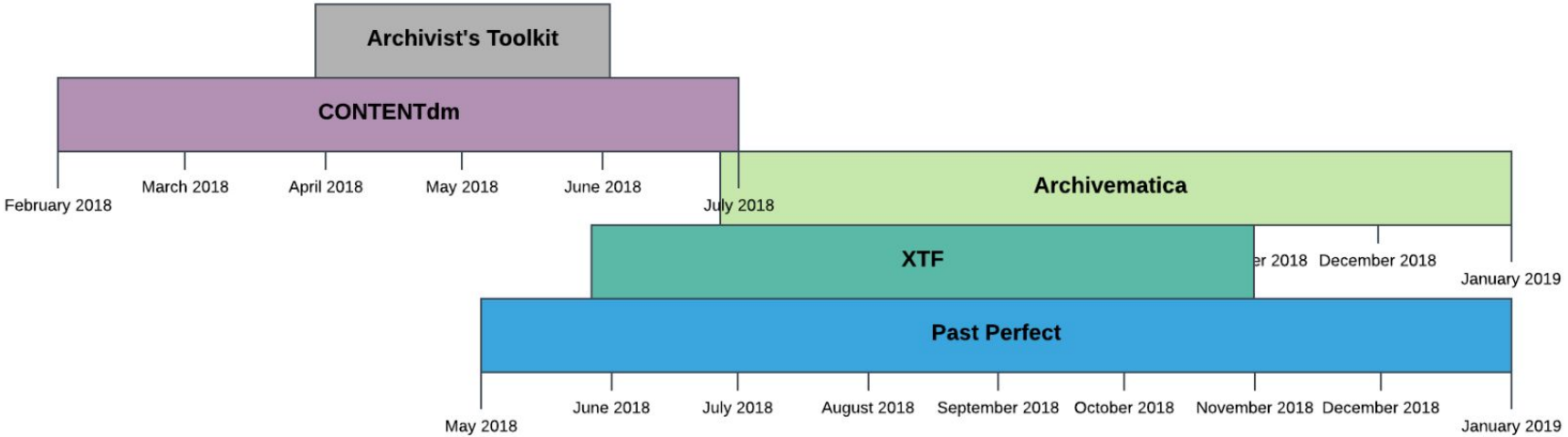
Eliminating Toil

Written by Vivek Rau

Edited by Betsy Beyer

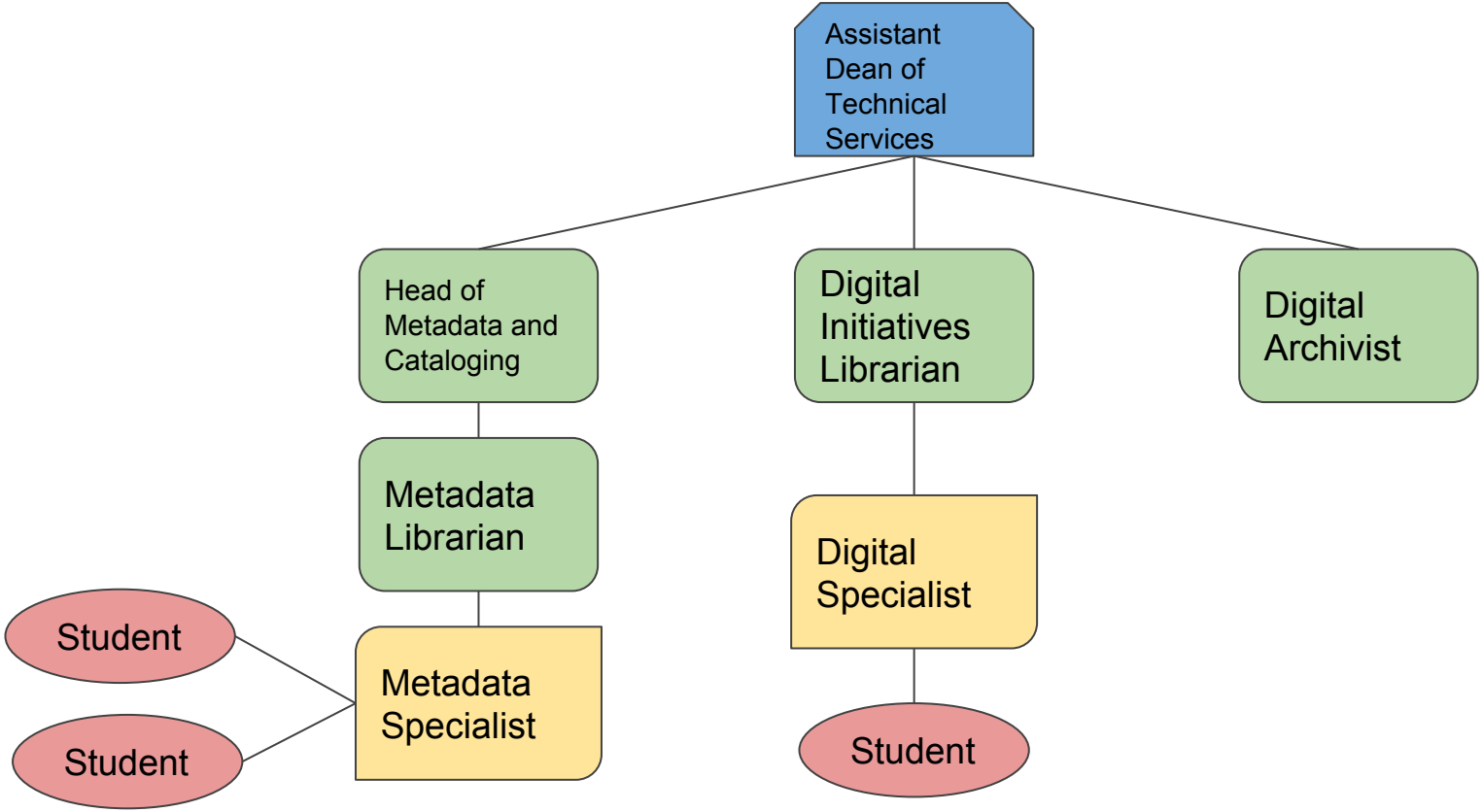
> “Toil is the kind of work tied to running a production service that tends to be manual, repetitive, automatable, tactical, devoid of enduring value, and that scales linearly as a service grows.”

Migrate, implement, sunset

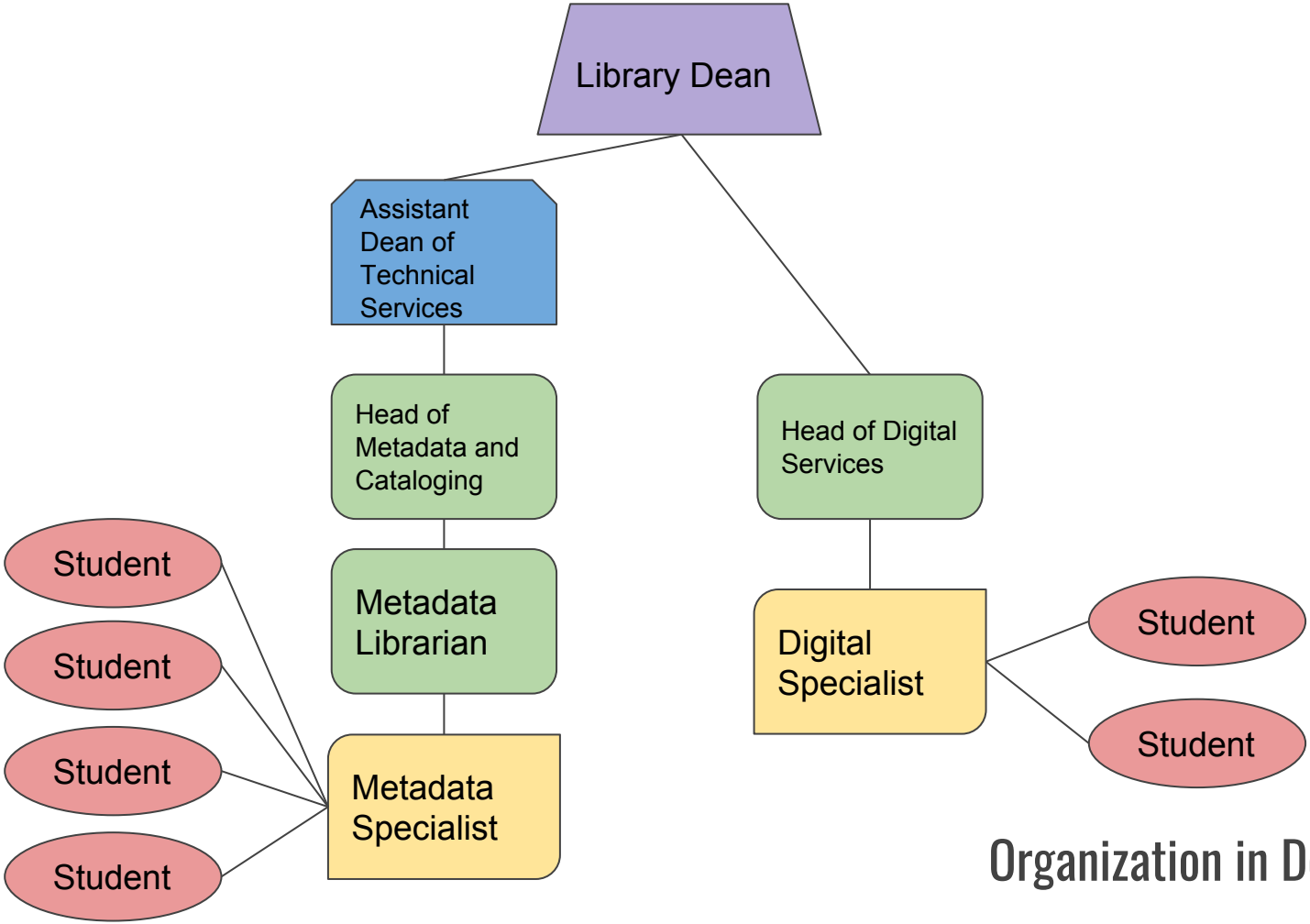


Workflows

- Workflow changes necessitate structural changes



Organization in February 2018



Organization in December 2018

Progress

- Reduced toil by 20%
- Reduced metadata duplication
- Gained a new streamlined department with decreased silos
- Over 600,000 views of digital materials in 6 months
- In 2019, slated to ingest 20,000 more digital objects then we did in 2018

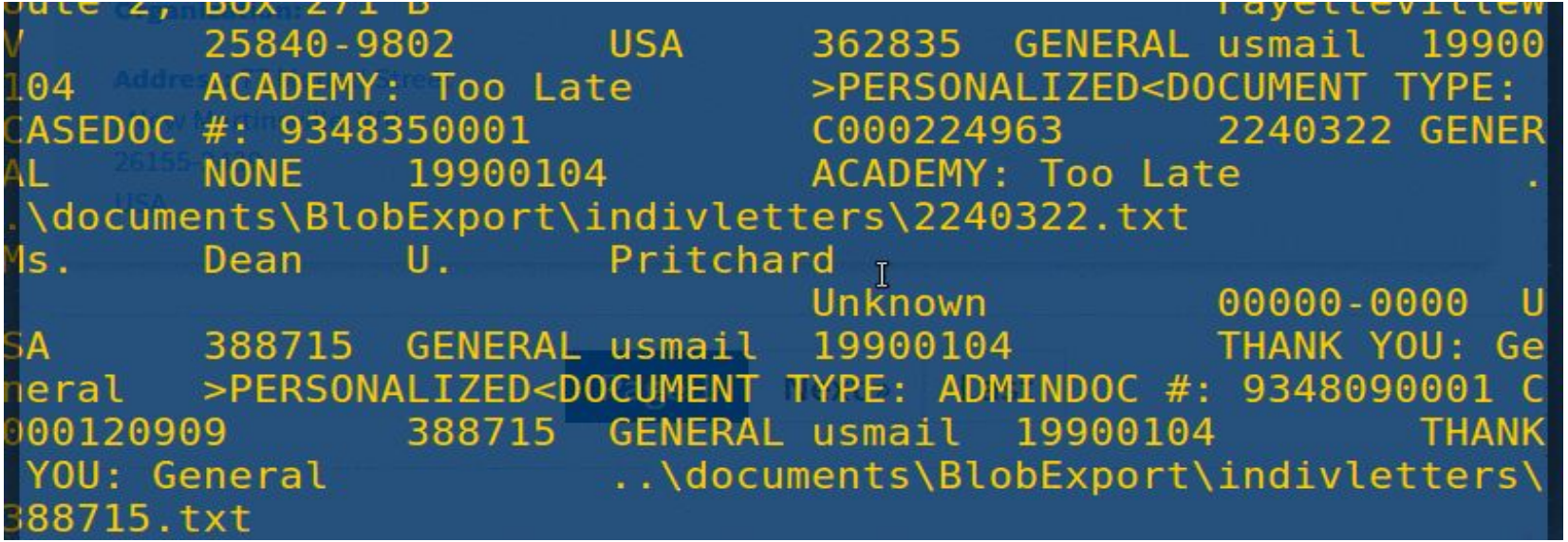
America Contacts Congress: CSS Data and West Virginia University Libraries

Jessica McMillen / Head, Digital and Web Services / West Virginia University Libraries

Data

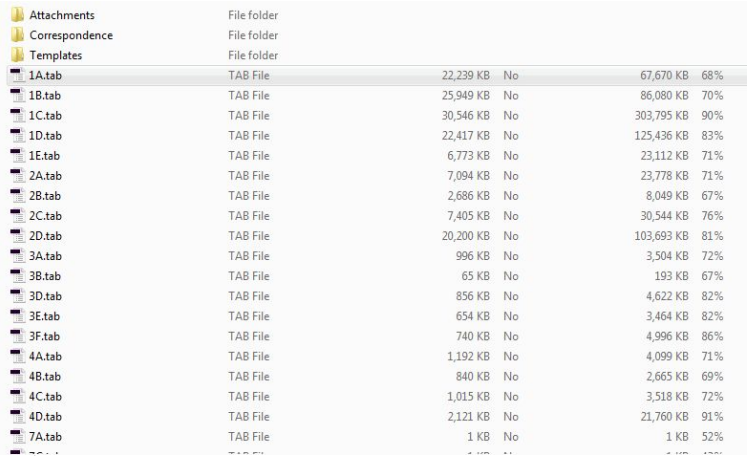
- Flat files
- CMS Data Interchange Standard database files

Flat File


V 25840-9802 USA 362835 GENERAL usmail 19900
104 ACADEMY: Too Late >PERSONALIZED<DOCUMENT TYPE:
CASEDOC #: 9348350001 C000224963 2240322 GENER
AL NONE 19900104 ACADEMY: Too Late .
.\documents\BlobExport\indivletters\2240322.txt
Ms. Dean U. Pritchard
Unknown 00000-0000 U
SA 388715 GENERAL usmail 19900104 THANK YOU: Ge
neral >PERSONALIZED<DOCUMENT TYPE: ADMINDOC #: 9348090001 C
000120909 388715 GENERAL usmail 19900104 THANK
YOU: General ..\documents\BlobExport\indivletters\
388715.txt

CMS Data Interchange Format

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	2D	6.69E+41	1.43E+14	1	CM	Conversion Data: 1 - OT -							
2	2D	6.69E+41	1.43E+14	2	STATUS	Closed							
3	2D	6.69E+41	1.43E+14	3	LETTERNA	Letters\I_1429596_							
4	2D	5.67E+40	1.55E+20	1	CM	05/02/2014 11:37:22 AM			- Meritorious Service Awards These				
5	2D	5.67E+40	1.55E+20	2	SUBJECT	Personal Letter-Kiwanis Award Dinner TY's							
6	2D	5.67E+40	1.55E+20	3	STATUS	Closed							
7	2D	5.67E+40	1.55E+20	4	LETTERCA	Personal Letter							
8	2D	5.67E+40	1.55E+20	5	LETTERNA	Kiwanis Award Dinner TY's							
9	2D	5.67E+40	1.55E+20	6	LETTERLAS	########							
10	2D	2.66E+41	1.48E+21	1	SUBJECT	Judiciary-USA Freedom Act Pro							
11	2D	2.66E+41	1.48E+21	2	STATUS	Closed							
12	2D	2.66E+41	1.48E+21	3	WEBSUBJE	IMA MAIL on WOMEN							
13	2D	2.66E+41	1.48E+21	4	WEBHEAD	Support the USA Freedom Act and Limit NSA Surveillance							
14	2D	2.66E+41	1.48E+21	5	LETTERCA	Judiciary							
15	2D	2.66E+41	1.48E+21	6	LETTERNA	USA Freedom Act Pro							
16	2D	2.66E+41	1.48E+21	7	LETTERLAS	########							
17	2D	6.75E+38	6.72E+19	1	SUBJECT	Environmental-Interior Appropriations Wildlife							
18	2D	6.75E+38	6.72E+19	2	STATUS	Queued							
19	2D	6.75E+38	6.72E+19	3	WEBSUBJE	FW: Website Contact Form Submission							
20	2D	6.75E+38	6.72E+19	4	WEBHEAD	As a resident of your district and a supporter of Defenders of Wildlife, thank you for voting a							
21	2D	6.75E+38	6.72E+19	5	LETTERCA	Environmental							
22	2D	6.75E+38	6.72E+19	6	LETTERNA	Interior Appropriations Wildlife							
23	2D	6.75E+38	6.72E+19	7	LETTERLAS	########							
24	2D	6.67E+45	1.61E+22	1	SUBJECT	Natural Resources-HR 3400 Pro							
25	2D	6.67E+45	1.61E+22	2	STATUS	Closed							



Attachments	File folder				
Correspondence	File folder				
Templates	File folder				
1A.tab	TAB File	22,239 KB	No	67,670 KB	68%
1B.tab	TAB File	25,949 KB	No	86,080 KB	70%
1C.tab	TAB File	30,546 KB	No	303,795 KB	90%
1D.tab	TAB File	22,417 KB	No	125,436 KB	83%
1E.tab	TAB File	6,773 KB	No	23,112 KB	71%
2A.tab	TAB File	7,094 KB	No	23,778 KB	71%
2B.tab	TAB File	2,686 KB	No	8,049 KB	67%
2C.tab	TAB File	7,405 KB	No	30,544 KB	76%
2D.tab	TAB File	20,200 KB	No	103,693 KB	81%
3A.tab	TAB File	996 KB	No	3,504 KB	72%
3B.tab	TAB File	65 KB	No	193 KB	67%
3D.tab	TAB File	856 KB	No	4,622 KB	82%
3E.tab	TAB File	654 KB	No	3,464 KB	82%
3F.tab	TAB File	740 KB	No	4,996 KB	86%
4A.tab	TAB File	1,192 KB	No	4,099 KB	71%
4B.tab	TAB File	840 KB	No	2,665 KB	69%
4C.tab	TAB File	1,015 KB	No	3,518 KB	72%
4D.tab	TAB File	2,121 KB	No	21,760 KB	91%
7A.tab	TAB File	1 KB	No	1 KB	52%

CCDT

Home Jessica Tapia Help

Table(s)

+ Create Table(s)

Load Data

1. correspondence belongs to rockefeller	1502899 Records	Upload Files
2. 1A1533936381 belongs to rahall	778287	

CCDT

Home Jessica Tapia Help

2D1533936464 Records

Browse through the records or search here.

Search... Search

id: 267794
Record Type: 2D
Constituent ID:
5468621686987056616665678525766470070127 04
Correspondence ID:
1183534568674566467682
2D Sequence Number: 1

id: 348547
Record Type: 2D
Constituent ID:
6669656936818815701878525769670080706686 6
Correspondence ID:
95781456867456531569
2D Sequence Number: 1

id: 354993
Record Type: 2D
Constituent ID:
2786436943666561664385257946500639156 9
Correspondence ID: 767804568674593862
2D Sequence Number: 1

BUILT BY WVULIBR

id: 348547

Record Type: 2D

Constituent ID: 666965693681881570187852576967008070668 66

Correspondence ID: 95781456867456531569

2D Sequence Number: 1

Text Type: CM

Correspondence Text: 07/24/2012 12:50:21 PM - [redacted] - Letter of support request due Aug. 1. From: [redacted] Sent: Monday, July 23, 2012 1:37 PM To: [redacted] Cc: [redacted] bject: FW: New Program [redacted] sent this asking for a letter of support for their veterans Upward Bound Program. I dont know who this is suppose to go to so I thought I would send to the two of you. Please let me know. Thanks! [redacted] From: [redacted] A [mailto:[redacted]c.edu] Sent: Friday, July 20, 2012 4:41 PM To: [redacted] Subject: New Program [redacted] e are endeavoring to bring a veterans Upward Bound Program to Huntington to assist all of southern WV. Please review the attachments and if the Congressman could do a Letter of Support, we would be honored. Respectfully, [redacted] This electronic message contains information from [redacted] llege which may be confidential or privileged and is intended to be for the use of the individual(s) named above. If you are not the intended recipient, please be aware that any disclosure, copying, distribution, or use of the content is prohibited. If you have received this electronic transmission in error, please notify the [redacted]

in_method: usmail

in_date: 20090202

in_topic:

in_text:

in_document_name: ..\documents\BlobExport\objects\755262.pdf View

in_fillin:

out_id: 9598108

out_type: GENERAL

out_method: usmail

out_date: 20090709

out_topic:

out_text:

out_document_name: ..\documents\BlobExport\formletters\425804.txt View

The Honorable Jay Rockefeller
531 Hart Senate office Building
Washington DC 20510

Dear Senator Jay Rockefeller

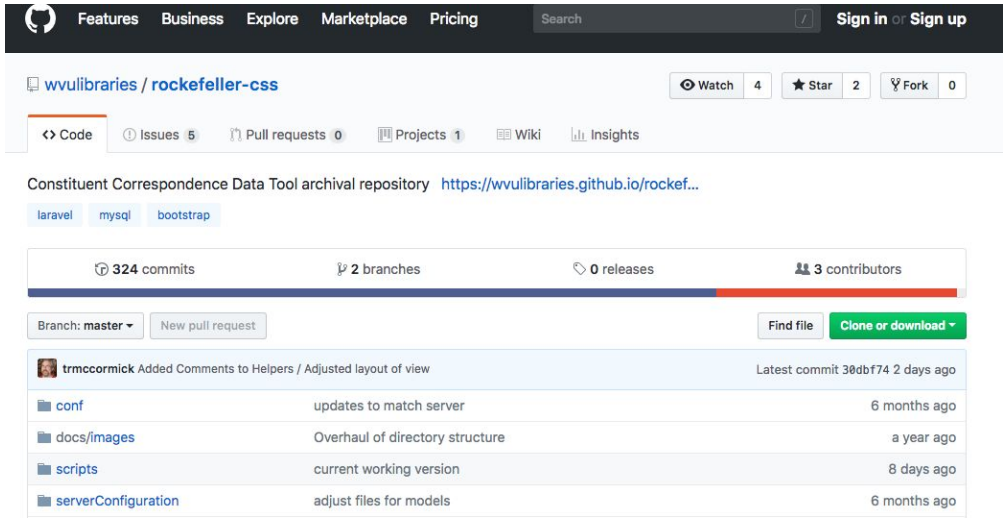
I am please to be writing to you in regards to a Program called TRIO. The fundamental principle of TRIO is to ~~help~~ college students that meet any of these criteria; Being ~~the~~ one person in your family to go to college, your parents have a low income or you have a disability. Not every individual on this earth was born with a silver spoon in their mouth. I and many other students are from a low income background. Once you get accepted into TRIO, you are at a good advantage of being successful. TRIO helps students pay for their books, allow you to print for free and take trips to visit different Universities for people that are interested in Graduate School. One may simply ask "Why are you writing me?" or "What does this have anything to do with me?" As you have been paying keen attention to the news, it stated that there will be major budget cuts. These major budget cuts does not only affect the amount of loan's being disperse to students, but it will have affects on programs such as TRIO, that are funded by the government. If this program ceases to be nomore, ~~students like me~~ it will become very difficult for students such as myself that is the first person in my family to go to college or a single mom struggling ~~to~~ to make ends meets to be successful in college, thus degrading my chances to be a Senator, Doctor, Lawyer or even the

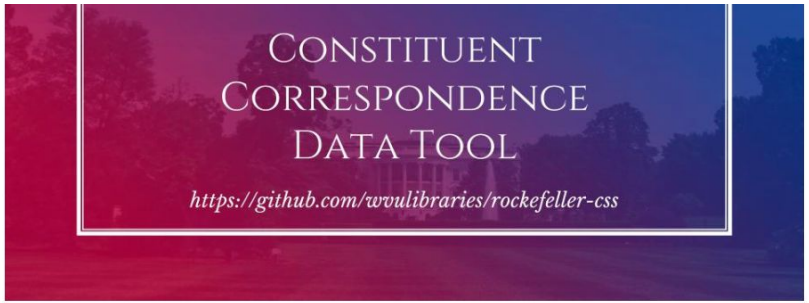
The Honorable Senator Jay Rockefeller
531 Hart Senate Office Building
United States Senate
Washington, DC 20510

Dear Senator Rockefeller,

I'd like to inform you of why TRiO is important, it gives teens like myself, the opportunity to learn about colleges, financial aid, college options and many other things about college life. It gives us the opportunity to grow, learn, and receive individual attention to our situations and what we can do about them. TRiO should recieve an increase in funding because of the outstanding accomplishments of the students and staff enrolled in the TRiO services. Students realise that they are able to become doctors, lawyers, artists and various other people. TRiO has had a major impact on in my life, I am in Upward Bound, a Junior in high school, and UB has helped me realise that I am able to go to college. I am able to become what I want to be, regardless of the income I have. It has helped me be less shy and more outgoing. I am making lifelong friends along the process of learning about colleges, recieving tutoring help in school and many other aspects of my life. TRiO is such a huge part in my life, I would be lost if I had it taken away from me, it's had having such an impact on my life, I can't possibly explain the happiness and joy that I get from being in →

Github Repository





Constituent Correspondence Data Tool is a platform to transform **congressional data** into **information** and **information** into **insights.** The projects aims to provide a holistic interface for importing `flat-files` and provide tools to **research and visualize.**

Status

build passing | Scrutinizer 9.17 | coverage 92 %

Lyrasis Catalyst Grant

“Access to US Congressional Correspondence Data” - $27,000 to complete a feasibility study that will assess and plan for the future collaborative technical infrastructure for an open-source congressional correspondence data access tool, to improve how libraries process and provide access to large data sets with sensitive information and how scholars and the public use data related to Americans’ civic engagement.

Consultant: Jodi Allison-Bunnell

Phase 1: Assess Existing Functionality (10/1-11/30)

- Initial meeting at WVU with congressional papers archivist and systems development department
- Consultant will fully review and document existing functionality
- Consultant will discuss and document potential future functionality, including inherent limitations of the data and systems
- Project planning with congressional papers archivist
- Consultant identifies end users to focus on and desired functionality/technical requirements of the congressional papers community

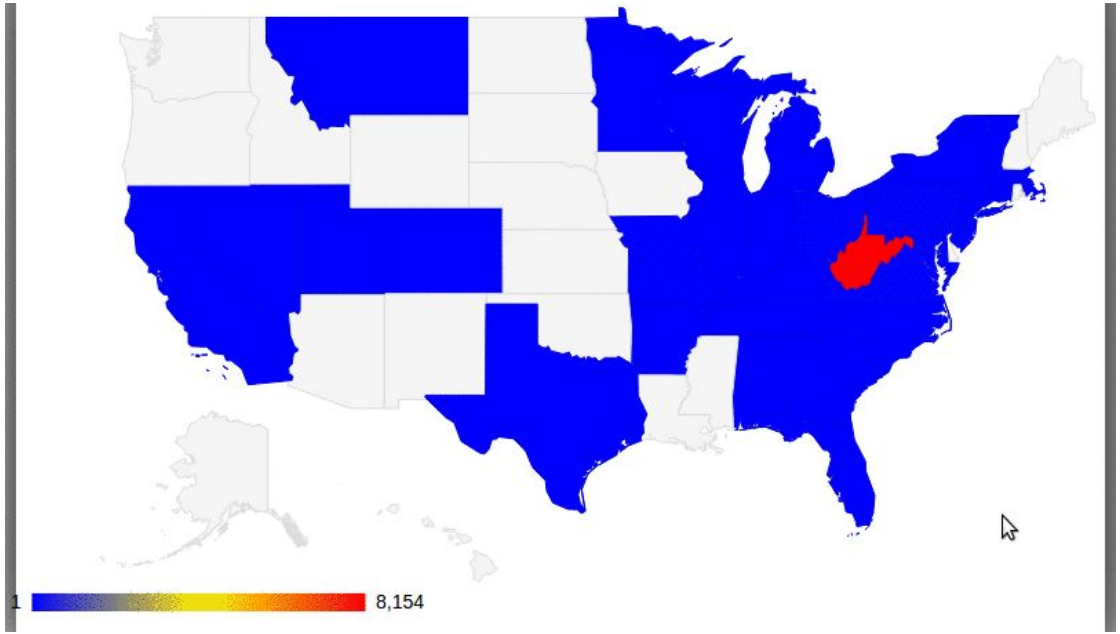
Phase 2: Develop User Requirements (12/1-2/28)

- Based on the plan for engagement created in Phase 1, consultant will work with identified user types to document user needs and use cases for the tool
- Possibilities for engagement may include individual interviews, focus groups, or other options. The consultant will work with members of the advisory board and other stakeholders to organize and support any distributed approaches. Most work will be conducted remotely, with some in-person work as that is financially feasible.
- Consultant will create user personae and/or use cases and test or verify with users
- Consultant will verify personae or scenarios with advisory board

Phase 3: Roadmap (3/1-4/30)

- Based on the user needs identified in Phase 2, and the assessment of the tool's current state, the consultant will characterize the uses of and need for the tool
- In consultation with the development team, the consultant will characterize possible development roadmaps for the tool
- Based on potential uses and development, the consultant will develop possible options for future funding models, administrative home, and administration

Analytics



Q&A