Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190025

The Rutgers Law Library U.S. Congressional Documents Digitization Collection

John JOERGENSEN
Rutgers Law School (U.S.A.)

Abstract. The motivations and processes developed at Rutgers Law Library for digitizing their print collection of United States Congressional hearings and committee prints, dating from 1967 to 2000 are discussed in this Chapter. Both the technical and collection goals of the project, and the important practical details of how it is being accomplished are described. The main theoretical goal was to show how a large scale digitization project could result in a useable, good quality, and sustainable collection while keeping costs at a scale that many institutions might consider affordable. The collection consists of over 25,000 documents. They are committee hearings and other print material that are generated as part of the U.S. Congress' legislative and oversight roles. Although the materials have been unbound, scanned, and checked for quality by hand, most other processes have been automated to minimize cost. Equipment and other expenses have also been kept to a minimum, but without compromise to overall readability, and archival quality.

Keywords. digitization: cost-effectiveness and practicality, U.S. congressional documents, law library

1. Introduction

The Chapter is a description of the Congressional documents digitization project that we have been conducting at the Rutgers Law School Library. This is one of several digital collections actively developed and maintained by the library. The Congressional Documents collection was our first, and most ambitious, attempt at large scale digitization from print. It will be ongoing for several more years, but the project has matured to the point where it can no longer be called experimental, and is an established part of our operations and our collection. The collection currently contains over 18,000 documents, consisting of over 5.5 million page images.

There are several reasons why we chose to develop this collection, and to do the kind of digitization what we have been doing. The first is practical: space in established libraries is always at a premium, so we needed to dispose of a large number of volumes to make room for new material. However, instead of merely throwing the older material away, we decided to digitize. The U.S. Congressional documents were a good candidate, because no one had digitized these materials already, and as government documents, there are no copyright or other restrictions on their distribution.

At the same time, the documents in this collection are a treasure trove of valuable information on matters the U.S. Congress has considered over the years. We have hearings and committee prints of congressional committees from the late 1960's through 2000. They are the transcripts of testimony and other reports generated as committees of the congress perform their governmental oversight functions, and consider new legislation. Matters from the impeachment of presidents to early considerations of global warming are all contained in these pages.

The technical justification for the project was to try to apply new equipment and techniques to find ways to reduce the cost of digitization. The old conventional wisdom about digitization was that readable and searchable document collections would cost in the range of $0.50/page to produce. The majority of this cost consists of the labor involved in either re-keying the printed material, or in performing detailed proofreading and correction of scanned and OCR'ed text[1]. However, with the advent of cheap, fast, high-quality scanning equipment, accurate OCR software, and inexpensive disk storage, it has become possible to change that equation considerably. In fact, our long-term goal for this project was to bring the cost of production to as close to $0.01/page as possible. If that could be done, then any institution wishing to participate in a digitization project could afford to do it. From a library perspective, having to spend millions of dollars on new shelf space, compared with several thousand on digitization, was an easy financial decision to make.

Since the library held these documents as a part of our role as a U.S. government documents repository, our first step was negotiating with the U.S. Government Publishing Office (GPO) for permission to withdraw our the documents from our depository. It is important to note that since there are no electronic copies of these materials from GPO, this was not a question of substituting the electronic for the print under existing depository rules[2] It was necessary to get permission to withdraw them before digitizing. The GPO was willing to consider the project, but required assurances concerning image standards, metadata, and a usable interface before allowing us to proceed[3]. The GPO requirements being met, we were allowed to proceed.

2. Procedures

The following procedures are what we have settled on as the most efficient and effective for our staff and facilities. It will be apparent that they represent a constant balance between quality and available resources. The guiding principle is not to let the perfect get in the way of the good. At the same time, we try to insure that the good really is good. The process starts with verifying our own catalog entry for the document, preparing the document for scanning, the scanning process, and post-scan processing.

2.1. Catalog Preparation

The first step in our conversion process is to insure that our online catalog remains accurate. So, when a set of books is taken off the shelves for digitization, they go straight

[1] Optical Character Recognition.

[2] See e.g.: http://www.fdlp.gov/collections/collection-maintenance/141-substitution-guidelines.

[3] See e.g. Draft Metadata report: https://www.fdlp.gov/news-and-events/321-metadata-report as an example of the standards that needed to be met.

to cataloging. There, a clerk checks them for an accurate record, including a Library of Congress Card Number (LCCN) (which we use as a key for accessioning in the digital collection), and writes the LCCN on the title page. The cataloger then changes the location code to ‘electronic document’ and inserts the URL where the document will be accessed. The above-mentioned URL is a reference to a CGI program that takes the document’s LCCN number as its parameter. It is of the form: https://njlaw.rutgers.edu/cgi-bin/lib/hearing.cgi?file=[LCCN]. At this stage of the processing, clicking on the URL in the catalog will result in a message informing the patron that the document is waiting to be scanned, and that they should inquire with the library should they need it.

2.2. Physical Preparation for Scanning

This a straightforward, if labor-intensive process. In order to scan efficiently at high volumes, a sheet fed duplex scanner is needed. This means, of course, that volumes must be unbound and pages trimmed.

Fortunately, most Congressional hearings and prints are very simply bound: most are folded and stapled with no cover (Figure 1). These are easily disassembled with a heavy-duty staple remover. The separated sections are then trimmed along the spine to get loose sheets using a standard guillotine paper cutter. For larger glued volumes, the covers are removed, the folded sections carefully torn away and then trimmed smooth.

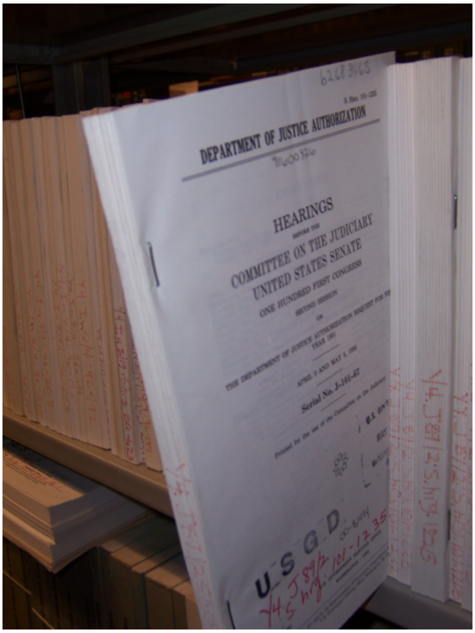

DEPARTMENT OF JUSTICE AUTHORIZATION
HEARINGS
COMMITTEE ON THE JUDICIARY
UNITED STATES SENATE
ONE HUNDRED FIRST CONGRESS
Serial No. J-101-47
U.S.G.D.

Figure 1. Congressional document bindings are pretty basic

After some practice, it was found that a little care at the paper cutter helped a great deal. Paper jams and skewed images could be avoided almost entirely by cutting to a consistent width for each volume, as well as being sure to cut enough to insure that all pages were properly separated, but while still leaving a decent amount of margin.

2.3. The Scanning Process

Once cataloged, unbound, and cut, a document is ready for scanning and is placed on a 'ready shelf'. Even with that, we keep standard settings for all documents, so the scan ning clerks only need to load the scanner and enter the LCCN number of the document before hitting the 'go' button. The software creates a new file directory for each new document, named by LCCN. Each scanned image is also named by LCCN, along with a 4 digit sequence number (i.e. LCCN-0001, LCCN-0002, etc.).

2.4. Post Scanning and Quality Control

After scanning, documents are rebanded, and kept until the final processed archival copy of the document has been burned onto DVD. This way they are available for rescanning should any flaws be discovered during quality control.

Once the electronic copy is completely processed, archived and available on the Internet, it is eligible for disposal.

In the meantime, a very basic and quick quality control check is done on images to insure that there are no obvious errors in the scans. This is done using using the Windows Explorer set to 'thumbnail' view (Figure 2). Using this thumbnail view, the checker can easily scroll through the set of images in a document, and identify faulty images. At this stage, defective images are obvious: folded pages, streaks or lines, super- imposed images, etc. The occasional error that is found is rescanned and the defective images replaced. Also verified at this stage are numbering and pagination errors. When all images are verified, a text file containing a copy of the document's library catalog record is placed in the document's directory, and the whole thing is moved to a 'ready' holding directory. The scanned document is now ready for automated processing. The catalog record is in standard MARC (Machine Readable Catalog) format, which is easily reformatted to XML/RDF.

2.5. Automated Processing

As stated earlier, there are two main goals in our preparation of images. First is a quality archive for long-term storage. The other is a readable image for viewing on the Internet. While not exclusive, these are very different goals. Although file size is an issue with the archived image, quality, robustness, metadata availability, and an open standard are more important. Display on the Internet, however, requires as small a file as practically useable, and a format that is easily viewed on most browsers. In addition, the documents should be searchable. Our solution is to create three files, each of which fills one of our goals.

One of the most important aspects of our procedures in this project is that all processing from this point forward is automated. Aside from someone to start the programs running and check on their progress, there is no more labor cost involved in the project. In addition, it should be noted that the processing scrips and programs are, with one exception, produced from open source (as in 'free') software. This also contributes greatly to savings.

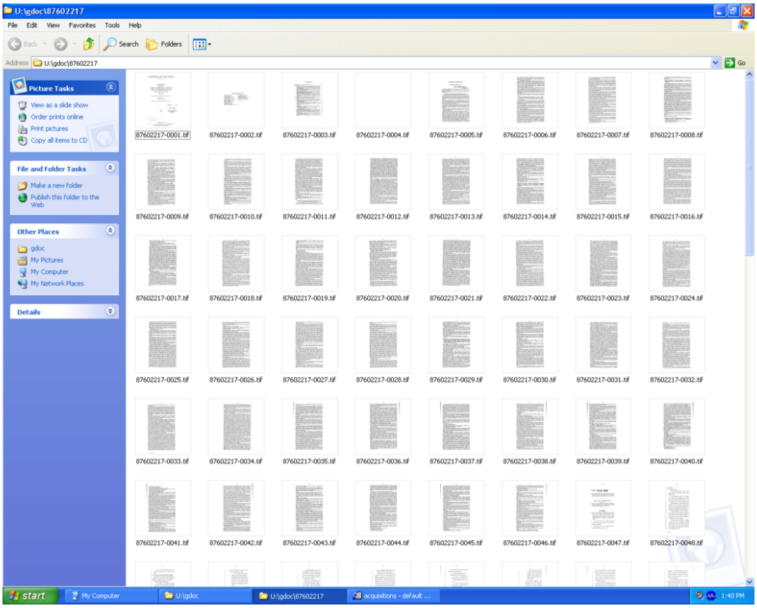


Figure 2. Quality Control: reviewing images for mis-scans

2.6. *Archive File*

One of the most important, and perhaps one of the most overlooked, aspects of successful long-term archiving of computer files is to preserve the connection between a computer file and the metadata which identifies that file. Typically, a digital library creates a metadata repository in the form of a database which has links to the objects identified, the same way a library catalog record has call numbers to direct users to the described book.

Unfortunately, the self-identifying nature of books does not regularly apply to digital records. At best, a whole document contained in a single computer file may or may not be sufficiently self-identifying to allow for some form of cataloging. In the case of something like the Rutgers congressional documents collection, however, each page of a document is stored as an individual set of digital files. If we were to rely on a separate metadata repository to identify items for long term archive purposes, we would be asking for disaster.

Fortunately, built into most current image formats is the facility for reliably embedding information directly into the image file. Although special software is required to write to it, most image formats have registers within the binary code in which descriptive metadata can be inserted. At Rutgers, we use Exiftool to embed bibliographic and other information directly into each image we will be using as an archival copy. This is free software, and can be used in batch-mode to edit many files at a time.

As noted earlier, during preparation of a document, we save a file containing a raw MARC record, taken from our catalog. During processing, this record is read, and a subset of information (author, title, LCCN#, SuDoc#, LC Subject Headings, etc.) are retrieved and reformatted as Dublin Core tags in RDF format. To this is added information on the image itself, such as dimensions, color depth, dots per inch, and where the image fits in the sequence of pages. An example of such a record is at Appendix I.

By embedding this information directly into each one of the images in the collection, each page of each document has sufficient identifying information so that the entire collection can be reassembled even without any external metadata repository. And, it can be done relatively simply. So, even if all filenames were changed and the database we use for searching were to be deleted, the entire library of documents can be reassembled by the computer using the metadata stored in each image file. As long as the image files themselves are maintained intact, we have good assurance of longevity for this collection.

The result is that our archived images are in a well supported open format that will be viewable for forseeable generations of computers. The holographic nature of the embedded metadata provides easily accessible and permanently available information which provides sufficient context to reliably place each image in its place within each document in the collection, no matter what may happen in the future to the external OPAC or interface database that are in place. They will never be just a jumble of images.

2.7. The PDF and OCR'ed Text Files

At the same time as the original scanned TIFF image is prepared for archiving, the processing script also creates two extra files. First, the TIFF is copied into a compressed PDF formatted image. Then, the TIFF is scanned by OCR (Optical Character Recognition) software, which generates a searchable plain-text file. The PDF and text file are saved along with the master TIFF image in the archive, but the PDF and text will mainly be used for the collection's user interface.

Of course, the PDF file is embedded with the same metadata as the TIFF image, also using Exiftool. In the case of the PDF, XMP format, which is particular to PDFs, is used.

As to the OCR'ed text: at this point, OCR software has become surprisingly reliable, 99% accurate in recognizing letters and words in images. With such accuracy, the OCR text can be searched with good reliability even without proofreading. Accurate for searching, however, does not necessarily mean readable. 99% accurate still means that out of an average 1,500 characters on a double-spaced page, there are still 15 errors. In addition, failures in font rendering and layout produce an inconsistent and unsatisfying result. For this reason, the compressed PDF image produced for presentation to the end user. Since it is a clear image, the original formatting, and letter renderings can be interpreted by the user themselves. This way, the OCR text can be indexed by our full-text search engine, providing accurate searching, but the user will view a readable, and quickly transmitted image file in PDF format.

It would, of course, be preferable to have a perfectly accurate and well formatted text file to work with. However, the cost of manual proofreading and reformatting makes this impossible. Given the good searchability of uncorrected OCR, and the viewability of both the PDF and TIFF image files, our method is an acceptable compromise. Complete proofreading, moreover, can only give marginal improvement to search accuracy, but with very large added cost. This must be compared to the presentation of an image, which, if of good quality, will always be a completely accurate representation of the original page.

This is, in fact, the process used in the JSTOR project, and Hein-on-Line, (not to mention Google Books), both of which have proven the viability of these methods[4]. It is not perfect by any means. However, at a generous estimate, we are able to digitize and make available to the world very large amounts of significant material at a cost of something near $0.02/page.

The software used to handle all this processing is, with the single exception of the OCR software, is free open source software that is widely available.

2.7.1. A Short Digression on PDF

The PDF format is very flexible as to content, and can be essentially whatever you need it to be. In the current project, the PDF files in question are pure, unsearchable page images, preserved in the PDF format so they can be conveniently viewed in a web browser. A PDF file can also contain pure formatted text, which would be searchable, or even text with an image superimposed (which is also searchable). In the context of the current project, a pdf/text file would display all the imperfections of the OCR result, and would not be desirable. A pdf image-on-text file would preserve display and be searchable, but suffers from the disadvantage of being a very large file. In the case of documents that are fairly small (say, 25 pages or less), this would be a very workable solution, but with book size documents, it is not practical. Our workstations cannot produce them, our Internet connection would break down transmitting them, and all but the most powerful machines would freeze on attempting to display them.

3. Preparation of a Searchable Document

At this point in the processing of each document, we need to address the problem of context in full text searching. The problem is that individual pages are too small a sample of the typical congressional hearing or committee print to preserve the context of a document as a whole. Searching an index of individual pages is not much use. Since there is no guaranteed regularity to the documents that would allow for any sort of sampling or breakdown, it was decided that the documents must be made searchable only as whole documents. To do this, the OCR'ed text files of each page are concatenated with embedded page break markers inserted at the beginning of each page to reconstitute a single text file of the whole document. The metadata set is then also included as HTML meta tags, as well as some basic HTML head and body tags. The Swish-e search engine then indexes each document as a whole, and creates secondary indicies for each of the meta tags included in the document headers. In this way, users can not only search the full text, but can include some useful fields in their query, such as title, author, year of publication, LC subject heading, and Sudoc number.

Finally, the processing script also concatenates the compressed PDF image files into a set of multipage files of about 50 pages/file. The user interface makes these files available to users who wish to download the enitre document.

In the archive, the compressed PDFs, page level text, large HTML, and large PDF files are saved along with the TIFF master image. All of this material is saved in a single directory per document, and burned onto DVD's as well as being stored on our remote

[4] https://about.jstor.org/ and https://heinonline.org/HeinDocs/HOLBrochure.pdf.

backup, both locally and to a commercial cloud service. The final steps in the processing are that the OPAC is updated to include a link to the online document, and a notation is included in the OPAC and in a separate collection database of the DVD serial number on which the document is archived.

4. User Interface

The core of the user interface for all of the Rutgers – Camden digital collections is the Swish-e search engine[5]. This is a highly configurable, and robust freeware search engine, which remains under active development. It supports boolean searching, can return KWIC (key word in context) sample results, and can maintain separate indicies of html meta tags.

As concerns our government documents project, the configurability of Swish-e is of vital importance. As mentioned above, Swish-e is actually indexing large quasi-HTML files made of contatenated OCR'ed text files. This allows for needed context for searching. When displaying results, however, it would be preferable to display the book page which actually contains the search results. This is something that had to be hacked into the operation of Swish-e. We did it in two steps. First, As mentioned, at the time the large HTML file is created, numbered sequential page break markers are also inserted. Then, we modified the way Swish-e handles its KWIC results feature.

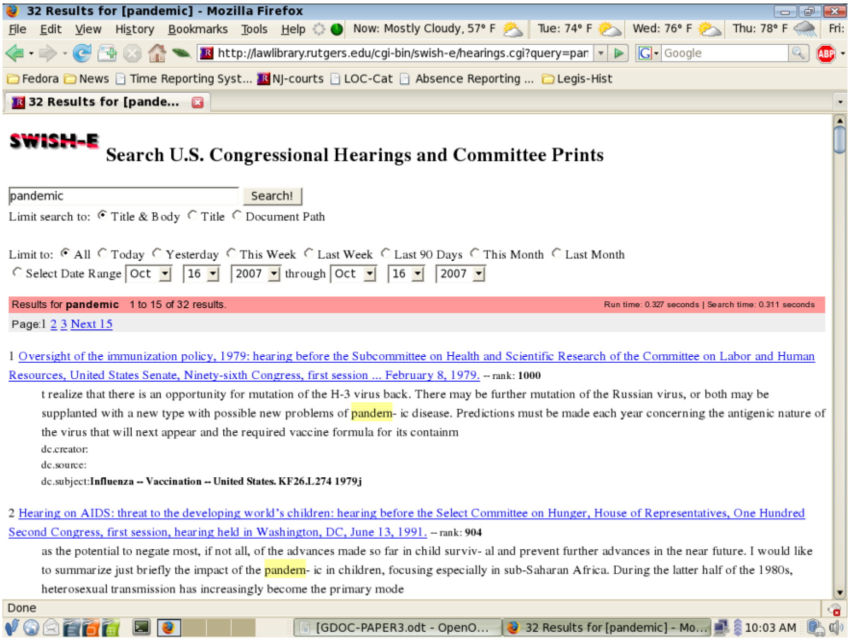
By default, Swish-e provides about 20 or so words before and after the first search term located, enough for most people who wish to look at a KWIC result. We started by expanding this to a much larger number, large enough so that we were sure to capture the numbered page break marker that would come ahead of the keyword. We then added code to the Swish-e script to scan the KWIC result for the page break number that is closest to, but preceding the keyword. We then inserted more code that, instead of providing a link to the indexed HTML document, calls another web script that will present the PDF image that corresponds to the page number, along with a side bar of navigation tools that allow the user to jump between individual page images quickly, or to download the entire document in one or more large 50 page files. That job done, we cut the KWIC result back down to a reasonable size for use on the Swish-e search results page (Figure 3).

Finally, the documents are displayed as seen in Figure 4. Individual page PDF Images are displayed for ease of reading and transmission, along with a navigation panel which allows for paging, etc. A link for downloading the entire document is also provided.

5. Cost and Practicality Issues

Those who might be familiar with producing digital text will note the greatly reduced amount of labor needed to produce a searchable document collection using the above methods. In fact, the labor savings are all the more dramatic when one considers the amount of multitasking that is possible using these methods. Given reasonably careful document preparation, we have experienced very few problems with paper jams, etc. in the scanning process. This means that the scanning devices, once loaded with 100 or so

[5] http://www.swish-e.org.


32 Results for [pandemic] - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Now: Mostly Cloudy, 57° F Tue: 74° F Wed: 76° F Thu: 78° F Fri:
http://lawlibrary.rutgers.edu/cgi-bin/swish-e/hearings.cgi?query=par
Google
Fedora News Time Reporting Syst... NJ-courts LOC-Cat Absence Reporting ... Legis-Hist
32 Results for [pande...
SWISH-E
Done
[GDOC-PAPER3.odt - OpenO... 32 Results for [pandemic] - Mo... 10:03 AM

Search U.S. Congressional Hearings and Committee Prints

pandemic Search!

Limit search to: ● Title & Body ○ Title ○ Document Path

Limit to: ● All ○ Today ○ Yesterday ○ This Week ○ Last Week ○ Last 90 Days ○ This Month ○ Last Month
○ Select Date Range Oct 16 2007 through Oct 16 2007

Results for **pandemic** 1 to 15 of 32 results. Run time: 0.327 seconds | Search time: 0.311 seconds

Page:1 2 3 Next 15

1 Oversight of the immunization policy, 1979: hearing before the Subcommittee on Health and Scientific Research of the Committee on Labor and Human Resources, United States Senate, Ninety-sixth Congress, first session ... February 8, 1979. -- rank: **1000**

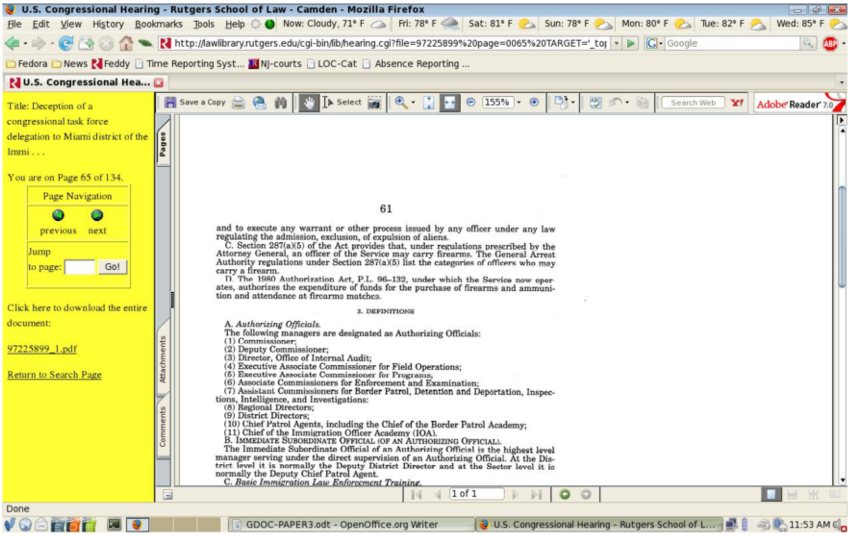
t realize that there is an opportunity for mutation of the H-3 virus back. There may be further mutation of the Russian virus, or both may be supplanted with a new type with possible new problems of pandem- ic disease. Predictions must be made each year concerning the antigenic nature of the virus that will next appear and the required vaccine formula for its containm

dc.creator:
dc.source:
dc.subject:**Influenza -- Vaccination -- United States. KF26.L274 1979j**

2 Hearing on AIDS: threat to the developing world's children: hearing before the Select Committee on Hunger, House of Representatives, One Hundred Second Congress, first session, hearing held in Washington, DC, June 13, 1991. -- rank: **904**

as the potential to negate most, if not all, of the advances made so far in child surviv- al and prevent further advances in the near future. I would like to summarize just briefly the impact of the pandem- ic in children, focusing especially in sub-Saharan Africa. During the latter half of the 1980s, heterosexual transmission has increasingly become the primary mode

Figure 3. Swish-e search results page


U.S. Congressional Hearing - Rutgers School of Law - Camden - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Now: Cloudy, 71° F Fri: 78° F Sat: 81° F Sun: 78° F Mon: 80° F Tue: 82° F Wed: 85° F
http://lawlibrary.rutgers.edu/cgi-bin/lib/hearing.cgi?file=97225899%20page=0065%20TARGET='_to|
Google
Fedora News Feddy Time Reporting Syst... NJ-courts LOC-Cat Absence Reporting ...
U.S. Congressional Hea...
Save a Copy Select 155% Search Web Adobe Reader 7.0
Pages Attachments Comments
1 of 1
Done
GDOC-PAPER3.odt - OpenOffice.org Writer U.S. Congressional Hearing - Rutgers School of L... 11:53 AM

Title: Deception of a congressional task force delegation to Miami district of the Immi . . .

You are on Page 65 of 134.

Page Navigation
previous next
Jump to page: Go!

Click here to download the entire document:

97225899_1.pdf

Return to Search Page

61

and to execute any warrant or other process issued by any officer under any law regulating the admission, exclusion, of expulsion of aliens.

C. Section 287(a)(5) of the Act provides that, under regulations prescribed by the Attorney General, an officer of the Service may carry firearms. The General Arrest Authority regulations under Section 287(a)(5) list the categories of officers who may carry a firearm.

D. The 1980 Authorization Act, P.L. 96–132, under which the Service now operates, authorizes the expenditure of funds for the purchase of firearms and ammunition and attendance at firearms matches.

3. DEFINITIONS

A. *Authorizing Officials.*
The following managers are designated as Authorizing Officials:
(1) Commissioner;
(2) Deputy Commissioner;
(3) Director, Office of Internal Audit;
(4) Executive Associate Commissioner for Field Operations;
(5) Executive Associate Commissioner for Programs;
(6) Associate Commissioners for Enforcement and Examination;
(7) Assistant Commissioners for Border Patrol, Detention and Deportation, Inspections, Intelligence, and Investigations:
(8) Regional Directors;
(9) District Directors;
(10) Chief Patrol Agents, including the Chief of the Border Patrol Academy;
(11) Chief of the Immigration Officer Academy (IOA).
B. IMMEDIATE SUBORDINATE OFFICIAL (OF AN AUTHORIZING OFFICIAL).
The Immediate Subordinate Official of an Authorizing Official is the highest level manager serving under the direct supervision of an Authorizing Official. At the District level it is normally the Deputy District Director and at the Sector level it is normally the Deputy Chief Patrol Agent.
C. *Basic Immigration Law Enforcement Training.*

Figure 4. View of a document. Compressed PDF image displayed with navigation controls to the left

pages at a time, run with very little attention. So, little, that a single staff member typically runs both our scanners simultaneously, while either preparing new books, or performing quality control from a third computer. The staff only pause from their other tasks to load more pages. In the end, therefore, the labor cost per page is actually quite low. The Panasonic duplex scanners have proven themselves to be durable and easy to use, and can be had for under $1,000. The large guillotine-type paper cutter was purchased on E-Bay for about $300.

The other, and even more significant factor controlling the cost is using the ‘dirty OCR’ process. The time and expertise required to proofread the volume of material that we are processing would easily run into the hundreds of thousands of dollars (a rough but reasonable estimate would be something like 800,000 images/year at $0.50/image on average = $400,000). Again, the principle is not letting the perfect prevent the good. The searches are very good, and the images are very readable.

6. Conclusions

Having access to the online information offerings of services like Westlaw and Lexis does not make a library a repository of information for now and the future. It only gives us current access to Thompson and Elsevier’s repositories. To the extent that others may be willing to prepare and actually sell their digital products, we can happily buy and own those items. The strong trend, however, is still in the direction of information being subjected to continuing control by vendors that will continue to charge for access. It seems, therefore, that in order to guarantee public access to the laws of our nations, libraries, LII’s and other organizations need to produce our own digital assets, whether individually or in consortium.

This Chapter is a description of how that can be done in a cost-effective manner, affordable by anyone with the will to start doing it. In the context of a larger consortium effort, the only limit is provided by copyright. Combined with a program of harvesting useful documents from the Internet for preservation and permanent access by the library, a large amount of digital assets that can be accumulated can be significant.

APPENDICES

APPENDIX I: Example RDF Record

```
<?xml version="1.0"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                xmlns:dc="http://purl.org/dc/elements/1.1/"
                xmlns:dcterms="http://purl.org/dc/terms/"
                xmlns:dctype="http://purl.org/dc/dcmitype/"
                xmlns:dcq="http://dublincore.org/2004/09/20/dcq">
 <rdf:Description rdf:about="URI:/77999007/77999007-0001.tif">
<dc:title>Nuclear order and human values, London, 1977:|report on the eleventh
meeting of members of Congress and of the European Parliament, July 11-13, 1977,
pursuant to H. Res. 313 ... </dc:title>
<dc:creator>
Committe on International Relations, United States House of Representatives
</dc:creator>
<dc:publisher>Washington:.S. Govt. Print. Off.,</dc:publisher>
<dc:source>Y 4.In 8/16:N 88/7 </dc:source>
<dc:contributor>
     <rdf:Bag>
   <rdf:li>United States.|Congress. </rdf:li>
          <rdf:li>United States.|Congress.|House.|Committee on International
```

```
                   Relations. </rdf:li>
           <rdf:li>European Parliament. </rdf:li>
           <rdf:li>Rutgers University School of Law - Camden</rdf:li>
     </rdf:Bag>
</dc:contributor>
<dc:subject>
     <dcterms:LCCN>
           <rdf:value>77999007</rdf:value>
     </dcterms:LCCN>
</dc:subject>
<dc:subject>
     <dcterms:LCSH>
        <rdf:value>
           <rdf:Bag>
              <rdf:li>Atomic power|International control. </rdf:li>
              <rdf:li>Atomic weapons and disarmament. </rdf:li>
              <rdf:li>Civil rights. </rdf:li>
           </rdf:Bag>
         </rdf:value>
     </dcterms:LCSH>
</dc:subject>
<dc:subject>
        <dcterms:LCC>
                <rdf:value>N/A</rdf:value>
        </dcterms:LCC>
</dc:subject>
<dc:language>
<dc:language>
        <dcterms:RFC1766>
                <rdf:value>EN</rdf:value>
        </dcterms:RFC1766>
</dc:language>
<dc:date>
        <dcq:created>
                <rdf:value>1977</rdf:value>
        </dcq:created>
</dc:date>
<dc:date>
        <dcq:issued>
                <rdf:value>2007-1-22</rdf:value>
        </dcq:issued>
</dc:date>
<dc:format>
        <dcterms:IMT>
                <rdf:value>image/tiff</rdf:value>
        </dcterms:IMT>
</dc:format>
<dc:format>
        <dcterms:extent>
                <rdf:value>64284 bytes</rdf:value>
        </dcterms:extent>
<dc:format>
        CompressionType=Group4
</dc:format>
<dc:format>
        ImageSize=3307 x 5423
```

```
</dc:format>
<dc:format>
        ImageDensity=600x600
</dc:format>
<dc:format>
        Colors=Bilevel
</dc:format>
<dc:format>
        ImageDepth=1 bits
</dc:format>
<dc:rights>
        Public Domain
</dc:rights>
<dc:relation>
        <dcq:isPartOf>
                <rdf:value>page 1 of 98</rdf:value>
        </dcq:isPartOf>
</dc:relation>
</rdf:Description>
</rdf:RDF>
```

APPENDIX II: Sample MARC record, source of RDF

```
001       NJRL06-B7891
008       061114s19777777dcu7777777777f000707eng7d
040     __ |a DGPO|DGPO|CStRLIN|NJRL
074     __ |a 1017
086     __ |a Y 4.In 8/16:N 88/7
245     __ |a Nuclear order and human values, London, 1977 :|report on the eleventh
           meeting of members of Congress and of the European Parliament, July 11-13,
           1977, pursuant to H. Res. 313 ...
260     __ |a Washington :|U.S. Govt. Print. Off.,|1977.
300     __ |a xiii, 83 p. ;|23 cm.
500     __ |a At head of title: 95th Congress, 1st session. Committee print.
500     __ |a Submitted to the Committee on International Relations.
500     __ |a Issued Oct. 1977.
650     __ |a Atomic power|International control.
650     __ |a Atomic weapons and disarmament.
650     __ |a Civil rights.
710     __ |a United States.|Congress.
710     __ |a United States.|Congress.|House.|Committee on International Relations.
710     __ |a European Parliament.
852     __ |a NjR-L|Y 4.In 8/16:N 88/7
852     __ |a NjR-L|E-document
856     __ |a |http://lawlibrary.rutgers.edu/cgi-bin/lib/hearing.cgi?
              file=77999007&page=0001|Access this document
902     __ |a ebook
```

APPENDIX III (Personnel Cost Estimates)

Productivity estimates, per 400 page item:	
Unbinding of text:	5 minutes
Scanning (time actually spent on task):	15 minutes
Quality Checking/copying MARC Record:	10 minutes
DVD burning:	1 minute
Processing	0 minutes (done in batches by computer.)
Total estimated time per book	36 minutes/item.
Personnel cost:	$18.00/hour * 0.52 hours/item = $9.30item.
	400 pages/item = $0.023/page.