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
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 “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
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

---
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
Assistant Dean of Technical Services

Head of Metadata and Cataloging
  Metadata Librarian
  Metadata Specialist
  Student
  Student

Digital Initiatives Librarian
  Digital Specialist
  Student

Digital Archivist

Organization in February 2018
Library Dean

Assistant Dean of Technical Services

Head of Metadata and Cataloging

Metadata Librarian

Metadata Specialist

Head of Digital Services

Digital Specialist

Organization in December 2018

Student
Student
Student
Student
Student
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 than 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
### CMS Data Interchange Format

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>6.69E+14</td>
<td>1.43E+14</td>
<td>1.14E+14</td>
<td>2</td>
<td>STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>6.69E+14</td>
<td>1.43E+14</td>
<td>1.14E+14</td>
<td>2</td>
<td>STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D</td>
<td>6.69E+14</td>
<td>1.43E+14</td>
<td>1.14E+14</td>
<td>3 LETTER</td>
<td>Lettera [142956]</td>
<td>Meritorious Service Awards These</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4D</td>
<td>5.67E+14</td>
<td>1.55E+14</td>
<td>1.15E+14</td>
<td>4</td>
<td>LETTERCA</td>
<td>Personal Letter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5D</td>
<td>5.67E+14</td>
<td>1.55E+14</td>
<td>1.15E+14</td>
<td>2 SUBJECT</td>
<td>Personal Letter</td>
<td>Kiwanis Award Dinner Ty’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6D</td>
<td>5.67E+14</td>
<td>1.55E+14</td>
<td>1.15E+14</td>
<td>3 STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7D</td>
<td>5.67E+14</td>
<td>1.55E+14</td>
<td>1.15E+14</td>
<td>4 LETTER</td>
<td>Lettera</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8D</td>
<td>5.67E+14</td>
<td>1.55E+14</td>
<td>1.15E+14</td>
<td>5 LETTERCA</td>
<td>Lettera</td>
<td>Kiwanis Award Dinner Ty’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>2 SUBJECT</td>
<td>Judiciary</td>
<td>USA Freedom Act Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>2 STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>3 WEBSUBI</td>
<td>IMA MAIL</td>
<td>on WOMEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>4 WEBREAD</td>
<td>Support</td>
<td>the USA Freedom Act and Limit NSA Surveillance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>5 LETTERCA</td>
<td>Judicial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>6 LETTERCA</td>
<td>USA</td>
<td>Freedom Act Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>7 LETTERCA</td>
<td>Queued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>1 SUBJECT</td>
<td>Environmental-Interior Appropriations Wildlife</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>2 STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>3 WEBSUBI</td>
<td>Fine</td>
<td>Website Contact Form Submission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>4 WEBREAD</td>
<td>As a resident of your district and a supporter of Defenders of Wildlife, thank you for voting a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>5 LETTERCA</td>
<td>Environmental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>6 LETTERCA</td>
<td>Lettera</td>
<td>Interior Appropriations Wildlife</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>7 LETTERCA</td>
<td>Queued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>1 SUBJECT</td>
<td>Natural Resources-HR 3400 Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>2 STATUS</td>
<td>Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25D</td>
<td>2.66E+11</td>
<td>1.48E+11</td>
<td>1.48E+11</td>
<td>1 SUBJECT</td>
<td>Natural Resources-HR 3400 Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>Create Table(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Load Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>correspondence belongs to rockefeller</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1502899</td>
<td>Records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upload Files</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>1A1533936381 belongs to rahall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>778287</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BUILT BY WVULIBRARIES
<table>
<thead>
<tr>
<th>id</th>
<th>Record Type</th>
<th>Constituent ID</th>
<th>Correspondence ID</th>
<th>2D Sequence Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>267794</td>
<td>2D</td>
<td>54686216669870566666567852576647</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>348547</td>
<td>2D</td>
<td>666965693681881570187852576967008</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>354993</td>
<td>2D</td>
<td>278643694366656166438525794650063</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Correspondence Text: 07/24/2012 12:50:21 PM
sent this asking for a letter of support for their veterans Upward Bound Program. I don't know who this is suppose to go to so I thought I would send to the two of you. Please let me know. Thanks.

From: [mailto: ] Sent: Friday, July 20, 2012 4:41 PM To: Subject: New Program
We 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, Corwyn Payne, Military Program Coordinator (304) 535-2219 payneo@me.com
This electronic message contains information from Southwest Community & Technical College 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 Southwest Community & Technical College Office of Human Resources & Employee Development at (304) 535-2219.
Dear Senator Jay Rockefeller,

I am pleased to write to you in regards to the Program called TRIO. The fundamental principle of TRIO is to help college students that meet any of these criteria: being the first 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. And many young students are coming 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 book, allow you to print for free and take for their book, allow you to print for free and take for their book. In addition, TRIO helps students use the resources that are available for them to succeed in college. These programs are designed to help students succeed in college and make ends meet. If this program was funded by the government, it would be a lot easier for students such as myself. It would be very difficult for students such as myself to graduate from college or receive a four-year degree in college. I am the first person in my family to go to college. I am the first person in my family to go to college. TRIO has been a major impact in my life. I am an Honorable Senator Jay Rockefeller.

531 Hart Senate Office Building
United States Senate
Washington, DC 20510

Dear Senator Rockefeller,

I would like to inform you of why TRIO is important. It gives teens like myself, the opportunity to learn about college, 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 situation and what we can do about them. TRIO should receive an increase in funding because of the outstanding accomplishments of the students and staff enrolled in the TRIO services. Students realize that they are able to become doctors, lawyers, artists, and various other people. This has had a major impact on my life, I am in Upward Bound, a junior in high school, and UBS has helped me realize that I am able to go to college. I am able to become what I want to be, regardless of the income I have. TRIO has helped me a lot. I am more optimistic and more outgoing. I am making lifelong friends along the process of learning about college, financial aid, and my college options. Other aspects of my life. TRIO is such a huge part of my life. I would be lost if I had it taken away from me. It has had such an impact on my life. I can’t possibly explain the happiness and joy that I get from being in TRIO.
## Constituent Correspondence Data Tool

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

### Status

<table>
<thead>
<tr>
<th>Build</th>
<th>Linter</th>
<th>Scrutinizer</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>passing</td>
<td>9.17</td>
<td>92%</td>
<td></td>
</tr>
</tbody>
</table>

---

### Code Repository

Visit the Constituent Correspondence Data Tool GitHub repository [here](https://github.com/wvulibraries/rockefeller-css).

**Repository Details**

- **324 commits**
- **2 branches**
- **0 releases**
- **3 contributors**

**Branches**

- master

**Features**

- laravel
- mysql
- bootstrap

**Latest Commit**

- `366bf74` 2 days ago

**Recent Updates**

- `conf` updates to match server (6 months ago)
- `scripts` current working version (8 days ago)
- `serverConfiguration` adjust files for models (6 months ago)
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.
Q&A