Digitizing History: Collaborating to digitize the United States Congressional Serial Set - Transcript

Please stand by for realtime captions.

Thank you for joining us to learn more about the collaborative effort today we've got four speakers, James Sweeney, Chris Berman, Bailey DeSimone, Elaine Lee. And Heidi and providing you with updates for now over to J for this presentation

---

Thank you and good afternoon thanks for joining us at this session James Sweeney from the Library of Congress pick grateful to be working with our partners for the publishing office to share with you our offer to digitize the U.S. Congressional serial set. After my introduction we will hear from my law library colleagues beginning with Chris who will walk us through the evaluation and workflow process that we've employed. Elina will show us what she's developed. To track the inventory and digitization process. Bailey would then escort us to this essential work of our procedures. Based on the documents we are discovering. Providing the public with easy online access to the complete U.S. Congressional serial set. Because of the legislative and historical significance is a high priority for the law library. The serial set is an official bound collection of official reports and documents from the house and senate of the U.S. Congress. It's contents include detailed information on the wide array of subjects as varied studies for wage and price is, immigration, women and child labor, unemployment, national security, conduct of war and civil rights. Key historical initiatives are also documented which provide riveting contemporary accounts of many efforts that changed our nation. The law library of Congress in collaboration with the U.S. government publishing office has launched a large multiyear effort to digitize and make accessible the roughly 60,000 volumes at the serial set dating back to the first volume published in 1817. The library is responsible for digitizing the volumes and GPO will provide the metadata, catalyzing each document and authenticating the digital files. The digitized serial set will be available to the Library of Congress and the GPO websites. Possibly Congress.gov, which we are exploring. We are well underway with a full digitization project. During this past fiscal year. Law library staff and contractors have reviewed serial volumes for completeness and condition and tracked volumes to the review process. With the goal of preparing a minimum of 875 volumes per quarter for shipment to the vendor. For quarters one and two, targets were exceeded. As a result of the pandemic and the limiting of staff on campus, performance, staff had to curtail the preparation with the volumes and beginning on July 20 the loss serial team resumed work at library to prepare and finalize another shipment of over 500 serial set volumes for digitization. We anticipate the next shipment of 600 volumes will occur in mid-November as of today our team has evaluated and selected 2693 volumes for digitization. For the current fiscal year I started earlier this year. The goal is to prepare a digitize another 3000 volumes and working with our GPO partners to publish a subset of the serial set is additional or a library project one collection on the Library of Congress website. Now, to explain how all this has been made possible let me turn to my colleagues beginning with Chris.

Hello hot this is Chris I'm going to describe the process that we go to for evaluation and some background information. As mentioned we are digitizing the 15th Congress to the 103rd Congress volumes. It's over 15,000 volumes roughly 12 million pages. There's larger volumes such as Atlas and many of the serial set volumes contain foldout maps or diagrams and also oversized maps. The process is to evaluate within the library's collection for digitization. Has up to six copies of each serial set number in our collections and through this process we inspect physical volumes for completeness. To do so we use the 1909 index created by the GPO and the schedule of documents to make sure that every document is within the serial set volume. We record any physical characteristics such as fold outs and maps and color plates or oversized maps and architectural drawings. If none of the on-site copies are complete we request volumes from off-site and is not the volumes within the collection are complete we send a request to GPO to see if they have the volumes in their collection. And if none of the volumes are complete we create a complete serial set number from multiple volumes. For digitization there digitizing to the performance level. We are creating each serial set volume with a digital object of master file which is a JPEG 2000 OCR which we use XML. A thumbnail for presentation which is a GIF and each document and volume is PDF. Each serial set volume is identified with a unique identification and this is a combination of the serial set number of part and volume. In this example it's serial set number part 6F and there's no volume number in it. We also use this unique guide ID from the volume PDF. And within each volume PDF directory is a directory for each document. For this we use the combination of a sequence for how the document appears within the volume. And the document number and the part if it contains that. All of these are delivered to us by our vendor in the structure. Once the serial set volume is delivered to us we just the bag into the library content transfer system which is our it digital content management system. We've perform quality of review with sampler. Library created software that allows for review of digital objects. And in this process we review the images for the directory structure. It correctly identifies what document it is. The images belong to and the completeness to make sure all the expected documents are within the bag. And some of the next steps in the project. Once we ask step the bags we send them to GPO and they will talk about GPO creating more records and returning to the Library of Congress where we will ingest them into our system again and present them in locked.gov which is project as mentioned. So now I will pass this presentation over to Alina who's gonna talk about a project tracking database.

[Audio issues on speakers side] The second issue was there was inconsistent for meeting of multiple user support. Based on the problem set I broke down each column in Excel and grouped the fields together starting the normalization process including creating tables and establishing relationships between tables based on our business we designed to protect the data to make it more flexible by eliminating redundancy and inconsistent dependency. So as you can see the relationship diagram [Audio issues on speakers side]

Sorry about that. Let me start from the beginning this is Chris. Somebody mentioned that in the chat. So when we started the project we had three major issues. First of all there was no traceability of the data. We couldn't tell who are how the data was modified or inserted. And secondly the data quality we had inconsistent formatting to the data and duplicated records in multiple Excel spreadsheets. It was pretty much a disaster. Lastly, there was no ability to support multiple users which means that multiple people could not enter data into a spreadsheet at a single time and based on that problem we had to come up with a software solution that was a perfect fit for this project and the solution that Alina proposed was a Microsoft access program to centralize all the information into a single database. It's designed to be similar to the SQL Server and web-based server using programming such as SQL and BVA. Her first step was to clean up the data and remove duplicate records. Incomplete or erroneous data and to do this she broke down each column in Excel and grouped the related fields together then started the normalization process from there. This included creating tables and establishing relationships based on our business workflow that we designed for the project and to make the data more flexible by eliminating redundancy and in consistent dependency. This is an example of the entity of relationship diagram or ERD. It a visual form that's created from the system. Our workflow after cleaning up. So this is our workflow here. We've got three major steps. The first is evaluation the second is data control and then data analytics to track everything. Collecting results in simple analytics such as aggregation and total percentage of the goal are all done within the access database. We then use this data to create data visualization that was created in tableau. These are examples of the serial subset database main pages from version one and 2.0. During the pandemic she was able to make some changes she wanted to make to the database so that's how version 2.0 came about. The different features within these include the volume locations for evaluation. The ability to update a status of a volume or bag and the ability to track the shipment. A digitization volume overviews that tells us everything we know about the volume. It contains information from the physical evaluation and from the digitized additional objects that we received from the vendor and it then we also have reports and and administration page. So the goal of the database is to centralize all this serial subset data in one place and in an automated data management system. The purpose is to retrieve information accurately and efficiently tracking data history and reducing human errors. Allowing multiple uses to update in real time. The access database makes it easy for us and helps to maintain our resources and creating supports and statistics and updating our reference manuals and's tableau statistics. It's totally customized for the team member and the reflected project needs. So we've got four different entry information pages. Doe s the inventory history and a very powerful search function that helps us to create clean data. It's accessible and editable by multiple users simultaneously. It's a way to help tree members track who has done work and elements and when. And when we have conducted review of the digitized objects as well. Because of this we can manage our data efficiently. In this slide we give an image of our digitization over page. When you select a serial set volume number on the left it automatically brings up all the information to the right and filters so we can check volume status. The page lets you define specific status like who assigned the batch names and it allows you to narrow down the data through search. The slide is the data entry page. For data monitoring we synchronize the database with reports run from CTS. And this allows us to vigorously monitor the data to make sure that it's clean and consistent between our content management system and the database. This is an example of one of the forms that we print for evaluation. The physical volumes practice is one of the first steps when we pull volumes from our shelves. It allows us to keep track of the information. We've got each serial set numbers and in this example we provide the information to GPO about what's missing from the serial side volume on for pages or in our collections. That we evaluated before requesting volume from them. The database allows us to create shipment reports including shipment number, box numbers and the shipment date. All of our observations about the physical volume that was selected for digitization. That we provide to our offender in both print and in Excel spreadsheet. So they have all the information that we have about each serial subset number. From the administration page of the database we can monitor any of our statistics and it converts the unstructured information into a comprehensive and logical result. We manage our data visually. This helps to maintain keeping everybody together on the same page and you can easily see the progress being made for we week. Alina takes that information and feeds it into tableau for data visualization. This allows us to visually check our progress such as goals and total goals for the whole project. The number of best copies sent in the number of volumes set to the vendor. It just allows us to take the data from the database and converted to understandable easy user-friendly absent websites. We also use the database to allow communication between the contractors library technicians and managers as well as supervisors. And then the last slide is an example of one of these reports. We have contractors assisting in the evaluation of physical volumes and digital bags. This just shows the progress every week. The number of problems evaluated and the numbers selected. The two volumes are being evaluated, based on the number of our shelves. So, that was the end of her slides that me past the presentation over to Bailey.

Thanks for hosting us today it's wonderful to be here I'm excited to be here sharing those projects for what I'm working on. We track our progress in three workflow guides. First we have the serial subset digitization guide that gets background information as a collection and as a foundation for the metadata retractor and the processes. Physical volume evaluation and this is outlined and shared with my brand contract staff involved with the project. Second is our sampler quality review work guide that outlines the process for files received from our vendors. Instructions on how to use sampler is included along with the specific workflow to evaluate each digitized volume. Metadata and organizations templates are included in this PR guide. Instructions for documenting a formatting meditate are also included in each guide and are regularly updated in order to reflect developments in the serial set over the course of the decade it was published in. Tracking metadata is especially important because it allows us to create a shared word bank of terms and phrases easily re-created by filter and query searches. The contractors and vendors about the formatting of metadata for communicating volume specific issues relies on our ability to track metadata in a clean and efficient way. The supplies to printed digitization errors that we record before determining that a volume is complete or acceptable for public use. Here's an example of how we identify metadata and one of the tables. Here's how we format it in the digitized volume evaluation table. Metadata is both qualitative and quantitative. Volumes also include foldout maps and content that need to be treated with special attention. We've identified these materials and develop guidelines for counting them. This helps us to ensure all volumes completion as identified by particularly interesting public use content. Because this requires physical access to the volumes our team has adapted the volumes and priorities with the collection of time for extended telework. As we return on site and adopted one of my workflows which is combining volumes to complete a single volume. By focusing on site volumes a copy received by GPO since it's currently on hold.

Working remotely with work to shift our focus for the efforts. Our staff is also that on site and we've been ensuring we are all on the same page. One of our outreach efforts can be seen right here. Since January of this year I've made an effort to pose an interesting or relevant document along with some general contacts also helping me out on one of these posts as well. We are in the process of story maps based on information from the collection. Story maps are immersive experiences that allow the user to have entitled city sketches as tracks should be published within the next couple of months. [Indiscernible] The team is focused on development with additional nonmarket metadata [Indiscernible] Let's talk about our big picture plans in GPO. First the law library is preparing volumes for digitization for obtaining replacement points upon request. Specifically working with discarding the repository libraries when the law library is missing a volume or the volume isn't complete. Take a look at her digitization we won't be able to complete the digitization without your assistance. The choice delivery for 445 volumes over the summer. Using this content to develop work force practices we will use going forward. Ensuring data integrity throughout our processes for gathering metadata as well as individual reports and documents. Were compiling requirements to design this collection set from the government. Today let's share an overview of our processing. Please keep in mind were still in the development phase of things could change.

After receipt a team member goes into the GPO building to conduct initial processing. Scanning for viruses which is a requirement. Following virus scan the team members upload so it's accessible to all team members. Simply running the software against the whole delivery. Chris mentioned the specification for file structure. Bagger has been developed by the Library of Congress to validate check stones. If any bags don't validate working the law library for replacements. Once the bagger elevation is completed and successful, the files for delivery for the next step. That flows to the creation Mark metadata. Is to path we used. Depending on whether it's a monograph for serial. For monograph titles and the volumes the metadata has downloaded records for documents and reports for the 15th to the hundred and third Congress. With each delivery the team will learn what we call the script. It is the match volume for records in the delivery. 12,901 and content for delivery one. This script generates [Indiscernible] The second number is for document type for example Senate executive or house report.'s numbers are created using the data in these fields. The script inserts a 500 note that states that the record was batch processed. It gives you potential duplicate errors luckily there's only 120 such errors in the record batch. Team members are reviewing so no correction is needed. The record download does not have any big records. The record is found and put through the matcher script. No acceptable record is found -- finally there is Q&A step where the graphic team members are removed. Once approved they place it in a common location for the government access. We've had to take a slightly different approach with cereals. The set contains quite a few. Team members are searching to look at records for serial titles in their delivery. Acceptable records can be located and created and overtime there will be a pool of marked XML then they will edit the MARC XML about the issue in the delivery. Delivery containing 55 issues of the house and senate journals. We added the whole number to the issue and edit issue information to the title field and added a full date and year month day. And on this slide you can see the added records. There will be a QA step to review edits and edited

Mark will be placed in that same common location. And speaking of that let me pass it over to Heidi to tell us more about this. Statement

---

So what's going to happen next. Basically in general the MARC XML is transferred and that's going to trigger metadata extraction jobs that produces more metadata. To give a little bit of background. Over the past few months GPO's been gathering as much metadata as possible about the serial set volumes and documents found within it. Specifically we've extracted serial set metadata from a lot of different resources. One such resource is the law library metadata. Something we refer to as a volume level master metadata. Existing serial set records, we also for each of these various resources it's important to note that all of them contain different things so they have this great rich metadata. When we combined all of that with the MARC XML it serves as a foundation for the serial set volume and document metadata. So what exactly are these resources that are talking about? That is giving us all this metadata that I keep saying. First up is what we reference as the law library delivery metadata. This is a bag delivery metadata. Is a file generated and provided to GPO from the law Library of Congress. The law library talks about all the workflows in the data that they are collecting that's a part of the digitization process. When they provide a delivery to the GPO, what were doing is this really awesome deliveries grade sheet contained within the bag. Has shown it's kind of an example of a sheet that contained it has a lot of information about the bags. Concluded things like serial number, serial part number, Congress session and title. A number of maps and foldout and notes from when they did their public inspection. This is all great information in addition to the law library data. Is also been gathering serial set data and from existing serial sets related publications. So as mentioned. We've generated and were calling this a volume level metadata spreadsheet in all of that it's what we've done, we pulled the data from the serial set from things like the serial set inventory and we've pulled off information from finding aids and from the U.S. of public documents. Is scheduled serial set volumes and numerical list. Metadata on this volume level that we have generated includes things like serial number, serial set title, document number seven within a volume. Notes about the volumes of various sources and agency information. A combination of this Mark XML and what's pulled from this to resources that I talked about. It's all going to be compiled to generate the serial set impacts. Basically once we've compiled it and captured it is placed within the back folder. Once the associated metadata is approved, at some point there will be a medication that the bag is ready for upload. So what that really means is, right now our plan is to extract metadata from all these resources and as a part of the preprocessing package preparation all of that will be done before we submitted to govinfo. So right now were adjusting our internal processing workflows as we dive deeper into the various aspects of this data. It's a very deep dive there's a lot of serial set data and in addition to that were continuing to analyze what we extract. Making incremental changes to make sure we got the right processes in place. So that we have quality metadata to individual documents contained within them. The reason the metadata is so important is because that metadata enables custom search and browse functionalities and govinfo to create mod XML files and create relationships for the serial set data for existing documents and it's important for us that we get this metadata and have it before we proceed. Right now when it comes to metadata we are focused on duplicating and validating the metadata. Resolving any gaps in the data for each of the volumes. Were also starting discussions about options for gov info and visitation. In addition to the metadata. Our team is just starting out the workflow packaging in processing steps. We talked about this preprocessing before we get to govinfo but were taking a look at reviewing what needs to be done with these bags as we try to move them into govinfo and make these packages. What we mean by that packaging and processing steps. Things to consider are validating during ingest and then recording validation results in premise. Performing in processing steps such as optimizing for web display and applying digital signatures. As you can probably tell we are still in the early stage of the collection development process. The discovery of what the collection could be, it's important to note that all these things we talked about what we've discovered, the preprocessing workflows and all the discussions that were having with individuals both internally and externally. Those are all very key in helping us to design a gov info serial set collection. So with that I will go ahead and head back over to Suzanne.

So whether we can you be able to start using this information. The first public release of content the fall 2021. With volumes from the 69th Congress. After that initial public release additional volumes of the digital series will be made available so please keep in mind it's going to take several years to get all 16,000. As we wait for the big day next fall, keep an eye out for updates on the serial set project page. As mentioned to be digitized all volumes. Please take a look at her needs list. The Library of Congress is also a great spot to learn more about the content for this serial. It's a fascinating multifaceted publication and a treasure trove of historic document reports on a wide variety of topics. With the law Library of Congress we've made this collection

---

Thanks Suzanne we've got a few questions that came in while you were presenting. So I can start off.

Refused the data as we mentioned were starting to talk about display and metadata's displays.

Some serial set volumes have numerous individually titled reports in side. Is it easy for users to find those reports through cataloging and direct page level links or other strategies.

The reports will be on gov info for individual reports in the document. There will be XML records for each and every important document.

What's the timeline for completion?

From the law library perspective. We have in this might answer some other questions that up and asked, we've received a specific funding for Congress to digitize the serial set. So we've launched and we see this is a five-year project. Roughly digitizing 3000 volumes per year in order to reach the 15,000+ volumes. So we are anticipating about 5 years for the digitization. Obviously for the metadata to catch-up it might be a little bit longer but our goal is, as we go to the workflow and complete these volumes is to, incrementally add them to our website. For the digitization process is going to be at least five years.

Who did you contact to get information for the experience.

We have an interesting relationship with ProQuest. Obviously they've utilized library collections and in creating their database and perhaps if I can speak to another question is how is why we are doing this. If There are databases available that contain the data set. Well this goes back to providing government documents freely for the people. In the public domain. So that's the impetus of making these congressional publications available. Not subscription. It freely available and government websites. We utilize the database of rejects and ProQuest to evaluate the completeness of the work that we are doing in the value process. Making sure we are not missing parts or fold outs or maps as we go process we haven't spoken with them directly but rely upon their products to ensure the completeness of our project.

In 2009 the law library announced its plan to digitize hearings. Is that project complete and is that why you're working onto the serial set? Or do these projects progress simultaneously.

The library as the website described, entered into a relationship with Google to digitize congressional hearings. These were digitized however, the library has decided not to pursue it. There were limitations on being able to put these hearings on the public site. Essentially limited to on-site. So the library is not pursuing this and were looking for other sources to deal with the same goal providing congressional hearings freely available to the public. This project is not related to the congressional hearings, it's proceeding independently and we are not involved with Google with the serial set of the digitization. We have specific money from Congress to be able to continue.

Let me mention that GPL is currently digitizing additional batches of congressional hearings so be on the lookout for those. [Indiscernible] Let me explain a little bit that early in the history of this project are first inaugural year we undertook a pilot and that project was undertaken by offender, Crowley, who digitized about 1000 serial set volumes. But, as we had conversations and developed a partnership with GPO. We determined we needed to break down the volume records into documents. Currently been involved in that as well. The vendor that undergoing the full digitization is Creekside digital.

Is there a category for the serial set? We do the tag from the serial set of research from there I can type into the chat as well and I can provide the link. You can find all those posts and for the foreseeable future the going to only be published under my name. This one from a colleague which I will also include as well. Any reports for early executive agencies are in the serial set.

There will be four serial records were making this decision this week is going to be one for the serials that as a whole, with the volume number and one for the document type and one with the agency number.

Is to go to be both data export by API?

As we mentioned were still in the early phases of gov info development. So it's definitely something were where is desired. So it is noted. The law library is early in web development but the Library of Congress website does have an API and hopefully this zero set is accessible. Will these marked records for the digitized serial sets be sent to MARC customers as well?

Once they are ready to go they will be invested into the catalog and available.

With images be in color?

The site is being digitized both in grayscale and in color. The typical serial set page of just documents is grayscale.if there's any color in a map or a table or a plate. Then the image will be in color and included.

Define PAG. Is it an electronic folder?

It's basically an electronic folder but it takes a folder and puts a rapper around it so it creates the additional information such as text that can be used to verify the contents of the bag to verify the content have occurred to transfer in or out of our system. And then it contains some additional information about the content volume. There's a file that contains the number of documents and document names. So it's basically a rapper that's created that goes around the file folder. That contains additional information is used for digital preservation. To make sure that no changes have occurred from the vendor to us and us to GPO and then us, within our systems.

Have you found that the content of the given serial set of volume sometimes differs from item to item. If so, how do you decide what to complete volume?

In the evaluation process we've definitely come across as serial set. Numbers containing documents that were unexpected to combat this. The list of documents for each serial set number. If there's a range of documents it will say that this volume should contain a range of documents that was published in another serial set number. After the 1909 index we've been using schedule of documents just to confirm that the expected documents are within each volume.

Next question. We need to have results for missions [Indiscernible] Is sometimes differs from item to item and if so how do you decide what's the complete volume.

Are you going to connect you digitized volumes to the board digital GPO volumes to make serial set portal? Has anybody used the inventory use for needs in exchange.

We have a request for evaluation. So if you're configured for optional we will max with us. Not seeing any more questions if anybody has a last-minute question please enter it now otherwise were going to wrap up. Okay so thank you to our presenters today for a wonderful program and thanks to the audience for participating in our virtual conference. Up Nexis introduction to CVO that's role services products and communications. And the other virtual meeting room the next program is documents 101 guide to lost federal documents. If you want to join that program using the other URL to enter the meeting room. These URLs can be found on our event page which is linked to the FPL.gov homepage and from the event page click on join the sessions. For now let's take a short break and pick up again at 4:45 PM Eastern time so in about 15 minutes.