America Contacts Congress Project

Report 2: User Requirements and Data Analysis

For the West Virginia University
AB Consulting
2019 March 22

Executive Summary

The *America Contacts Congress* project is a feasibility study of a tool to manage constituent correspondence and case files from the U.S. Congress. As articulated in Phase 1, the focus of Phase 2 (January-March 2019) was user requirements and data analysis to determine the extent to which this data can meet those requirements.

The constituent correspondence forms an important part of congressional papers that may be transferred to a repository after a member leaves office or loses an election. They are part of, but by no means all of, the electronic records that a congressional office generates. And--like analog congressional papers--they have considerable research potential and barriers to access.

Based on a focus group with processing archivists/librarians and one-on-one interviews with quantitative and qualitative researchers, we found that user requirements comprise a short list rather than an expansive one. The archivists and librarians need the tool to do a few things well and to integrate with (and not duplicate) other tools that are already part of their workflows. Qualitative researchers are interested in a tool that eliminates the tedious search process currently associated with analog congressional correspondence. Quantitative researchers are unconcerned about inconsistent data and would like to download CSV or JSON files in order to do their own analysis. In addition, these researchers have strong interest in being able to perform analysis across data sets from multiple congressional offices.

Based on a still very limited data analysis, it appears that the data could meet most of the identified user needs. There is potentially enough consistency to support ingest, keyword search of the metadata, date search, and export for either analysis (by researchers) or curation (by archivists/librarians). Less clear is the capacity for subject browsing or searching the text of attachments. Making the House Interchange Standard and SCDIF data useful for research will require additional data analysis and development.

Researcher interviews clearly showed that this data is valuable and fills identified needs, particularly those of political scientists. Many questions remain--the extent to which the data has been “sanitized” by offices is at the forefront--but it is most likely through access to this data that those questions will be answered.

With the results of this project phase, we have the useful and credible information that will guide informed decision making in Phase 3 and project wrap-up.

Key Terms[1]

Archive Format: A 32-field data export format used for transferring data from Constituent Services Systems to repositories beginning in about 1990.

CSS/CMS: Constituent Services Systems and Correspondence Management Systems. Proprietary systems used in (respectively) the U.S. Senate and U.S. House to receive, store, index, and send correspondence with constituents. Some offices also use these systems for scheduling and casework.

House Interchange Standard: A data export format for CMS data systems. It became an export choice for House offices that are preparing to send their CMS to a repository, but is not as well documented as the Senate format.

Processing Archivist/Librarians: Throughout this report, terms for cultural heritage professionals who focuses on management and curation of unique and special content.

Qualitative Researchers: Throughout this report, refers to faculty and graduate students in history, communication studies, geographical psychology, political science, and allied fields whose primary mode of interaction with sources is search/access/read.

Quantitative Researchers: Throughout this report, refers to faculty and graduate students in history, communication studies, linguistics, geographical psychology, data science, political science, and allied fields whose primary mode of interaction with sources is large-scale and computational.

Repository: Throughout this report, a cultural heritage institution that manages congressional papers after a U.S. Congress member leaves office.

SCDIF: Senate CSS Data Interchange Format. A data export format for CSS systems with over 200 fields. Originally designed for migration between CSS systems, it became an export choice for Senate offices that are preparing to send their CSS to a repository in 2016.

[1] For a more complete list of terms associated with this data and further details, please see the *Archiving Constituent Services Data of the U.S. Congress* report, https://www2.archivists.org/sites/all/files/2017_CSS_CMS_Report_0.pdf.

Findings

Function Decision Points

In Phase 1, we said that in order to fully evaluate the potential relevance, success, and sustainability of the tool, we needed to address four major decision points. The findings from this phase support the following conclusions:

Is the future tool functionality basic or expansive?

For both processing archivists/librarians and researchers, the ***desired tool functionality is very basic***. More important than expansive functionality is that it integrate well with both tools used for curation and tools used for large-scale quantitative analysis. This conclusion echoes the findings of Greene and Meissner: Keep processes as simple as possible and make the data available, even in "messy" form, to researchers.[2]

What data is most critical to the priority users?

The answer to this question depends on whether the researcher is a primarily quantitative or qualitative one.

- ***Quantitative researchers*** feel they ***can do a great deal with the metadata*** exported from the CMS/CSS.
- ***Qualitative researchers*** imagined ***few to no uses for the metadata without attachments***. As a result, a filtering function that allows them to exclude metadata without attachments is important to them.

What is the relative importance of quantitative versus qualitative research?

The answer to this is not either/or. Instead, the researchers we interviewed use a spectrum of tools to work with data like this. They felt strongly that this data can drive research questions for both types of research (or even both simultaneously) through different means. For all audiences, the transformative moment comes with access to the data. Research in this data becomes more attractive for qualitative research because of keyword search. Structured data that is primarily born-digital and large-scale opens numerous doors for quantitative research.

What are the capacities and limitations of the data?

As noted in the *Archiving Constituent Services Data of the US Congress* report, the data at hand is known to be incomplete, inconsistent, and sparsely documented, and that office practices vary widely. Through both interviews and data testing, we wanted to understand how much of a barrier that represented for potential users, particularly researchers. For the researchers we worked with, ***incomplete data was neither***

[2] Mark Greene and Dennis Meissner (*2005*) More Product, Less Process: Revamping Traditional Archival Processing. The American Archivist: Fall/Winter 2005, Vol. 68, No. 2, pp. 208-263.

particularly surprising nor a significant barrier. As you would expect, they would prefer to have complete documentation of the processes by which data was eliminated, but a lack of documentation does not generally render the data useless. In particular, quantitative researchers are accustomed to working with less-than-perfect data and routinely perform remediation themselves or hire others to do it. They will tolerate missing attachments in up to 25-30% of the data. Qualitative researchers found incomplete data less useful, but felt generally that the gains in search and retrieval were significant enough to outweigh concerns about integrity. ***So long as the tool can ingest the data, resolve the structure, and provide basic search, it should render this data useful for research to an unprecedented degree.*** Fully exposing the potential for enabling some of the most enticing research of congressional office practices, responsiveness, and constituent influence will require further analysis of much more CMS/CSS data.

User Requirements for Highest Priority Groups

Processing Archivists/Librarians	**Qualitative Researchers**	**Quantitative Researchers**
Ingest data that is 1 GB and up efficiently Search and browse by date/date range Search and browse by subject Keyword search the CSS/CMS data (and, less critically, attachments) Generate reports on the status of the data Integrate with other tools for curation	Search and browse by subject Search and browse by date/date range Keyword search attachments Filter results to limit to records that have attachments	Aggregate and access large amounts of data across collections Export structured datasets in CSV or JSON for analysis in other tools Access attachments as text files for analysis Allow them to perform data cleanup

Data and Ability to Meet Requirements

At a high level, the Archive format or SCDIF data can meet these requirements, as follows:

- Subject search/browse: Yes and no. Identifying and searching subject fields only should be quite feasible. Browse of subjects is doubtful given the degree of inconsistency.
- Date search/browse: Yes. Dates are generally present in a normalized format (YYYYMMDD).
- Search attachments: Maybe. The data sample was insufficiently large to determine what percentage of attachments are or could be rendered searchable.
- Filter by presence/absence of attachments: Yes, so long as we can ensure during ingest that a larger percentage of attachments will retain their connection with their associated metadata.
- Choose to search single set or across multiple ones: Technically likely, but administratively thorny. Privacy and institutional relationship concerns may make this difficult to negotiate, but with the right parties and agreements could be possible and very valuable for researchers.
- Export data as CSV or JSON for large-scale analysis: Yes. Since the data is already structured, this seems quite feasible so long as administrative concerns about personally identifiable information can be addressed.
- Integrate with tools for curation: Yes. This need and feasibility is closely parallel with the ability to export CSV.

Data and Research Interests of Interviewees

Looking broadly at the potential research value of the data compared with the stated research interests of the interviewees, it appears that the data could support quantitative and qualitative inquiries in the following ways:

- Congressional staffing and the revolving door with lobbying firms:
    - ***Significant potential in SCDIF and HIS, but not in Archive format.*** The tool could increase ability to search efficiently on staff names and to identify the specific roles of individuals so long as the SCDIF or HIS 8A table is intact. This information is not present in the Archive format.
- Large-scale analysis of broadcast communications from congressional offices (and stated desire to have a comparative data set):
    - ***Considerable potential in both formats.*** All three can support analysis of correspondence received and sent during a specific time period and/or on a topic and to compare with broadcast communications from that same office.
- Congressional leadership:
    - ***Little Potential in any format***, as leadership decisions are not likely to be documented in constituent correspondence. If the set is SCDIF and the office used the 7 tables for schedule data, or HIS used 7A table for schedule data, the potential could increase somewhat.
- Legislator responsiveness and representation:

    - ***Considerable potential in all formats.*** This data (Archive format, SCDIF, and HIS) fills a known gap in knowledge of the relationship between constituent communications and legislative action.
- Congressional office management of constituent communications:
    - ***Considerable potential in SCDIF, some in HIS, but not in Archive format.*** SCDIF data that has the 1, 2, 3, and 8 table present (and, for completeness, the 6 table) and with some data in the 1A, 1B, 1F, 2A, 2C, 3A, and 8A fields can support significant insight into office processes. In the HIS data, table 3A identifies staff working on casework files.
- Congressional history, both political and social:
    - ***Considerable potential in all formats.*** The ability to search congressional correspondence by keyword, subject, and date rather than manually reading through files arranged chronologically is a significant advance that could vastly increase the use of these underutilized resources. However, this potential depends on the ability to identify, search, and read the attachments.

Selection of End Users for Phase 2 Focus

We selected end users for special focus during this phase based on the priority ranking we developed during Phase 1: Processing archivists/librarians, Quantitative Researchers, Qualitative Researchers, and IT Staff. Although the Phase 1 ranking for students was the same as that for IT staff, we chose them rather than students for special focus in Phase 2 in order to split our inquiry equally between internal and external users. We also felt that by including graduate students in the Researchers category, we were focused on the most likely users for this complex data.[3]

Processing Archivists/Librarians

Methodology

During the workshops we held during Phase 1, we posited that the needs of processing archivists and librarians were:

Processing Archivist/Librarian	Cultural heritage professional who focuses on management and curation of unique and special content
Scenarios: Prepare collections for researcher use by arranging and describing. Protect rights of privacy and other restrictions placed by agreement or statute. Ensure data is authentic.	

[3] For more details on rankings of potential users, please see the Phase 1 report: https://wvrhc.lib.wvu.edu/collections/physical/congressional/correspondence.

Top Functions:	Constraints:
● Ingest data ● Identify PII ● Search by date/date range ● Browse by date/date range ● Document any changes to data	● Time: Low ● Resources: Low ● Expertise: Very High

Our task during this phase was to test these hypotheses. The group was composed of individuals with curatorial responsibilities for CSS/CMS data from members of Congress with a focus on ingest, description, and processing. AB Consulting developed the recruiting materials, scripts, and visual materials and moderated the sessions. Members of the Advisory Board recruited subjects from members of the Society of American Archivists Congressional Papers Section who were not among those who participated in the November workshops and served as note takers and additional interlocutors during the focus group. AB Consulting conducted short pre-interviews with all individuals who volunteered to be part of the focus group and selected five individuals for participation. Two individuals who expressed interest were not selected: One lacked hands-on recent experience with CSS/CMS but their comments were summarized in writing; the other found that their institution holds other electronic records from a member of Congress, but that they are not CSS/CMS.

The focus group participants were four archivists (either political papers or digital initiatives) and one digital preservation librarian. They represented three academic institutions, all of whom have, or will very shortly receive, CSS/CMS data. Fortuitously, two institutions were able to offer a pair of individuals: an archivist and a digital preservation/digital initiatives specialist. The one-hour focus group worked on two exercises:

- The first focused on prioritizing tool functions for managing, preserving, and preparing data sets for research use.
    - One day before the focus group, participants received a list of potential functions most likely to be important for processing archivists/librarians. Our aim was to test those assumptions and to possibly narrow the number of functions that need to be developed or enhanced.
    - Focus group participants worked together on a virtual whiteboard to categorize functions into “Must,” “Should,” and “Could” categories. (For a screen capture of both the function priorities and the functions that were not a priority, please see Appendix B.)
- The second asked participants to envision where the tool is likely to fit into their existing curatorial workflows for collection management, ingest, preservation, and access. This exercise was designed to test how expansive the tool needs to be and to identify the extent to which potential functions are already performed by other tools. Because of

unanticipated time constraints and a tornado warning that required two participants to leave halfway through the session, focus group members addressed the final areas of inquiry by email after the session.

All focus group participants received a $25 gift card in consideration of their expertise and time. For recruiting and focus group materials, please see Appendix B.

Findings

Overall, we found that the needs of this group were both consonant with and very different from what the workshop participants perceived. We found that the constraints we envisioned for the subjects were correct: they are very time constrained; long closures on collections make it difficult to prioritize work on them; and while they have very high expertise in many areas, relatively little exposure to this data means that it is difficult for them to fully envision their needs.

The most compelling difference between our hypotheses and the results of the focus group was the number of things the tool needs to do. Subjects specifically cautioned against building a tool to "do everything," which could have a number of negative effects: creating redundancies or conflicts with tools already in use that perform the same or similar tasks; increasing the complexity of development and implementation; and decreasing the effectiveness and likelihood of adoption. Instead, the toole needs to perform a few functions well and integrate with other tools that are part of repositories' ingest, processing and preservation workflows.

Specifically, the tool must:

- Ingest massive quantities of Archive format or SCDIF data--1 GB and up--efficiently and without being overwhelmed by the size;
- Search and browse by date;
- Search and browse by subject;
- Keyword search the CSS/CMS data;
- Generate reports related to these beginning stage ingest processes (e.g. to confirm the integrity of the tool and the data in the tool), particularly any changes that occurred on ingest.
- Integrate with other tools.

The group felt that the tool should also search the text of attachments, but that this was not as high a priority as searching the CMS/CSS data. Additionally, the tool could search by media or file format of attachments, but that was in the "could" category.

The reason for this significant difference was revealed very clearly in the second exercise, in which the group looked at where this tool was likely to fit into curatorial workflows and what other tools it should integrate with. When the group looked at a potential workflow diagram

with four basic categories (ingest, collection management, preservation, and access), they found as a group that the following tools and tasks formed useful groupings:

- Ingest
    - Functions: digital forensics, space utilization, cleanup; opening data sets; scanning files and generating report; identifying PII; initial checksum; virus scan; file format identification
    - Tools: BitCurator and its modules; FITS, JHOVE, DROID, bulk extractor, and others
- Process
    - Functions: accession, describe at the collection or item level
    - Tools: Collection management (including, but not limited to, ArchivesSpace, Archivist's Toolkit, Archon)
- Preserve
    - Functions: data packaging, generate PREMIS metadata
    - Tools: Bagit, Preservica, Archivematica
- Output
    - Functions: Facilitating staff or end user access; creating Dissemination Information Package (DIP)
    - Tools: Archon, ArchivesSpace, Archivematica

The group as a whole expressed a strong need for the development of this tool and vouched for its uniqueness. Additionally, the group expressed a preference for an open-source tool that would integrate with other open-source tools, be built sustainably, and be a collaborative effort.

So, to return to and update the Processing Archivist/Librarian persona based on these findings:

<table>
<tr><td>Processing Archivist/Librarian (Updated)</td><td>Cultural heritage professional who focuses on management and curation of unique and special content</td></tr>
<tr><td colspan="2">Scenarios: Prepare collections for researcher use by arranging and describing. Protect rights of privacy and other restrictions placed by agreement or statute. Ensure data is authentic.</td></tr>
<tr><td>Top Functions for Tool:

● Ingest data that is 1 GB and larger efficiently
● Search and browse by date/date range
● Search and browse by subject
● Keyword search the CSS/CMS data (and, less critically, attachments)</td><td>Constraints:

● Time: Very Low
● Resources: Low
● Expertise: Very High</td></tr>
</table>

• Generate reports on the status of the data • Integrate with other tools for curation	

Qualitative and Quantitative Researchers

Methodology

We chose interviews for the potential users who are time constrained and difficult to schedule: the faculty and graduate students who make up the Qualitative and Quantitative researcher groups.

Advisory Board members engaged deeply in recruiting so that we would work with subjects who were already engaged in and passionate about congressional correspondence and other closely related data as a central part of their research. The members compiled a list of likely candidates and sent personal invitations to fourteen of those individuals. We also posted invitations to the listserv of the Association of Centers for the Study of Congress. Once we received responses to the screening survey, AB Consulting reviewed the individual's online presence; did fifteen-minute screening interviews with each individual to discuss their research, sources, methods, and availability; and scheduled one-hour interviews.

AB Consulting led the interviews and followed a script. One Advisory Board member attended each interview to take notes and to ask clarifying questions. We did not make audio or video recordings.

All interview participants received a $25 gift card in consideration of their expertise and time.

For recruiting and interview materials, please see Appendix D.

Subjects Interviewed

Quantitative: Four subjects

- Two professors (political science)
- Two PhD candidates (political science)

Qualitative: Six subjects

- One administrator (historian)
- Three professors (political science)
- One PhD candidate (political science)

- One government agency historian

We did not collect demographic information, but perceive from online CVs, publications, and appearance that we had a fairly evenly distributed age range of 20-50 and and a 6:4 ratio of males to females.

The majority of the subjects--eight out of ten--were political scientists who use a mix of qualitative and quantitative methods in their research. We were only able to recruit two historians. Neither of them were graduate students, and both were in administrative positions rather than research faculty. This is unsurprising given the general perception that congressional collections--and correspondence in particular--are a seriously underused resource in a number of disciplines.[4]

A significant number of the interviewees were former congressional interns who had interacted with CMS/CSS systems during their time on the Hill, either in the 1990s or within the last ten years. They showed evidence of understanding how these systems are used and the origins of data inconsistency.

Summary of Research Interests and Methods

All subjects had either direct experience with congressional correspondence or with closely related data sets. All use sources that were originally created for some other purpose. The primarily qualitative researchers were more likely to work in analog congressional collections, agency communications, and surveys/interviews. The primarily quantitative researchers were more likely to use sources that are readily available online, like disbursements, newsletters, staff directories, or official biographies. Although we classified subjects as qualitative or quantitative researchers, most were using a mixed toolset.

Research interests included:

- Congressional staffing and the revolving door with lobbying firms;
- Large-scale analysis of broadcast communications from congressional offices;
- Congressional leadership;
- Legislator responsiveness and representation;
- Congressional and policy history;
- Congressional office management of constituent communications.

Most of the subjects perceive or know from direct experience that constituent correspondence is a significant source for their work. Both the barriers to access to the analog collections and the

[4]There are no archival studies quantifying the use of congressional collections or change in use over time. Perceptions are based on anecdotal evidence. On historians preferring "less bulky" sources on Congress to large congressional collections, see Mark A. Greene, "Congressional Records at the Minnesota Historical Society," in *An American Political Archives Reader*, p. 183. On the nature of congressional collections and doubts cast on research value, see Linda A. Whitaker and Michael Lotstein, "Pulling Back the Curtain: Archives and Archivists Revealed," in *Doing Archival Research in Political Science,* 108-111.

lack of availability of the CMS/CSS data leaves a "black box" for many topics of inquiry. They make do with other sources, but feel the limitations of those sources.

Data Samples

Subjects first saw an example of a fairly complete record (the incoming letter, the CMS/CSS metadata, and the outgoing letter). The interviewer talked through some key features including subject, date, and the nature of both the incoming and outgoing letters and invited the subject to discuss and react to the sample. Subjects then saw an incomplete record: the CMS/CSS data only with no attachments and a subject field that contained only a code with no indication of meaning.

The subjects had strong positive reactions to the complete data sample. The metadata was of interest, particularly ZIP codes. As a whole, access to the complete package (incoming plus metadata plus outgoing) was intriguing because paper collections are rarely that complete. It's essential to have the attached files as text (therefore searchable), or being able to re-render a PDF as text. PDFs are OK but more difficult to work with unless there is a text layer.

Quantitative researchers are not bothered by incomplete data up to a point: Interviewees perceived that even if a good portion of the data was incomplete, they could infer some trends and match them against other data sets. The number of requests each office receives is very much tied to responsiveness. Even mystery subject codes could be useful since they might be able to figure out the code.

Qualitative researchers--the two historians and two of the political scientists--found little or no value in just metadata without the attached correspondence.

Reaction to Draft User Persona

During Phase 1, we posited that the needs of quantitative and qualitative researchers were:

<table>
<tr><td>Qualitative Researcher</td><td>Faculty and graduate students in history, communication studies, geographical psychology, political science, and allied fields</td></tr>
<tr><td colspan="2">Scenarios: Primary needs are search/access/read; manual search interface is the primary entry point; similar to working with paper materials</td></tr>
<tr><td>Top Functions:

● Subject search
● Search by date/date range
● Subject browse
● Browse by date/date range</td><td>Constraints:

● Time: May vary, but generally motivated and tenacious
● Resources: Low to moderate
● Expertise: Very high</td></tr>
</table>

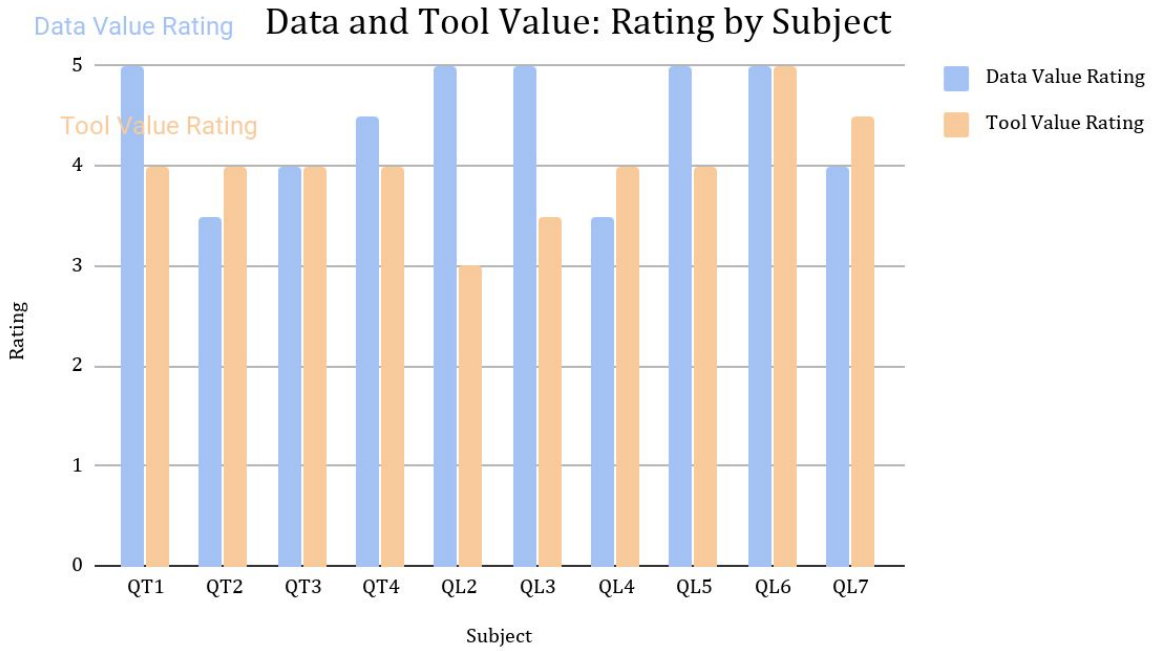
• Filter results	

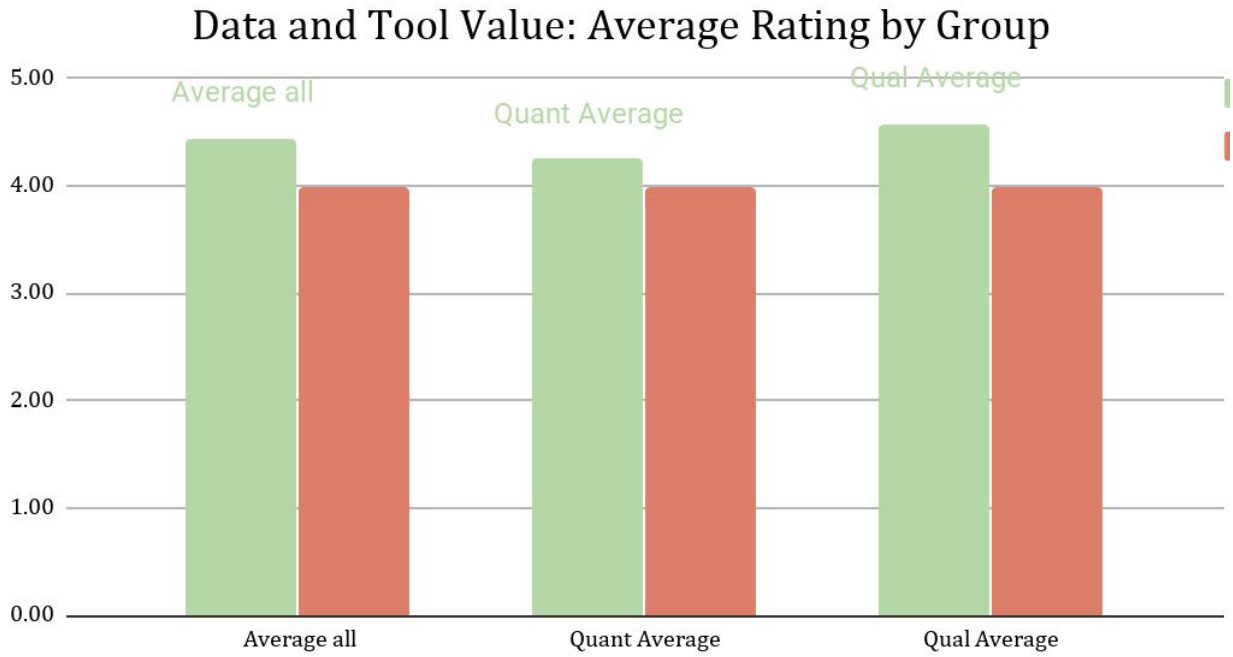
Quantitative Researcher	Faculty and graduate students in history, communication studies, linguistics, geographical psychology, data science, political science, and allied fields.
Scenarios: Big data people who use emerging technologies and data science to do research that is not feasible from analog records. Entry point is programmatic. Aim and expertise is to decipher patterns not discernible at any other scale. Create outputs that may include text, visualizations, or cutting edge approaches to presentation.	
Top Functions: • Aggregate and access large amounts of data; • Export structured datasets based on search/browsing results • Search and analyze CSS/CMS data • Combine natural language and programmatic queries to explore the data • Sort data	Constraints: • Time: May vary, but generally motivated and tenacious • Resources: Low to moderate • Expertise: Very high

In the next section of the interviews, subjects were asked to react to the scenario and function that corresponded to the type of researcher they are. Without significant exceptions, the subjects responded that the scenario and needs that most closely resembled them were complete and correct.

Perceived Data and Tool Value

Last, subjects were asked to rate the importance of the CMS/CSS data and the tool for their research on a five-point scale. Subjects perceived high value in both with no significant differences between qualitative and quantitative researchers.





Findings

Findings for Both Researcher Categories

- The CMS/CSS data fills a gap in information that is both important and timely: Response to constituents compared to other influences. Currently, the political scientists who study this rely on agency requests from congressional offices, franking, trips, and newsletters. None of these are ideal sources, but they are available. They know that there are significant differences between offices, but the relationship between constituent opinion/requests and actions taken by the office is currently a black box.
- Most subjects perceive that congressional correspondence is an underutilized resource for both political science and history. They cite changes in both fields, the perception that congressional collections are only valuable for political history, and challenges with access.
- Overall, subjects expressed great excitement about the data and the gaps it has the potential to fill. This data is potentially transformative not only for their work, but that of the field by reviving some topics that were abandoned decades ago. However, the value of data is somewhat unknown until researchers have access to it (is it all just form letters? Do the offices sanitize to such a degree that what's left is useless? How has it changed over time?).
- Almost everyone commented positively about having the incoming letter, CMS/CSS data, and the outgoing letter, and particularly observed that completeness isn't always present in the analog collections. One researcher (who has extensive experience in analog congressional collections) described themselves as "flabbergasted."
- Value of CMS/CSS data: The metadata is useful for narrowing down a search beyond just keyword. There was a strong interest in analyzing ZIP codes and the dates that correspondence was received and responded to.
- All subjects were enthusiastic about the project and the data--not surprising to hear from this self-selecting group, but heartening nonetheless.
- All subjects had insightful comments on availability versus quality of data, and how much weight they give to making deliberate choices about what sources they choose. Certainly it's a long-documented information seeking behavior to choose accessible sources over higher quality sources, and not unique to this user group.
- More than one also thought that the presence of a tool would entice congressional offices to take more care with constituent correspondence and to be more likely to place it with a repository.

Findings for Qualitative Researchers

- Those who have used analog congressional papers cite the bulk and arrangement and description of correspondence as a major barrier to research. The prospect of even keyword search of the metadata is transformative. They want structured search, retrieval, and access, seeing the potential to make what is now tedious and

time-consuming much more effective and efficient. They perceive that these materials, currently an underutilized resource, would be used more with a search function available.
- They were more likely to be satisfied with working on one data set at a time. Some want to search across multiple collections (or had even assumed that was the unspoken intention).
- Comparing the CMS/CSS metadata and the attached correspondence, they felt that the attached correspondence was what they found most valuable. They had difficulty imagining uses for metadata without attachments.
- The perception from the workshops that this group primarily needs to search, access, and read appears to be correct.

Findings for Quantitative Researchers

- Both the CMS/CSS metadata and the attachments are useful to this group. Some felt that metadata alone (with no attachments) is also useful to do analyses based on date and ZIP code.
- Incomplete data is unproblematic. So long as the missing data is random, they can infer at a high level from what remains. One subject said that as little as 20% completeness would yield a valid analysis if you compared it to parallel data.
- Inconsistent data is also a non-issue for this group. They are accustomed to either normalizing data themselves using tools like OpenRefine, or contracting with an organization (ICPSR, ProPublica, and others) to do it for them. Graduate students particularly stated that pooling resources to hire out data cleanup is very common.
- These researchers are more interested in data export as soon in the process as possible and will do analysis, cleanup, and other functions in a different tool. They do not want the tool to do any of the analysis. Their preferred formats are CSV and JSON, but the only firm requirement is that the data be text and that attachments are either plain text or PDFs with text layers. They could approach the data through an API or write code to scrape a web interface.
- Most subjects were interested in working across data sets from multiple offices.

For more details on subjects' responses and summaries, please see Appendix D.

Taking these findings and modifying the Phase 1 user personae yields the following results:

Qualitative Researcher (Updated)	Faculty and graduate students in history, communication studies, geographical psychology, political science, and allied fields
Primary needs are search/access/read; manual search interface is the primary entry point; similar to working with paper materials	

Top Functions: ● Subject search/browse metadata ● Date/date range search/browse metadata ● Keyword search attachments ● Filter results to limit to records that have attachments	Constraints: ● Time: May vary, but generally motivated and tenacious ● Resources: Low to moderate ● Expertise: Very high

Quantitative Researcher (Updated)	Faculty and graduate students in history, communication studies, linguistics, geographical psychology, data science, political science, and allied fields.
Big data people who use emerging technologies and data science to do research that is not feasible from analog records. Entry point is programmatic. Aim and expertise is to decipher patterns not discernible at any other scale. Create outputs that may include text, visualizations, or cutting edge approaches to presentation.	
Top Functions: ● Aggregate and access large amounts of data across collections ● Export structured datasets in CSV or JSON for analysis in other tools ● Access attachments as text files for analysis ● Allow them to perform data cleanup	Constraints: ● Time: May vary, but generally motivated and tenacious ● Resources: Moderate to high ● Expertise: Very high

Data Testing

Methodology

In the Phase 1 report, we noted that the tool's basic functionality was based on just two data sets (one Archive format, one House Interchange Standard) and that we needed to test a broader set of data in order to understand the target data and what is—or is not—possible to do with it. This

is a shift from the original project plan (which called for a primary focus on user testing in Phase 2) to an approximately equal focus of user testing and data testing.

Members of the Advisory Board began to set up data testing in December and continued through the end of February. The process was intentionally narrow in scope to ensure that the process yielded the information that the project needed and did not inadvertently result in either a drain on members' time or evaluatively testing an emergent tool that is not ready for that.

Testers installed the tool based on the available instructions and with assistance from the developers at West Virginia University. They loaded one or more data sets into the tool and filled out a simple form to record the following information:

- Vendor/system used in original office (if known)
- Name of set (using an alias if needed)
- Approximate size of set (in MB, GB, or TB)
- Approximate dates associated with set
- Format of data set (flat file, CMS output, Archive Format, SCDIF, HIS)
- Whether the respondent was able to ingest the data set with the tool
- What the column headers were (e.g. name, date, subject)

The intent of data testing was to characterize:

- The size of data set that the tool will need to be able to ingest and search;
- What formats existing data sets are in;
- How the data is structured. This is a particularly critical point; evidence to date suggests that the data structures in this material are very inconsistent. In order to meet potential user needs (e.g structured search, large-scale data analysis, efficient normalization/de-duplication/redaction), we must be able to discern data patterns.

Together with the user testing, data testing allows the project to more accurately characterize the relationship between user needs and the data's capacity to meet those needs.

Data Sets Tested

Data Set Name	Size	Creation Dates	Format	Notes
Senator Harry Reid	1.1 GB (Tables); 31 GB (Attachments)	2010 - 2016	Senate CSS Data Interchange Format (SCDIF)	Reid was the first senator to export data as SCDIF. This CSS data has had at least 6 major updates. After a major update in 2010, some of the data migrated to the new version did not copy over cleanly, and some correspondence cannot be found or identified accurately.

				Challenges with data ingest: the data included a voluminous set of attachments. Working to get said data associated with tables has raised some logistical challenges. The only manner to move any in is through the back end. However, even then it has taken work with the commandline to move the files, given that most of the directories have a quarter million or more files. As a result, I'm not sure that I was able to view the data to its full extent, though I was able to ingest all of the tables.
Senator Saxby Chambliss	1.31 GB; 32.4 GB (attachments)	2003-2015	Flat file, Archive Format (32 field)	Tables were so large that uploading them using the tool's web interface was not an option. The .dat flat files had to be uploaded via command line (ssh), or using an ftp client. Similar situation with the attachments, with over 5 million files, these had to be uploaded on the backend. This is not inherently a problem, in my opinion. Because of the multi-level directory structure of the attachments, the tool was not able to establish clickable "view" links between all data records and their associated attachments. Flattening the structure of the directory containing the attachments to 1-2 levels seemed to help with this. Overall responsiveness of the tool's web interface is slow when trying to navigate through search results, presumably due to the size of the tables.
Senator Jay Rockefeller	437.9 MB; 40 GB (attachments)	1995-2015	Flat file, Archive Format (32 field)	Gave 8 or 16 GB of RAM and a II core processor; took 1-2 hours to import; 1,502,899 records; 1,301,592 files (attachments)

Congressman Nick Rahall	842 MB (tables); 1.3 GB (attachments)	?-2015	House Interchange Standard	23 tables imported (but should be 25); 11,067,185 records; 300,526 files

Findings

The data testing findings are preliminary since Advisory Board members were only able to test four sets (two of which--Rockefeller and Rahall--had already been used to build the tool). We had expected to test at least two more, but that was not possible. With the 2017 Congressional Papers Section report documenting 15-21 repositories with this data at that time--and knowing this number was augmented considerably with the outcomes of the 2018 election--this represents a very small portion of the data sets that exist in, or are in the process of being transferred to, repositories. We cannot regard this as a representative sample, but it is an advance in tangible knowledge of both Archive format, HIS, and SCDIF.

Testing confirmed the following:

- These data sets are indeed massive, with most sets' attachments exceeding 30 GB. Sets contained over a million metadata records.
- Ingesting the records takes considerable computing power and some time.
- Advisory Board members had to have technical expertise to ingest the files, working with command-line and file structures to associate the metadata and the attachments.
- The flat file Archive Format data sets were generally unproblematic with little to no variation in table structure and field usage. Linking the metadata and attachments was more challenging, but yielded results after some restructuring.
- The SCDIF data from Reid, and the HIS data from Rahall, are complex. The multi-table format from the relational database offers enticing possibilities for understanding congressional office operations at an unprecedented level, but size and complexity also present barriers. The Rahall data had evidence of fields that were re-named by the office: Person ID became Constituent ID, for instance. More data testing could reveal some patterns that would enable the tool to work with these variations.
- The Rahall data in particular is close enough to the SCDIF standard to provide value; the variations in use of and renaming of fields are within reason to accommodate. The Reid data is less certain as the Advisory Board member was unsure that he was able to view the data to its full extent.
- The SCDIF data may be consistent enough to support the types of research our interviewees envisioned. However, this will require additional development, as the tool was built for the (much simpler) Archive Format.

For more information, including detailed analysis of each data set against either the Archive format or SCDIF, please see Appendix A.

IT/Systems Survey

Methodology

During the Phase 1 work, we both identified Systems/IT administrators as an important user group and decided to conduct a brief survey and focus group with Systems/IT administrators from each Advisory Board member institution. Our reasons for engaging with this group were twofold:

- Any institution that implements the tool will need to include their Systems/IT administrators in that process. The developers at West Virginia University asked that we confirm some basic elements of infrastructure, institutional constraints, and preferences so that they could ensure that they are developing the tool in a way that is likely to work well for implementers.
- One of the future options for further development could be open source collaboration. We need to know the level of enthusiasm for engaging in co-development.

We sent a survey to five institutions in December and asked for responses by early January. A summary of the answers:

- These libraries have a relatively high level of control over their IT resources--either sole management of the vast majority of resources or close to it--and are able to make choices of operating systems and software.
- Most have already, or are in the process of moving to, more cloud-based and virtual machines rather than on-premises hardware.
- Preferred server platforms are mostly Windows Server, RedHat Linux, or Ubuntu.
- The preferred virtualization platform is VMware.
- Only two institutions answered the question about containerization platforms, but both use Docker.
- The most common programming languages that institutions could support were Python and PHP.
- The desktop operating system most preferred was Windows.
- Network security requirements vary to some extent among institutions.
- Four of five institutions are required to use only applications that are accessible, with some variation on stringency (e.g. whether there are provisional approvals, whether VPAT 2.0 is required).
- Two institutions indicated that the potential RAM and disc space associated with the tool was a non-issue. Others indicated that they had some limitations, and would certainly require a compelling use case.
- Most institutions characterized themselves as neither risk takers/early adopters or risk averse/late adopters, but somewhere in between.

The survey data provided some value despite the very small sample size. However, our attempts to follow up with the individuals were less than fruitful. Ultimately, the project director and AB Consulting (with affirmation from the Advisory Board) decided that neither our development or implementation plans are concrete enough to have this discussion at this point in the project. Clearly, and within certain constraints, personnel who manage congressional data who need a particular or specialized tool must make the case for their needs with their IT/Systems personnel, and how effectively they advocate is the most impactful factor in those decisions.

For more details on the IT/Systems survey responses, please see Appendix C.

Appendixes

A: Data Testing Materials

Instructions
Survey
Responses
Analysis of Responses

B: Processing Archivist/Librarian Focus Group Materials

Recruiting and Other Messages
Screening Survey
Focus Group Script
Focus Group Slides
Functions Visuals
    - Prioritized
    - Not Prioritized
Workflows Visual

C: Systems/IT Survey Materials

Survey and responses

D: Quantitative and Qualitative Researcher Interviews

Recruiting and Other Messages
    - Qualitative
    - Quantitative
Screening Surveys
    - Qualitative
    - Quantitative
Interviewees
Interview Script
Interview Slides
    - Qualitative
    - Quantitative
Notes and Analyses