Digital Content Contributor Partnerships, Collaborating with GPO, and Building the National Collection

Jessica Tieman, Digital Preservation Librarian, GPO
David Walls, Preservation Librarian, GPO
Sara Sessler, Boise State University

The National Collection of U.S. Government Public Information

- Individual FDLP collections are developed and managed by Federal Depository Libraries in response to local user community needs.

- Ensuring future generations have access to Government information requires collaboration at both the regional and national level, and in coordination with GPO.

- Digital Content Contributors
- Boise State University Albertsons Library provides public access, via GovInfo, to digtized publications of regional interest.
- Oklahoma State University Edmon Low Library provides public access to the 7 volumes of Indian Affairs: Laws and Treaties (Kappler's) through GovInfo.
- University of Florida Libraries provides public access to various publications from the Panama Canal Commission (PCC) and Isthmian Canal Commission (ICC) through GovInfo.
- U.S. Department of the Treasury Library provides access to digitized content from the Treasury Library through GPO's GovInfo.
- Washington State Library provides public access on GovInfo to publications digitized to fulfill user or ILL requests.

Digital Content Contributors

- Crowd-sourced digital imaging
- Specifications for content creation ensure discovery, access, long-term preservation

National Digital Collection: GovInfo.gov

GPO provides the repository for hosting and preserving the content

NEWS RELEASE

GPO

FOR IMMEDIATE RELEASE: January 13, 2021   No. 21-02
GPO MEDIA CONTACT: GARY SOMERSET 202.512.1957 | mb 202.355.3997 | gsomerset@gpo.gov

GPO'S GOVINFO BECOMES THE ONLY ISO CERTIFIED TRUSTWORTHY DIGITAL REPOSITORY IN THE WORLD

WASHINGTON – The U.S. Government Publishing Office (GPO) has once again made history by becoming the only organization to maintain the highest global standard of excellence for digital repositories. GPO successfully completed its second yearly surveillance audit that is required to maintain its ISO 16363 Trustworthy Digital Repository certification for **govinfo**, the one-stop site for authentic information published by the Government. GPO achieved the certification by meeting official criteria for trustworthy digital repositories as defined by experts in the field.

"This recertification of **govinfo**'s status as a trustworthy digital repository confirms to the world what GPO already knows: **govinfo** is the one and only resource for authentic, trustworthy information published by the Federal Government," said GPO Director Hugh Halpern. "This certification is difficult to earn and maintain, and I am proud of our team for their hard work to keep **govinfo** fully certified. It is a huge step towards fulfilling our mission of *Keeping America Informed*."

Digital Content Contributor Partnerships

Digital Contributor Partnerships are right for your organization if...

- You already have digitally imaged government information that you want to contribute to GovInfo

Or!

- You are considering a digital imaging project and the content, or some of the content, falls into scope of GovInfo and you are interested in contributing content files to GovInfo

Is this the right partnership for your FDL?

Digital Content Contributors provide digital content to GPO for ingest into GovInfo. The Contributors provide content that meets GPO specifications for ingest into the repository. GPO provides ongoing access and preservation of the digital content. If you have existing digitally imaged publications, or are planning to undergo a digital imaging project, your FDL may already be in a great place to reach out to GPO to discuss opportunites.

How do I get started?

- GPO provides guidelines on how content files are preferably digitally imaged, named, and organized for efficient submission to the GovInfo repository

- These guidelines can be helpful for local digital imaging projects of any kind and provide insight into the level of effort required between both the Partner and GPO for ingest

Review GPO's Guidance Documents

Visit https://www.fdlp.gov/about/partnerships to review GPO's Guidelines on Digital Imaging and Packaging content for submission to GovInfo. The documents may help you either plan your digitization requirements for a project, or will help guide the conversation with GPO on how to get these publications into a ready-to-ingest state.

We will cover these guidelines in more detail later in the presentation!

How can I find more details?

- Establishing a partnership can be an iterative and collaborative process

- Preparing digitally imaged content for ingest is a cooperative effort – GPO acknowledges that different institutions require different levels of assistance in managing the imaging, processing, bibliographic description, and organization of content

Talk with GPO!

GPO welcomes initial conversations at any point of the process or any state of readiness. Ultimately, if your FDL would like to proceed to become a Digital Content Contributor, GPO will work with your institution to formalize an Memorandum of Understanding.

What effort, exactly, is required?

- Regardless if there is a single publication, or large volume, all content is arranged, prepared, and submitted to GovInfo in the same manner

- Submission to GovInfo ideally includes the following:
  - PDF (PDF/A; OCR'd; Optimized)
  - JPEG2000
  - JPEG thumbnail image
  - MARC.xml

Hands-on with Digital Publications

Preparing content for submission to GovInfo is a joint effort between the Partner and GPO. GPO will enable SFTP transfer for your content files. Content is submitted to GovInfo in the final state of a package containing PDF, a thumbnail image, JP2 master files, and marc.xml metadata. GPO can provide recommendations and guidance at any step of the process.

A more detailed look at content packages:

- At minimum, a high-quality scan of a PDF/A is required; this constitutes the version of the publication that is available for both access and long-term digital preservation

- A JPEG2000 master image file is preferred, as this provides more assurances of long-term preservation if PDF viability is ever uncertain, or PDF documents must be regenerated in the future; this is preferred over TIFF due to the smaller storage required

- MARC.xml is used for creating descriptive metadata within GovInfo; if there are no full marc records for the content, GPO can work with your local resources to settle on an alternative solution for generating bibliographic information

Access U.S. Government information on the go and on the shelf.

Guidance for Packaging Digital Content for Submission to FDsys/govinfo

The following guidance is for Digital Content Contributer Partners to utilize when packaging or preparing digital content for transfer of the digital content to the U.S. Government Publishing Office (GPO) for inclusion in **govinfo**. In order for GPO to submit content to the preservation repository, packages must be properly formed, and files must be named appropriately. Following the packaging guidance will assist GPO to submit content into govinfo effinciently.

govinfo Packaging Structure

Content in **govinfo** must be packaged in a specific structure. A package must include:

- tiff-submitted or jpeg2000-submitted folder: for the master preservation files (TIFF or JPEG2000) if master files are being ingested
- pdf-submitted folder: access derivative files (PDF or PDF/A)
- metadata-submitted folder: contains the marc.xml metadata and template.xml (XML) files (if provided)
- thumbnails-submitted (80x129 pixels jpeg) (if provided)

Other renditions can also be included in a package but are less common; additional file formats may be approved or requested by GPO.

govinfo File Naming Conventions

All files within the same package should be named with the same "stem" name. For example, if a given package represents a bill called "HR 1," all files within that package should be named similarly:
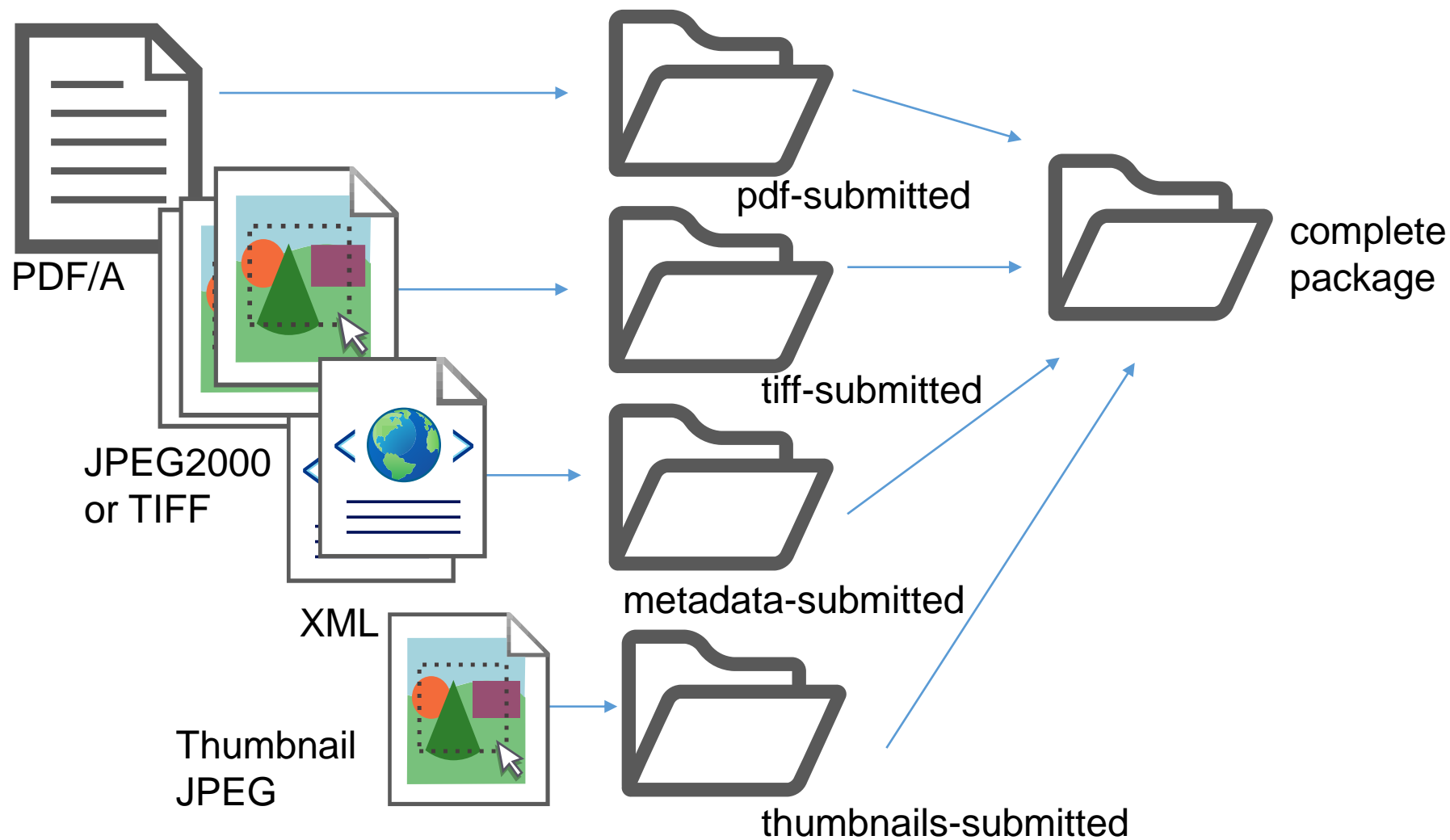
hr1.pdf
hr1.txt
hr1.html

Using these naming conventions will ensure that each package makes its way into the proper rendition and that the system can recognize that all files belong to the same package.

All packages, and files within packages, should be named with the following convention:

Package Contents

PDF/A

JPEG2000
or TIFF

XML

Thumbnail
JPEG

pdf-submitted

tiff-submitted

metadata-submitted

thumbnails-submitted

complete
package

Other considerations for package creation

- GPO provides recommendations for file-naming conventions. If your local institution already uses preferred file naming structures, GPO may be able to use these file names if they are consistent and unique
- GPO prefers marc information in XML format, but other formats may be accepted as suitable
- GPO can advise on scope determination of publications
- GPO can provide JPEG2000 file validation
- Content files may be other formats in addition to PDF; GPO can work with your institution on how best to provide description and access to these file types

Access U.S. Government information on the go and on the shelf.

Guidance for Packaging Digital Content for Submission to FDsys/govinfo

The following guidance is for Digital Content Contributer Partners to utilize when packaging or preparing digital content for transfer of the digital content to the U.S. Government Publishing Office (GPO) for inclusion in **govinfo**. In order for GPO to submit content to the preservation repository, packages must be properly formed, and files must be named appropriately. Following the packaging guidance will assist GPO to submit content into **govinfo** effinciently.

govinfo Packaging Structure

Content in **govinfo** must be packaged in a specific structure. A package must include:

- tiff-submitted or jpeg2000-submitted folder: for the master preservation files (TIFF or JPEG2000) if master files are being ingested
- pdf-submitted folder: access derivative files (PDF or PDF/A)
- metadata-submitted folder: contains the marc.xml metadata and template.xml (XML) files (if provided)
- thumbnails-submitted (80x129 pixels jpeg) (if provided)

Other renditions can also be included in a package but are less common; additional file formats may be approved or requested by GPO.

govinfo File Naming Conventions

All files within the same package should be named with the same "stem" name. For example, if a given package represents a bill called "HR 1," all files within that package should be named similarly:

hr1.pdf
hr1.txt
hr1.html

Using these naming conventions will ensure that each package makes its way into the proper rendition and that the system can recognize that all files belong to the same package.

All packages, and files within packages, should be named with the following convention:

Examples of file naming conventions

All packages, and pdf and image files within packages, should be named with the following convention:
"GovernmentAuthor-TitleofWork-DesignationofWork"
For example: "CDC-AnnualReport-Volume1"

When possible, use common abbreviations for authors and titles in the file names to make the file names as short in character length as possible. The maximum character length is 30 characters, not including dashes. Other examples of properly named package folders (or files within package folders) include:

"NOAA-ActionPlanFishRelease-Vol161"
"NLM-NativeVoices-Vol2"
"DOT-BeyondTraffic-Vol3"

Access U.S. Government information
on the go and on the shelf.

Guidance for Packaging Digital Content for Submission to FDsys/govinfo

The following guidance is for Digital Content Contributer Partners to utilize when packaging or preparing digital content for transfer of the digital content to the U.S. Government Publishing Office (GPO) for inclusion in **govinfo**. In order for GPO to submit content to the preservation repository, packages must be properly formed, and files must be named appropriately. Following the packaging guidance will assist GPO to submit content into **govinfo** effinciently.

govinfo Packaging Structure

Content in **govinfo** must be packaged in a specific structure. A package must include:

- tiff-submitted or jpeg2000-submitted folder: for the master preservation files (TIFF or JPEG2000) if master files are being ingested
- pdf-submitted folder: access derivative files (PDF or PDF/A)
- metadata-submitted folder: contains the marc.xml metadata and template.xml (XML) files (if provided)
- thumbnails-submitted (80x129 pixels jpeg) (if provided)

Other renditions can also be included in a package but are less common; additional file formats may be approved or requested by GPO.

govinfo File Naming Conventions

All files within the same package should be named with the same "stem" name. For example, if a given package represents a bill called "HR 1," all files within that package should be named similarly:

hr1.pdf
hr1.txt
hr1.html

Using these naming conventions will ensure that each package makes its way into the proper rendition and that the system can recognize that all files belong to the same package.

All packages, and files within packages, should be named with the following convention:

MARC metadata guidelines

All MARC.xml files should include the following:

- Full SuDoc value in the 086 field
- Dates within the 260 or 264 fields must be written in year-month-day format (YYYY-MM-DD); if the publication only lists a month and year, use the first day of that month. If the publication only lists a year, use the first day of that year.
- The 245 field must include subfield information to include the designation of the title if it is not a monograph.
- The 260 and 264 subfields "a," "b," and "c" must appropriately reflect the bibliographic information of the individual title within a series or as part of a multipart publication if it is not a monograph.
- A 500 field reading "This content was digitized by the [your institution's name] and is made publicly available through a partnership between the U.S. Government Publishing Office and [your institution's name].
- A 710 field with your library or institution's name. 710 _ _ $a Institution Name.

Access U.S. Government information **on the go** and **on the shelf.**

Guidance for Packaging Digital Content for Submission to FDsys/govinfo

The following guidance is for Digital Content Contributer Partners to utilize when packaging or preparing digital content for transfer of the digital content to the U.S. Government Publishing Office (GPO) for inclusion in **govinfo**. In order for GPO to submit content to the preservation repository, packages must be properly formed, and files must be named appropriately. Following the packaging guidance will assist GPO to submit content into **govinfo** effienciently.

govinfo Packaging Structure

Content in **govinfo** must be packaged in a specific structure. A package must include:

- tiff-submitted or jpeg2000-submitted folder: for the master preservation files (TIFF or JPEG2000) if master files are being ingested
- pdf-submitted folder: access derivative files (PDF or PDF/A)
- metadata-submitted folder: contains the marc.xml metadata and template.xml (XML) files (if provided)
- thumbnails-submitted (80x129 pixels jpeg) (if provided)

Other renditions can also be included in a package but are less common; additional file formats may be approved or requested by GPO.

govinfo File Naming Conventions

All files within the same package should be named with the same "stem" name. For example, if a given package represents a bill called "HR 1," all files within that package should be named similarly:

hr1.pdf
hr1.txt
hr1.html

Using these naming conventions will ensure that each package makes its way into the proper rendition and that the system can recognize that all files belong to the same package.

All packages, and files within packages, should be named with the following convention:

Example from Panama Canal, contributed from University of Florida

Example from Panama Canal, contributed from University of Florida

jpeg2000-submitted

- annualreport1990unit_Page_001.jp2
- annualreport1990unit_Page_002.jp2
- annualreport1990unit_Page_003.jp2
- annualreport1990unit_Page_004.jp2
- annualreport1990unit_Page_005.jp2
- annualreport1990unit_Page_006.jp2
- annualreport1990unit_Page_007.jp2
- annualreport1990unit_Page_008.jp2
- annualreport1990unit_Page_009.jp2
- annualreport1990unit_Page_010.jp2
- annualreport1990unit_Page_011.jp2
- annualreport1990unit_Page_012.jp2
- annualreport1990unit_Page_013.jp2
- annualreport1990unit_Page_014.jp2
- annualreport1990unit_Page_015.jp2

metadata-submitted

- marc.xml
- template.xml

pdf-submitted

- annualreport1990unit.pdf

thumbnails-submitted

annualreport1990unit_Page_001thm.jpg

Creating Lasting Impact Through Partnerships

GovInfo

Enhancing Access to the National Collection

Your participation as a partner is truly appreciated and enables GPO to enhance access to the National Collection of Government Information for generations to come through long-term digital preservation and stewardship in GPO's ISO 16363:2012 certified repository, GovInfo.

Albertsons Library
Boise State University

A Complete Novice's Journey Through The GPO Digital Content Contributor Partnership

Sara Sessler, Library Assistant 3
Serials, Continuing Electronic Resources, and Government Documents

Background Information

Boise State joined the Digital Content Partnership Program in September of 2019.

Albertsons Library has been a Federal Depository since 1966. As of 2021, we had roughly 81,056 government documents in print in our collection.

Our government document selection focuses on publications that directly pertain to the fields of study going on at our campus, such as from the Forest Service, Congress, Census Bureau, FBI, CDC, and FDC. *Everything* geosciences is especially desired by our faculty.

Even though we are selective with our choices, we try our best these days to ensure we're only "retaining" the electronic version of things when possible, otherwise we won't have any room for it all!

Why Did We Join the Program?

We first heard about the Digital Content Partnership Program from the FDLP newsletter. This began internal talks with library leadership about if the program would be right for us.

With our rapidly decreasing floor space, our mission has become to not to simply collect publications, but to promote information literacy. We reasoned that joining the program would not only preserve publications that were in danger of being lost, but also create access and increase discoverability for these items.

We concluded that our goals aligned with GPO's own and set out become a Digital Content Partner.

How GPO Aided Us in Joining the Program

There was one big snag when we first talked with GPO about joining: We were unable to catalog and create metadata for our submissions due to a staff shortage.

GPO graciously worked with us to submit our digitized content regardless of the missing pieces of data. With that, we were able to move forward with our first submission to govinfo.

GPO staff members have always been quick to respond to our questions, even for what we might consider a "stupid" question. For example, checking to see if a document does fall into scope for inclusion into govinfo. It can be hard to single out a federal agency sometimes if multiple state agencies were involved in the authorship of a document!

A Paraprofessional's Crash Course into Digitization

With the partnership in place, it was time to scan our test submission and learn the ropes of digitizing documents. I had previous experience with scanning images, but not at this level.

While the library's Interlibrary Loan, Archives, and Scholarly Communications units were already digitizing materials, this was the Serials' unit first foray into such. We relied on their knowledge (and some of their equipment) to begin our work.

We first started with ILL's scanner as well as Archive's DSLR camera, for oversized items that the scanner couldn't handle. Both units walked us through using the equipment, but it was up to us to create a workflow. This was accomplished purely by trial and error.

"Mastering" a 10 year-old scanner

It still does a good job though.

A Paraprofessional's Crash Course into Digitization, cont.

It quickly became apparent that the software already used by the library was insufficient for properly packaging our contributions to govinfo.

One key example is that Adobe Acrobat saves image files in the *wrong version* of the JPEG2000 format, so our first document was initially rejected by GPO. We scrambled to find a fix! Due to our limited budget free software, such as IrfanView, became our best friends. We still use IrfanView to this day to reformat image files and make thumbnails.
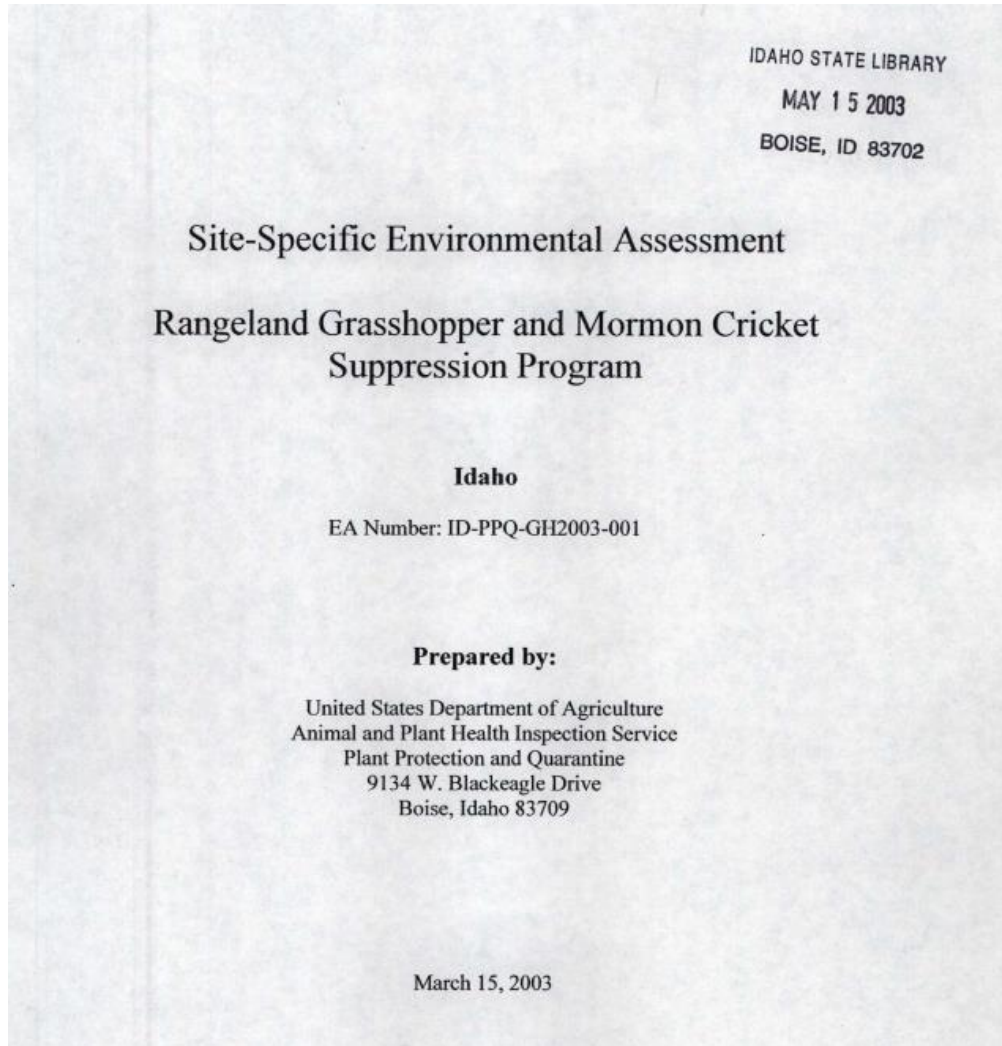
Be extra careful with the free stuff though, our first pick for a FTP program was treated as a virus by our systems. Thankfully we were able to find one that worked well.

Explaining what a JPEG2000 is to OIT

It is apparently a very obscure file format, it makes me feel like a hipster.

A Paraprofessional's Crash Course into Digitization, cont.

IDAHO STATE LIBRARY
MAY 1 5 2003
BOISE, ID 83702

Site-Specific Environmental Assessment

Rangeland Grasshopper and Mormon Cricket
Suppression Program

Idaho

EA Number: ID-PPQ-GH2003-001

Prepared by:

United States Department of Agriculture
Animal and Plant Health Inspection Service
Plant Protection and Quarantine
9134 W. Blackeagle Drive
Boise, Idaho 83709

March 15, 2003

This was our first submission!

We chose this document for our sample scan due to the following factors:

- Even, modern typeface to learn OCR with.

- "Book surgery" would not be needed to prepare the pages for scanning.

- Recent, so not as fragile as older items.

- Subject matter was indicative of the sort of documents we wished to submit to govinfo.

Visit the document on govinfo:

https://www.govinfo.gov/app/details/GOVPUB-A101-9142046c1fd3c318cc5e52869d2199be

A Paraprofessional's Crash Course into Digitization, cont.

The biggest challenge of all was figuring out how to fit digitization work into our daily tasks. None of us in the unit are dedicated solely to content submission, we have many, many other balls to juggle, so to speak.

Optical Character Recognition (OCR) alone can take many man hours to us perform depending on how degraded the text is or how unique the typeface is on the original document. Due to our budget constraints, we do not currently have software that will do this for us, but we believe that a license for OCR software is in the works.

The primary solution we've found was to have more people chip in when they can; when we started I was the only person to touch a document from start to finish during the digitization process. Now, other team members scan the documents in for me. We are also now training student assistants to perform OCR.

I hate OCR, and it hates me

But accessibility is extremely important.

A Paraprofessional's Crash Course into Digitization, cont.

Draft Digitization Workflow for GPO

1. Select a document to be scanned
2. Remove the binding if needed
3. Use the flatbed scanner or the camera to scan the document
4. Create a folder using the GPO naming convention and save files in it
5. Save raw files as JPEG2000 (.JP2) or uncompressed TIFF
6. If needed, use IrfanView to convert TIFF files to JP2
7. Build a PDF file from the TIFF images
8. Run OCR on the PDF, correct mistakes whenever possible
9. When done save PDF as a PDF-A
10. Create a thumbnail of the title page of the document
11. Send entire folder of files over to GPO via FTP
12. Let Jessica @ GPO know there is a new document to review

Through many trials, tribulations, and revisions, we created a workflow document that anyone in the room could follow. This is the table of contents for it.

It is still a work in progress, but such documentation is vital for training others so that they may help or take over.

GPO's own document for submitting digital files to them was essential, both in figuring out where to start and as a reference guide. I referred to it often for the thumbnail size guidelines alone since I keep forgetting the exact dimensions they need to be.

It is 80 pixels by 129 pixels, by the way.

Your computer kindly asks you not to use "inches".

Our Future with GPO and govinfo

Since starting our partnership, we were able to properly convey our need to library administration for a scanner of our own. While we still need more equipment and manpower to fully realize our potential, we remain optimistic that our contributions to govinfo will only increase as time goes on.

The program allows institutions of any size, from the smallest college to the largest university to preserve knowledge for future generations. Information that Americans as a whole can then access.

Not only are we helping to advance research at our university, we are aiding our government with our contributions. It is a pretty good feeling!

fi-7700, my precious

It still hates the JP2 format though.

Thank you for listening!

I'd like to thank my unit at Albertsons Library for their understanding and aid:

Nancy Donahoo
Library Section Manager, Serials, Continuing Electronic Resources, and Government Documents

Gerry Bryant
Library Assistant 2, Serials, Continuing Electronic Resources, and Government Documents

Mary Aagard
Associate Professor, Head, Acquisitions & Collections

Also a *huge* thank you to Jessica Tieman at GPO for her infinite patience and help.

Digital Content Contributor Partnerships, Collaborating with GPO, and Building the National Collection

01/30/2023