ACIR Digitization Project: Working Manual Version 2

University of North Texas

LIBRARIES

Information Technology Services,
Digital Projects Unit

March 2006

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

1. Introduction

The University of North Texas (UNT) is one of over 60 federal depository libraries in Texas and 1,250 in the nation, storing and providing free public access to government information. In the 1990s, many government agencies and commissions changed their publishing practices and began providing access to their publications in a digital format over the World Wide Web.

Since 1997, the UNT Libraries has formed partnerships with state and federal agencies to preserve their electronic publications. One such partnership is the *CyberCemetery*. In an agreement with the U.S. Government Printing Office (GPO), UNT preserves the Web sites of federal agencies that have ceased operation. This project presents many challenges related to the capture and long-term preservation of this heterogeneous collection.

In 1997, a Memorandum of Understanding was signed by UNT and the U.S. Government Printing Office (GPO) to outline respective roles in the preservation of the electronic publications of a government agency that had ceased operation, the Advisory Commission on Intergovernmental Relations. This was the first "dead" agency ingested into the site, which soon became known by its users as the "CyberCemetery." The *CyberCemetery* page is growing continuously, and presently (in 2004) it provides public access to the Websites of 22 "deceased" agencies. At this time, work is underway to ingest 15+ government agencies that were folded into the complex structure of the new Homeland Security Department.

Although ACIR was the first "dead" agency ingested into the *CyberCemetery* site, the ongoing work towards retro-digitization of the print reports (ACIR Policy Reports and ACIR Information Reports), are still underway. As can be seen from Table-1 below, out of the total of 327 published ACIR documents (130 Policy Reports and 197 Information Reports), as of March 2004, only 137 documents, (i.e. 69 Policy Reports.+ 68 Information Reports) were digitized. That means 190 documents (58%) still need to be converted to digital form.

In light of this, this working manual provides a step-by-step explanation of how to convert and create digital representations of documents in the *Adobe* Portable Document Format (.PDF) for Web based electronic access. The procedures manual is also supplemented by a series of appendices that provide further descriptions, including glossary of terms and links to useful Web sites.

Report ↓ Year →	**1960s**	**1970s**	**1980s**	**1990s**	**Total***	**Grand Total***
Policy Report (P.R.)	A1-A35	A36-A70	A71-A113	A114-A130	**130**	Out of **327** published reports, the digitization of **137*** documents completed.
Total No. of Published P.R.	35	35	43	17	**130**	
Total No. of Digitized P.R.	2	17	34	16	**69**	
Info. Report (I.R.)	M1-M47	M48-M116	M117-M170	M171-M197	**197**	
Total No. of Published I.R.	47	69	54	27	**197**	
Total No. of Digitized I.R.	1	23	19	25*	**68***	

* Some of the ACIR documents' series were published in multiple volumes.

Table-1: ACIR Reports' Published and Digitized by Year (March 2004).

2. General Background

The Conversion of textual, visual, or any other information sources to digital form encompasses a range of activities, procedures and of course technologies. Of course the degree and levels of sophistication varies depending on a number of factors including: the nature of the source material, the current and potential users' behaviors, the format and the potential nature of the digital products, (i.e. how digital resource will be described, delivered, and preserved).

As you proceed through this manual, it is helpful to bear in mind that the ultimate purpose of the work described is to make document images available in the form best suited to the needs of all actual and potential users. Therefore: the digital documents are intended to be:

- Faithful representations of the original work;
- Quick to download;
- Fully searchable;
- Easy to navigate;
- Accessible to the blind or sight-impaired.

By following the steps outlined below, it is possible to create PDF files meeting all of these criteria. A proper and systematic file management system always facilitates the management of the ever-growing file size. Hence, to keep close track of the various files you create at various points in the process, you need to create a new directory for ACIR documents on your personal (I) drive and store your work there until it is completed. It may be useful to create subdirectories in which to place each scanning project document separately. Bear in mind, however, that no one but you have access to the files on your personal drive. As projects are completed, move the files to the shared ACIR directory on the Government Documents departmental drive (H: or P://Dept/Govdocs/ACIR). From there, files can be loaded onto the Web server and backup copies can be saved (or burned to CD/DVD) even for archives.

There are two distinct processes for creating online PDF files. In the *first* scenario, actual scanning of text will have been outsourced to a commercial service bureau that makes page images (.PDF) file available to us for downloading from their Web site or send to us on external storage medias. In the *second* document process scenario, the scanning is performed *in house.*

3. DOCUMENT PROCESSING PROCEDURE-I: DOWNLOADING PDF FILES

If the files you intend to use have already been scanned by a commercial service bureau and come to us as a complete, bound document file, (usually in PDF format) then follow the steps outlined below.

3.1 Acquire and open the file

- Step-1: Go to the download site or appropriate storage media: http://plutonium-erl.actx.edu/contentdev/northtexas.html (for first phase of ACIR Project.). Other vendors may be used in other phases and may provide documents on different storage media: CD-ROMs, DVDs, etc.

- Step-2: Look for new documents that have not yet been downloaded or already digitized. Note that these will normally be in the form of a bound, PDF file.

- Step-3: Download any of these documents and save them to your personal directory. When you save the file, make sure the file name is descriptive and consistent with local naming convention (see also Section-5).

3.2 Cleaning a blemished pages

Although the main purpose of the digitally converted document is to be faithful representations of the original work, often the first few pages have stamps on them or other damaged areas. To correct this there needs to be some touch-up. *Adobe Acrobat* software allows you to edit PDF document in a variety of ways. For instance, you can edit text and graphics within a file, crop, rotate, insert, or delete pages. You can also renumber pages or rearrange the order of pages in a document. The only down side of *Adobe Acrobat* is it does not reduce file size. If the TIFF files are available and replacing or recreating the PDF files are feasible, then see Section 4.3 for cleaning TIFF files using Adobe Photoshop. The following are the common cleaning steps using *Adobe Acrobat*:

- Open the file that you will work on in *Adobe Acrobat*.
- Select the form tool.
- Draw a box around the damaged area. A menu box will pop up.
- Name the area something simple e.g.: “w1” for white button 1.
- Match the background color to the color of the selected area. If the area has many colors, you may need to draw many boxes to properly cover any damaged area.
- Sometimes the letters of the title are partially covered. You can repair this by either inserting text, trying to match the font and color as closely as possible, or drawing many boxes and filling them with the text color to make it look like the letter. It is often hard to match the color exactly on the cover. Use the custom colors, and use trial and error until the color matches.

3.3 File Organization

Once the PDF document has been cleaned, then the next step would be working on file organization process for better aesthetic effect and of course, ease of navigation. To maintain all PDF files in consistent way, they need to be organized in the same format as the locally digitized documents in *Procedure-II*. To do so, jump to section-5 and continue the file organization process from *Section 5.2* through *Section 5.5* as appropriate.

4. DOCUMENT PROCESSING PROCEDURE-II: IN-HOUSE DIGITIZATION PROCESS

The main purpose of this section is to offer general scanning hints and tips intended to help with fundamentals and other basic scanning information so as to maintain consistency in all digitization projects at the Government Documents and Digital Projects Departments of the UNT libraries. The process described below takes several factors into account. Some of the most important considerations include:

- **The objectives of our departments,** (e.g. meeting users' needs, providing high quality digital objects and metadata information, ensuring long-term access, etc.)
- **Characteristics of our collections** (e.g. color, condition, format, type [e.g. document in paper, microfilm/microfiche etc.] and related information)
- **The needs of our users** (e.g. full text search, internal navigation and other functionality,) etc.

4.1 Initial startup: preparing to scan hard copy documents

- Make sure that the paper document is as clean as it can be for the best possible scan. (If the document is on microform, a high-quality copy must be made for scanning).
- Clean the scan bed glass.
- Place the document to be scanned on the scanner. Make sure the page is flattened and both vertical and horizontal edges of the document are aligned correctly with edges of the scan bed. If the document is oversized, try to fit the text and/or images on the scan surface, letting any blank margins extend beyond the scanning surface, if need be.

4.2 Scanning

We believe that the scanning basics are all the same, although the implementation details may be slightly different in some cases. You can use various Scan software commands to run your different scanners, but to make image editing (cleaning, rotating, cropping, etc.) easier, Adobe Photoshop is preferable. In our case, we have been using a variety of hardware and software for capturing both black and white and Color documents. : The great high-speed flatbed scanners includes: Microtek Scanner (ScanMaker 6400XL) and duplex Fujitsu Scanner (fi-4340C), which offer excellent document capturing features for our specific requirements.

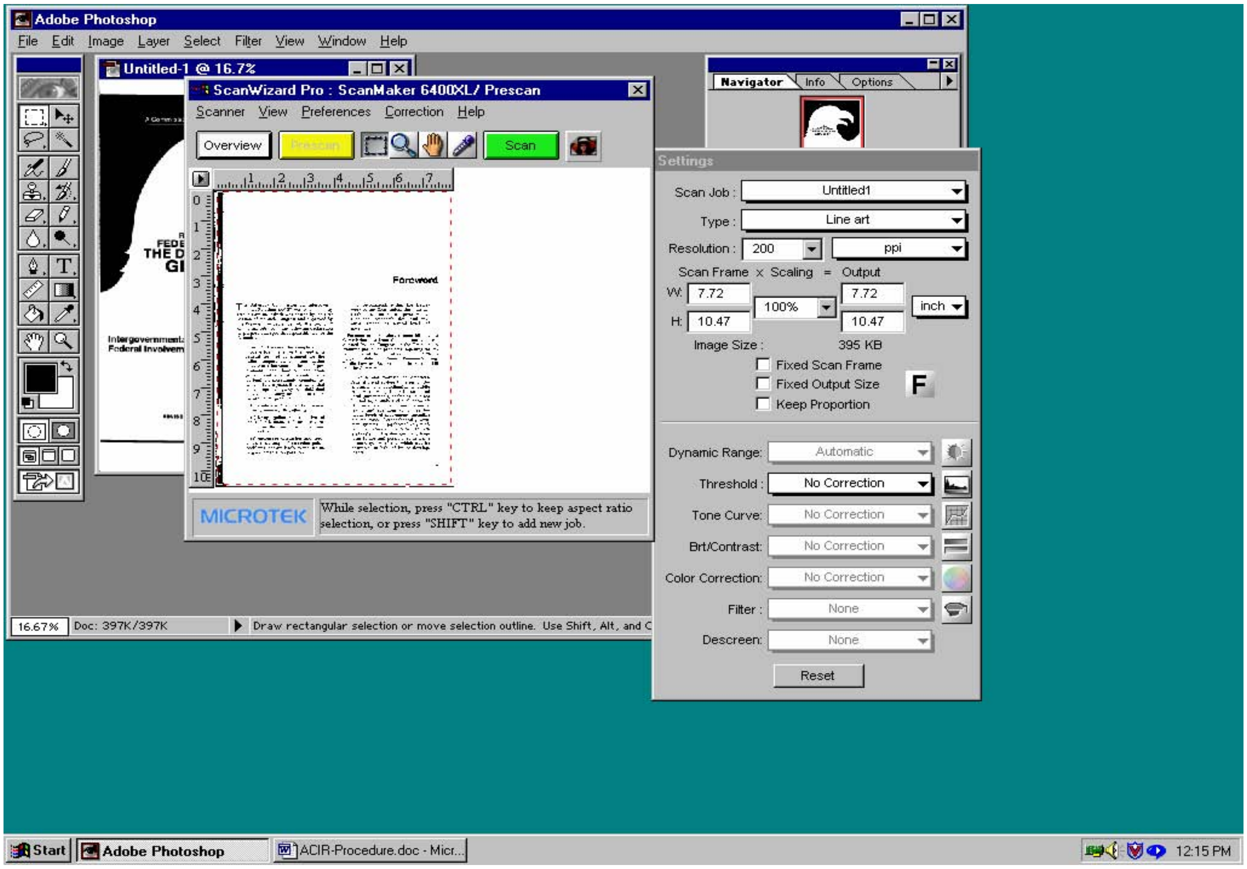
4.2.1 Scanning black and white pages

These are the steps for scanning black and white pages:

- Launch *Adobe PhotoShop*, double-click on the *Adobe PhotoShop icon* or open it from start menu.
- Go to the *File* menu and select *Import [ScanMaker, Fujitsu et al] --TWAIN 32*.
- Verify that the scanning set up is as follows: *300 dpi resolution* and *Line Art* type. Please note that although 300 is an optimum dpi recommended for the best OCR, we experimented with 200 dpi and it worked OK for clean and normal size documents. We have made the assumption that this is because of the recent development in OCR software. The biggest common problems for OCR are colored backgrounds, and smudgy print. If you are scanning tiny characters, smaller than 8 points, then do use more resolution to make the characters be normal size, perhaps up to 400 dpi. Otherwise, for typical text sizes we encourage to use 300 dpi for consistency, long-term preservation and related standard purposes.

- Click on either *Overview* or *Prescan* button and check the position of the document. After selecting the text area (making some preliminary cropping), then click on the *Scan* button. (See illustration below in Figure-1).

- Scan all the black and white pages of the document in the same procedure.


Adobe Photoshop
File Edit Image Layer Select Filter View Window Help
Untitled-1 @ 16.7%
ScanWizard Pro : ScanMaker 6400XL/ Prescan
Scanner View Preferences Correction Help
Overview
Prescan
Scan
Foreword
MICROTEK
Navigator Info Options
Settings
Reset
Start
Adobe Photoshop
ACIR-Procedure.doc - Micr...
12:15 PM

While selection, press "CTRL" key to keep aspect ratio selection, or press "SHIFT" key to add new job.

16.67% Doc: 397K/397K

Draw rectangular selection or move selection outline. Use Shift, Alt, and C

Scan Job : Untitled1

Type : Line art

Resolution : 200 ppi

Scan Frame x Scaling = Output

W: 7.72 × 100% = 7.72 inch

H: 10.47 × 100% = 10.47 inch

Image Size : 395 KB

□ Fixed Scan Frame

□ Fixed Output Size

□ Keep Proportion

Dynamic Range: Automatic

Threshold : No Correction

Tone Curve: No Correction

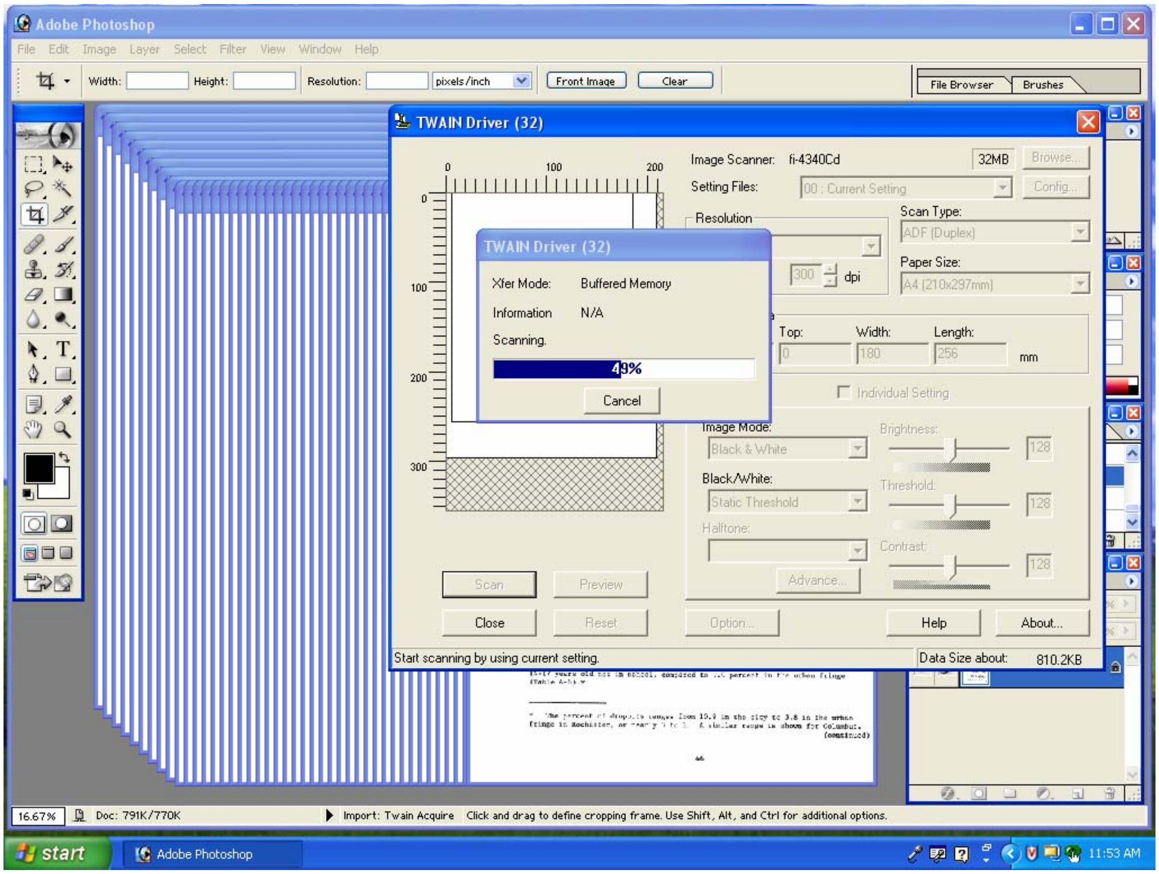
Brt/Contrast: No Correction

Color Correction: No Correction

Filter : None

Descreen: None

Figure-1: Scanning Settings for Black and white page



Width: Height: Resolution: pixels/inch

Image Scanner: fi-4340Cd 32MB

Setting Files: 00 : Current Setting

Scan Type: ADF (Duplex)

Paper Size: A4 (210x297mm)

300 dpi

Top: 0 Width: 180 Length: 256 mm

Image Mode: Black & White

Black/White: Static Threshold

Brightness: 128

Threshold: 128

Contrast: 128

Xfer Mode: Buffered Memory

Information N/A

Scanning.

49%

Start scanning by using current setting.

Data Size about: 810.2KB

16.67% Doc: 791K/770K

Import: Twain Acquire Click and drag to define cropping frame. Use Shift, Alt, and Ctrl for additional options.

Figure-2: Scanning Process Using Duplex Fujitsu Scanner

- When this part of the work is finished, you must *Exit* the scan driver and the next step would be editing (cleaning, further cropping etc) and then saving the files. See the next sections for details, but if you need to remember to save every twenty or so pages, so that you can avoid losing valuable time if the program crashes or there is a power interruption. It is recommended to automate saving processes by using batch saving functions of the Photoshop. Particularly, if you are using the auto duplex Fujitsu Scanner, which can scan 300+ pages at a time, automating the process has paramount importance.

4.2.2 Scanning Color Pages

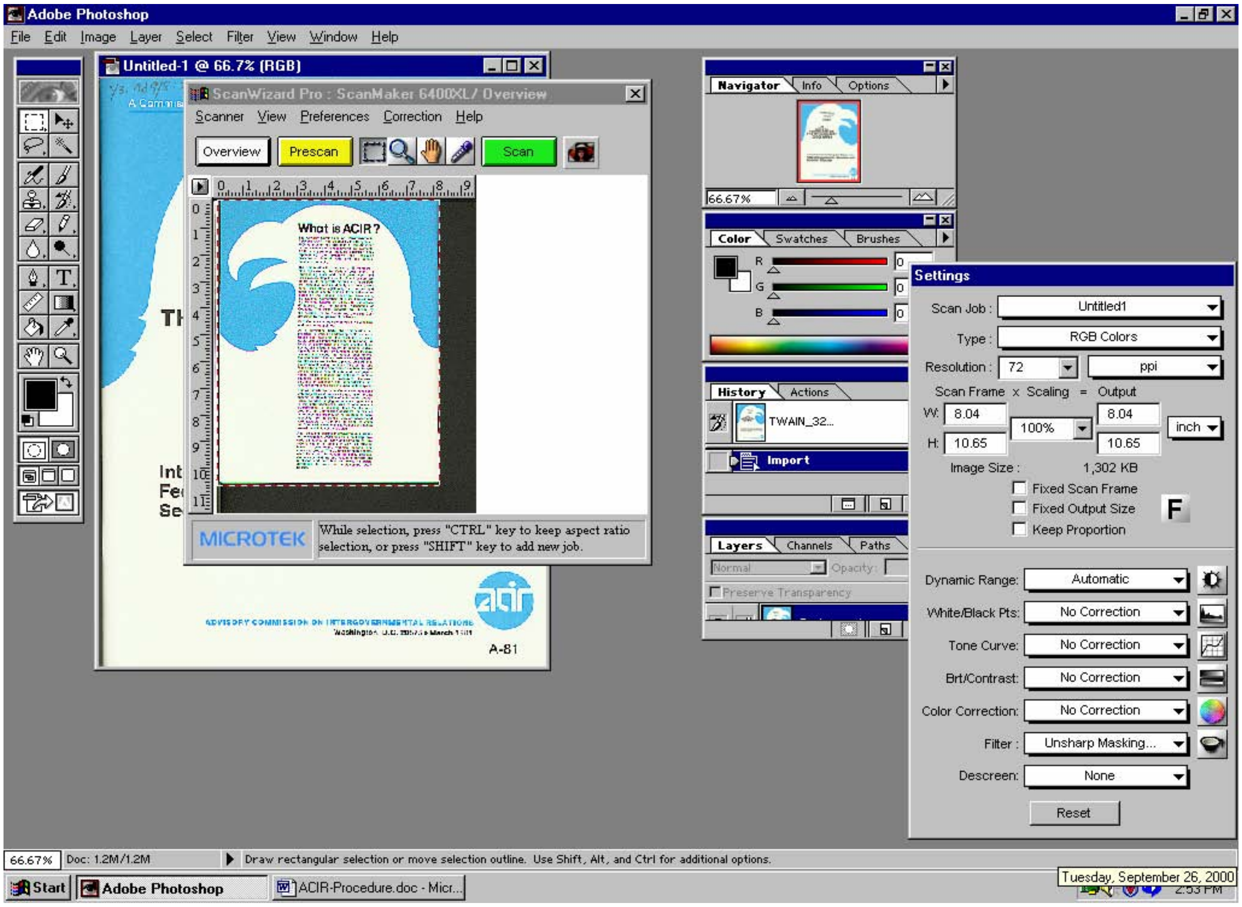
The set up is different for scanning pages containing background colors that would cover up text, any item having a table with background color, or any item with color pictures (e.g.: the cover page). By following the steps outlined below, you can create a graphically pleasing image that is quick to download.

For *color document covers* - which must be especially fast to download and which contain relatively little detailed text –As shown in Figure-2 below, make sure the scanning set up has the following settings:

- *72 DPI*, *Color Photo RGB*, Save as the file, (under a descriptive name, in the same folder on your personal- directory as you did with the black & white scans).

For *color pages other than the front document cover*, set up the driver with the following settings:

- *200 DPI*, Color Photo RGB. Use the filter to make the image file size smaller. Then clean each page and save as cpage# using the true page number of the scanned page in your personal directory as a TIFF file.



While selection, press "CTRL" key to keep aspect ratio selection, or press "SHIFT" key to add new job.

Settings

Scan Job : Untitled1

Type : RGB Colors

Resolution : 72 ppi

Scan Frame x Scaling = Output

W: 8.04 100% 8.04

H: 10.65 10.65 inch

Image Size : 1,302 KB

□ Fixed Scan Frame

□ Fixed Output Size

□ Keep Proportion

Dynamic Range: Automatic

White/Black Pts: No Correction

Tone Curve: No Correction

Brt/Contrast: No Correction

Color Correction: No Correction

Filter : Unsharp Masking...

Descreen: None

Reset

66.67% Doc: 1.2M/1.2M

Draw rectangular selection or move selection outline. Use Shift, Alt, and Ctrl for additional options.

Figure-3: Scanning Settings for color (Cover) pages

4.3 Cropping, Cleaning, and Saving *TIFF* Files

Although it is always good to scan the documents as clear as possible, sometimes it is hard to exclude all ugly edges, property stamps and other spots and marks. While many software applications, apparently provide facilities for cropping and cleaning, by looking at the before and after files sizes (accurately displayed in Microsoft Explorer file manager) it is apparent that most of these do little more than mask the black smudges from view. PhotoShop, however, is an exception since it actually reduces the file size. The trade-off here is that once you erase or crop the file, you normally cannot undo the operation and revert to the previous image, so care is indicated. (See also Section 3.2 of this document for cleaning a blemished page using by *Adobe Acrobat* software).

In order to achieve an economical file size, it is necessary to eliminate the black smudges that normally appear on the margins of page scans. Cropping such types of pages that you already created with *Adobe PhotoShop* involves a number of steps:

- Double-click the crop tool from the toolbar or choose *Image* and select *crop*.
- Drag a cropping rectangle (a square around the text along with any white space you wish to retain).
- If you want to change the page boundaries, you can specify the area to crop by selecting a handle at a corner of the cropping rectangle, and drag to the correct size.
- Click on *crop*.

After completing the necessary editing process (cropping, cleaning etc.), the final step would be saving the file. Make sure all files are being saved as *TIFF* files. You need to save each successive file with appropriate and descriptive file name, (preferably with their actual page numbers), so that you can avoid losing track of their order.

Now the documents (TIFF-files) are ready to convert to PDF. Note that the files should be converted to .PDF only at the time when you are importing them into *Adobe Acrobat* and bind them into a single file. The next section (*Section 5*) of this document, will describe this conversion process. As there is value in retaining the cleaned *TIFF* files, you need to keep them on the departmental directory, until they moved to their permanent location (usually burn to CDs or DVDs). It is the UNT libraries' preservation policy to keep the original *TIFF* files, for their potential archival value. As TIFF files are not compressed, they would be relatively easy to migrate them to a new file format. (See also Appendix-I).

5. CONVERTING SCANNED DOCUMENTS TO PDF

Once the documents are scanned and all the images are saved as *TIFF* on your personal directory, the next phase of the process is creating PDF files, which involves importing *TIFF* files into *Adobe Acrobat* and bind them into a single *PDF* document.

5.1 Creating a Single PDF File from Multiple *TIFF* Files

By using the previous versions of *Adobe Acrobat*, importing and binding the TIFF files into a single PDF document is little bit different process. There were some limitations such as importing 50 pages at a time. Using the current version, *Adobe Acrobat 6.0 Professional*, conversion can be done as follows:

- Go to the *File* menu > Choose Create PDF > and then From Multiple Files, or click the Create PDF button and choose From Multiple Files.
- Click Browse and go to the appropriate File Directory and select the (TIFF) files to be imported and converted
- To change the order of the files in the list, highlight the filename and click the Move Up or Down button. To remove a file click Remove.
- Click OK to convert and consolidate the files into one Adobe PDF document
- Check each page against the print document to insure that they are in their proper page order. The very first page usually interchanges with the last page. You can use Thumbnails option, which display more pages and helps to jump quickly to and edit (move, insert, replace, delete, etc.) a selected page. Moving or deleting thumbnails actually moves or deletes the corresponding page. To show the Thumbnails palette: go to *Window* and click *Show Thumbnails*. You can also click the *show/Hide* Navigation button and click *Thumbnails* Tab.

5.2 Customizing PDF for Ease of Navigation

A PDF document created from the scanned (TIFF) files is like any other PDF document; you can navigate through the document and add enhancements to it. However, for the most efficient workflow, it is best if you implement enhancements for your document as the last stage, after your PDF document is complete in content and organization. Because, any major editorial tasks might cause you to have to redo the navigation procedures all over again.

Therefore cycle through the pages and double check to make sure that all pages are clean. Crop any pages that contain black smudges around their outer edges, (by choosing *Document* and then select *Crop* from the main menu or by selecting the *crop tool* and drag a cropping rectangle and specify the area). By doing so what remains is a clean-looking document surrounded by as much white space as feasible. Please note the fact that the cropping feature in Adobe Acrobat does not delete the cropped part of the page. It only masks that portion of the page so that the viewable portion is made more attractive to the user. Hence, this sort of cropping (unlike Adobe Photoshop) has no effect on file size and on download time.

5.2.1 Creating Internal Document Navigation

Bookmarks allow you to jump within a PDF document (or to another document), among other actions. Although *Acrobat* generates bookmarks automatically from the table of contents of documents (created by most desktop publishing programs), you need to create bookmarks manually for imported images such as TIFF.

To show the bookmarks palette and set up internal document navigation:

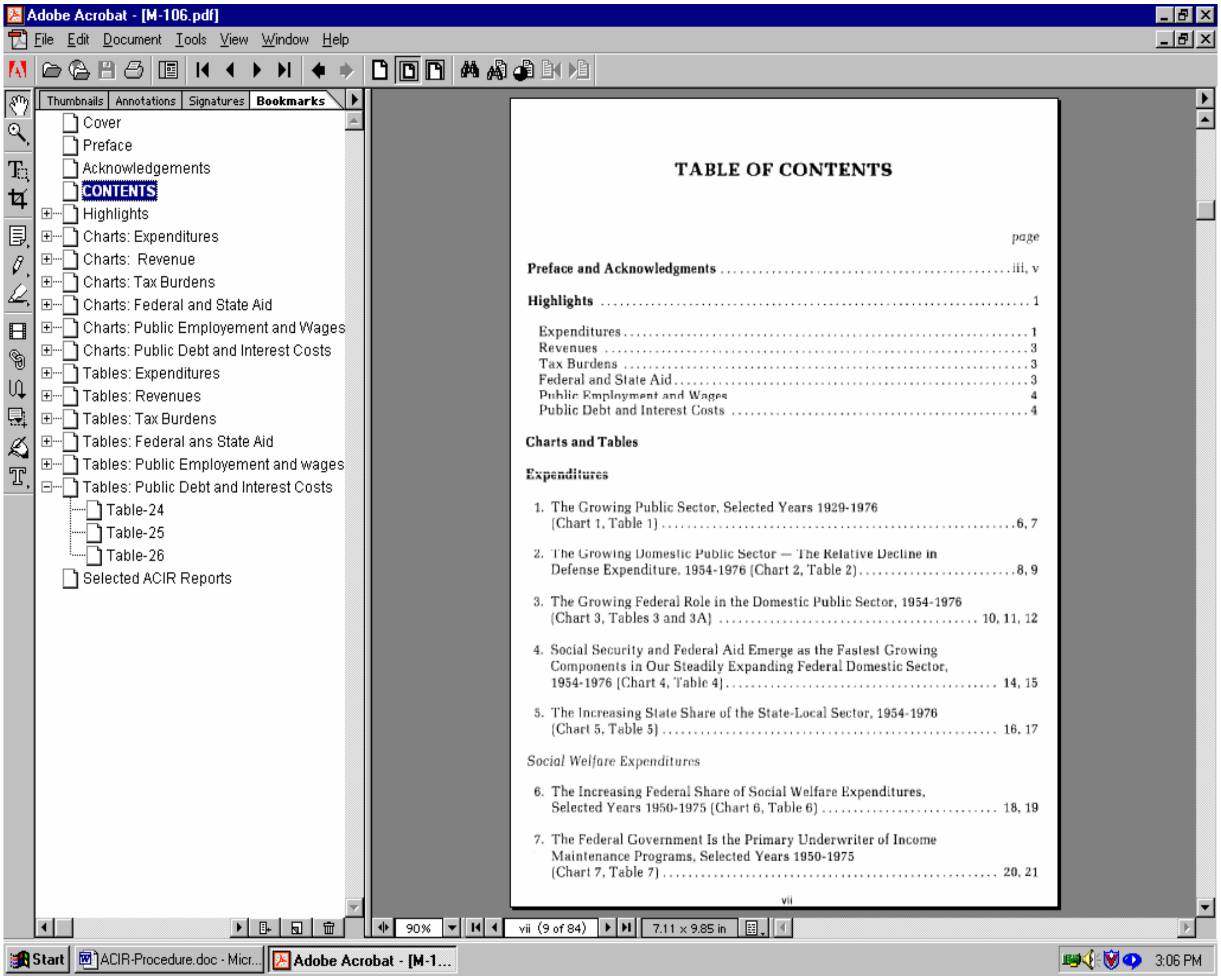
- Go to *Window* menu and click on *show bookmarks*. You can also click the *show/Hide* Navigation button and then click the *bookmarks* tab.
- Cycle through each page, *bookmarking* important pages (Ctrl + B), taking cues from the contents and naming the bookmarks accordingly.
- Order and nest the bookmarks by highlighting each one and dragging it to its proper location. If you correctly nested the bookmark, an arrow will appear next to the upper level bookmark. To select several bookmarks, hold down the shift key while clicking on the page icons next to the bookmarks. You may then drag and drop many bookmarks at once. To better understand this process, in addition to the following figure (Figure-3), it is suggested you actually examine one of our ACIR documents to see how the bookmarks have been set-up.

Once the document has been bookmarked, you are ready to hot link the Table of Contents. To hot link the Table of Contents:

- Select the *Link* tool from the selection bar, whose icon looks like two links of a chain.
- Draw a box around one of the sections. [IMPORTANT: At this point, a dialogue box will pop up].

1) You must select the *"Invisible Rectangle"* key in the *Appearance Type* box, so the system will render the boxes you draw invisible to the end user.
2) You must also ensure that the *Go to View* is selected in the *Action Type box.*
3) Make sure the magnification is set to *Fit in Window*.
4) While the *Create Link* box is still there, using the contents or the page cycle keys, go to the page that you are linking to. When you find the place you want to link to, click the *Set Link* button.

Continue this process for the entire Contents. If the document has a back-of –the-book style *Index*, then you will need to repeat the steps outlined above for each indexing term. Then, as depicted in Figure-4 below, the next step would be to specify the display options.



- Cover
- Preface
- Acknowledgements
- CONTENTS
- Highlights
- Charts: Expenditures
- Charts: Revenue
- Charts: Tax Burdens
- Charts: Federal and State Aid
- Charts: Public Employement and Wages
- Charts: Public Debt and Interest Costs
- Tables: Expenditures
- Tables: Revenues
- Tables: Tax Burdens
- Tables: Federal ans State Aid
- Tables: Public Employement and wages
- Tables: Public Debt and Interest Costs
  - Table-24
  - Table-25
  - Table-26
- Selected ACIR Reports

TABLE OF CONTENTS

page

Charts and Tables

Expenditures

Social Welfare Expenditures

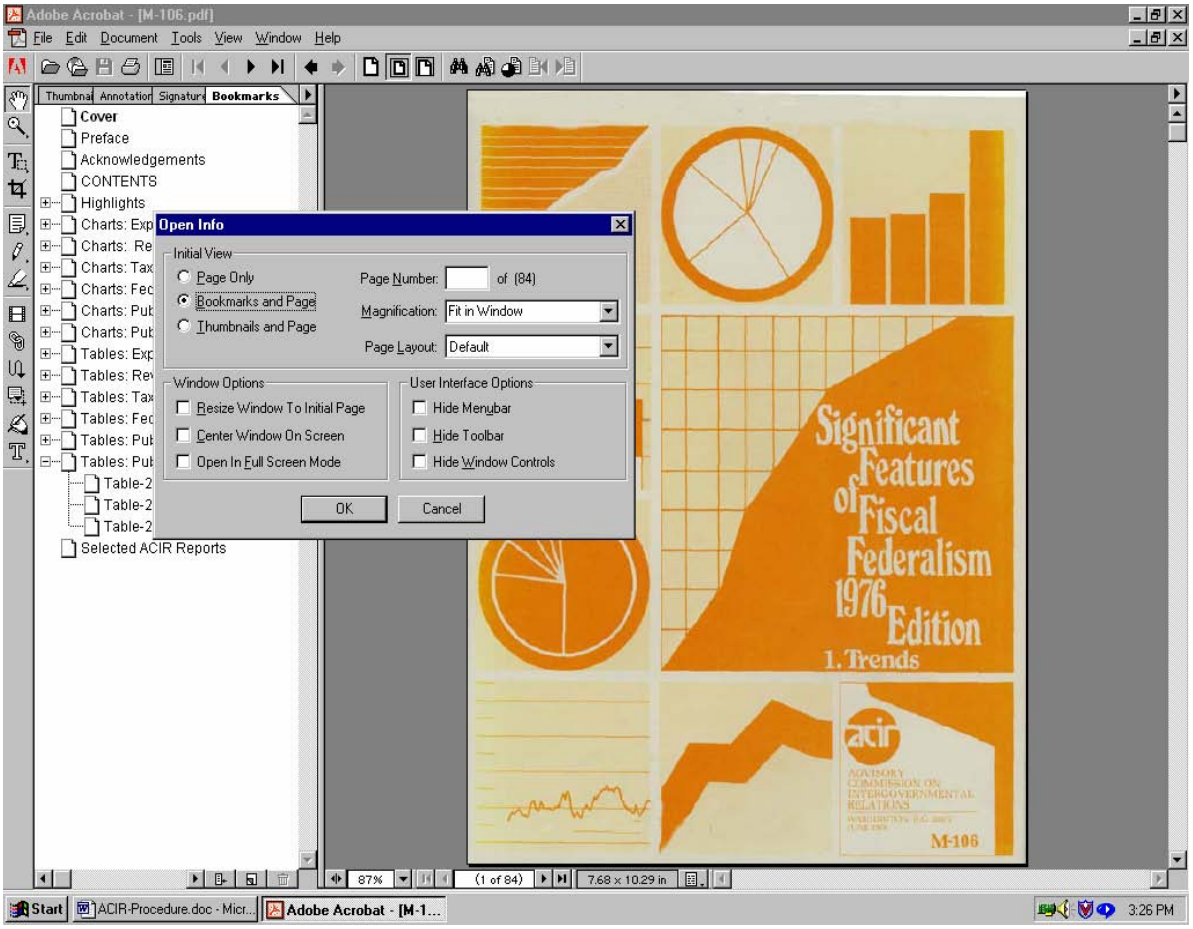
vii

Figure-4: PDF document: with hot links & bookmarks

5.2.2 Specifying Opening View

Once the document is linked, then the next most important step is to determine how the document will appear to the user when it is viewed. To specify how the page will display to the user:

- Go to the *File* menu and select *Document Info* and then choose *Open*
- *Select Bookmarks and Page* for the *Initial View*. (This opens the navigation pane with bookmarks in front).
- Under *Page Number* enter an opening page number. (Usually Title Page)
- For *Magnification* select *Fit in Window*. (The page fit entirely in the window)
- For *Page Layout* accept the *default*. Then click *OK*.



Open Info

Initial View

○ Page Only
● Bookmarks and Page
○ Thumbnails and Page

Page Number: of (84)
Magnification: Fit in Window
Page Layout: Default

Window Options

□ Resize Window To Initial Page
□ Center Window On Screen
□ Open In Full Screen Mode

User Interface Options

□ Hide Menubar
□ Hide Toolbar
□ Hide Window Controls

OK Cancel

Figure-5: PDF document with Display options

5.3 Adding Metadata to the PDF File

So far you did all possible enhancements and determine how the document will appear to the users when it is viewed. The next step, perhaps the most important step, is to set up how your document will appear to the search engines. This process enhances access and discovery of the digital documents by the users. Appendix-IV, of this document also provides detail description with specific examples. To add searchable Information and set binding:

- Go to the *File* menu, select *Document Info*, and choose *General*.
- Enter the *Title* of the document, the *Subject*, the *Author*, and *keywords* with a comma and no space. Please note that many Web search engines use the title to describe the document in their search results list. If you do not provide a title, the filename will appear in the results list instead. It is currently considered best practice to not to repeat any one word more than four times, since search engines interpret metatags that repeat words too many times as an attempt to spoof them and toss them out. When you are done, click OK.

5.4 Capturing Pages to Convert to Searchable Text

The next phase of the work involves “capturing” the text to make the document searchable. Since text capturing process is a memory-intensive operation, it is suggested that no other program be running on your computer. To use the Capture Pages command, here are the steps:

- Choose the *Tools* menu > select *Paper Capture* > then select *Capture Pages*.
- Go into *Preferences* in the pop-up box and make sure that *Primary OCR Language* is "English (US)" and "Original Image with Hidden Text" is selected in the *PDF Output Style* box and that *Downsample Images* is enabled.
- Select whether to capture all pages, the current page only, or a range of pages. Capturing all pages is preferable; however, some pages that were scanned at less than 200 DPI, (such as the cover and blank pages) cannot be captured. Thus, you will need to begin text capture starting at page 2 and click 'skip' button for the blank pages.
- Page capture may be stopped and saved partway through the process and then be restarted again, at the point where the process was halted. If you attempt to capture a page twice, you will see an error message. Occasionally, pages, for whatever reason, will not be captured. Provided these are relatively few, simply skip these pages and proceed.

When you capture pages, as can be seen from Figure-5 below, Adobe Acrobat performs optical character recognition (OCR) and font and page recognition on the pages and assigns a confidence rating to each word it finds. For the Capture Pages command to be able to recognize text, the text must be upright and not rotated. Please also note that you can only use capture pages for images, which were scanned with the resolutions from 200 to 600 dpi. For most pages, scanning at 300 dpi produces the best captures. See also Appendix-V for further information on capturing and related issues.

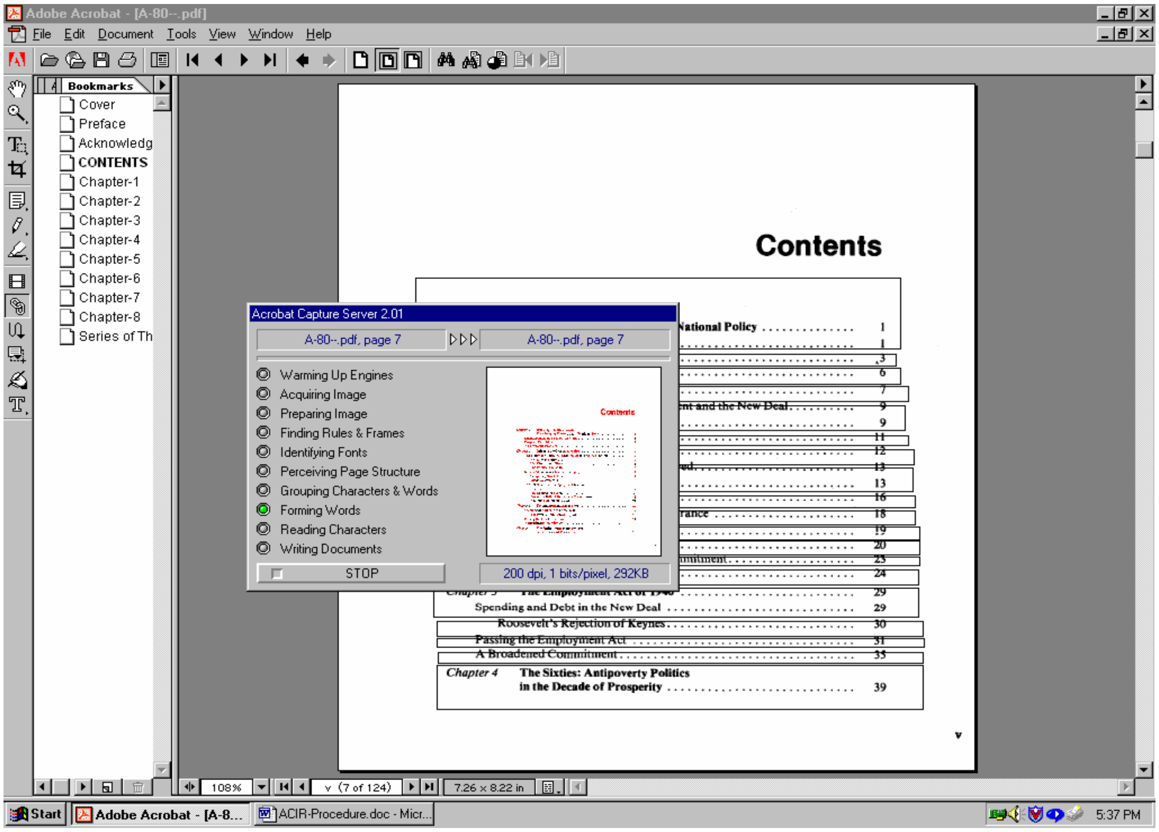
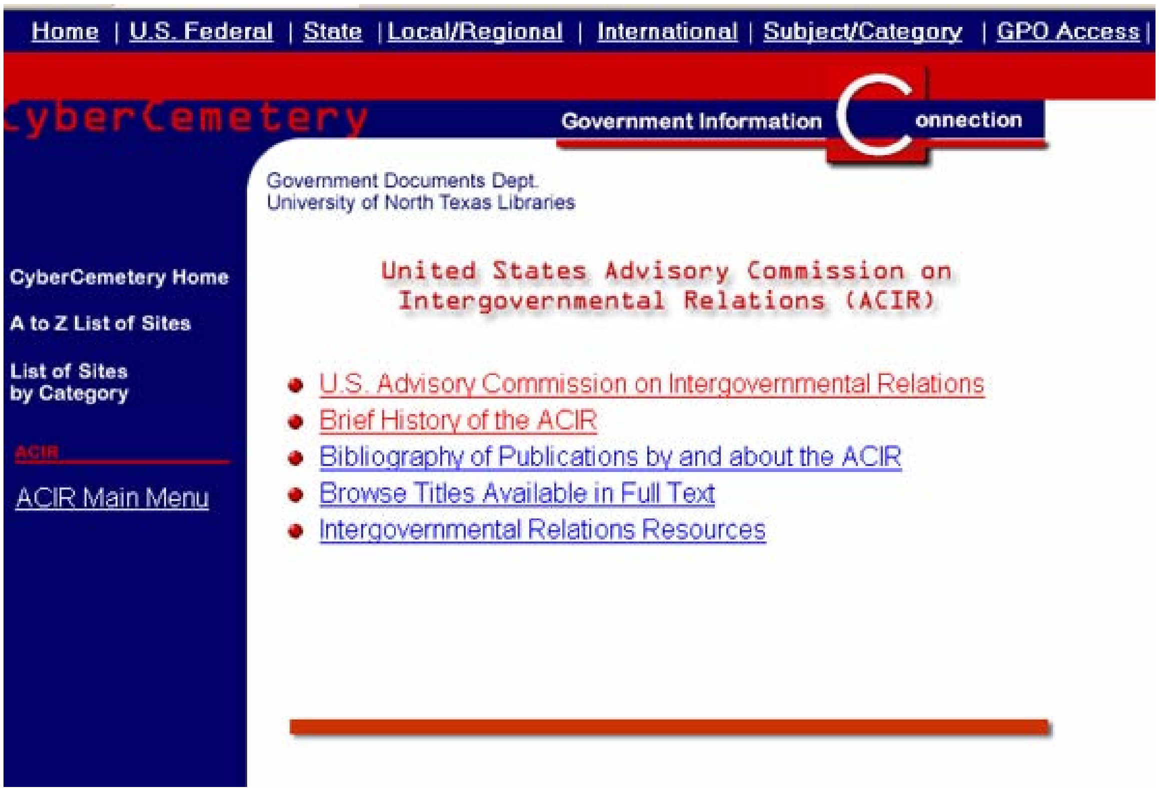


Figure-6: Text Capturing (OCR) process

5.5 Naming and Saving PDF Documents

Before you save your final copy, once again check to make sure that the content and layout in the documents are complete and correct, and that all links, bookmarks, and other enhancements are in place. Currently Digital project department of the UNT Library is exploring the feasibility of library wide file naming convention. In the meantime, we encourage continuing to use the old Microsoft Disk Operating System (MS-DOS) file naming convention, which is up to 8 characters followed by 3 extension. This relatively shorter but descriptive file-naming system, in addition to helping proper in-house file management, it also avoids truncation of long filenames, which is common in many network programs. More importantly, as Microsoft Windows have been configured to associate "PDF" with Acrobat or the Web browser plug-in, it helps to launch the appropriate reader application.

Save the final copy of the document by selecting the *Save As* menu item and making sure that the *Optimize box* is checked. Save it to the appropriate directory on the appropriate drive, (preferably on the shared H or P drive, which are backed up.). Normally, as you work, you will wish to save the document often, by using Ctrl + S or Save. While this will not optimize the document for byteserving, it will ensure that, in the event of a system crash, you do not lose your all of work. To make the PDF files small enough for network distribution, in addition to optimizing files, instead of distributing one large document, it is recommended to distribute a collection of small documents with links between them. Once you save your final version to the appropriate directory, then the document is ready to be uploaded on the server.


Home | U.S. Federal | State | Local/Regional | International | Subject/Category | GPO Access |
CyberCemetery
Government Information Connection
Government Documents Dept.
University of North Texas Libraries
CyberCemetery Home
A to Z List of Sites
List of Sites by Category
ACIR
ACIR Main Menu

United States Advisory Commission on Intergovernmental Relations (ACIR)

- U.S. Advisory Commission on Intergovernmental Relations
- Brief History of the ACIR
- Bibliography of Publications by and about the ACIR
- Browse Titles Available in Full Text
- Intergovernmental Relations Resources

Figure-7: The UNT Libraries' ACIR Website (as in January 2004) at:
http://www.library.unt.edu/gpo/Acir.html

6. Summary

This *Guide Document* has already described the digitization process and it is clear that the process is resource intensive. Therefore it is imperative to manage the digital resources effectively so that their long-term access and preservation will be ensured. The UNT library's Digital Projects Department is in the process of creating an infrastructure or archive system that integrate all its digital resources centrally and describe them with detail and quality metadata to ensure that their long term access is preserved. In light of this endeavor, (implementing digital management system), this *Guide Document* plays its part by promoting standardization at the very early (creation) stage of the workflow in the life cycle of the digital resources.

=== // ===

APPENDICES

Appendix–I: Glossary

ACIR – The **A**dvisory **C**ommission on **I**ntergovernmental **R**elations.

- This Federal commission underwent several changes of name, initially being known as the Hoover Commission, after President Hoover, who appointed its first members. Until it ceased operation in 1996, the *Advisory Commission on Intergovernmental Relations* served as a vital source of information about the impact of federal programs on state and local governments. In January 1998, *The American Prospect* wrote, “Often the ACIR was the only place people interested in urban policy could look to for data about city versus suburban fiscal resources. The commission's members included governors, mayors, state legislators, county officials, and members of Congress—all of whom served two-year terms. As a nonpartisan group, the commission earned a reputation for producing accurate and unbiased reports.” It remains an important source of historical information, particularly for social scientists. A brief history of the agency is available online at http://www.library.unt.edu/gpo/ACIR/acirhist.html

JAWS: is acronym for Job **A**ccess **W**ith **S**peech. (See also Appendix-.III)

- JAWS provide speech technology to provide access to today’s popular applications for blind and visually impaired individuals. It installs an enhanced, multi-lingual software speech synthesizer, "Eloquence for JFW" that works with today's standard sound cards. JAWS supports popular applications such as e-mail programs, word processors, spreadsheets, web browsers, project management and research tools, contact management software, presentation software, web development tools, software development tools, database management software, sound editing software, and much more. You can find a complete list of application and support at: http://www.synapseadaptive.com/henter2/default.htm

TIFF - Acronym for ***T**ag (ged) Image **F**ile **F**ormat*.

- TIFF is one of the most widely supported file formats for storing bit-mapped images on personal computers .TIFF graphics can be any resolution, and they can be black and white, gray-scaled, or color. Files in TIFF format often end with a .tif extension. Every choice of file format entails certain trade-offs. An important characteristic of TIFF files is that they store data in an uncompressed format. This makes for a very large file size, but also means that the files are relatively robust from an archival standpoint. This is because uncompressed file formats are far easier to migrate to new file format standards, as these become available.

PDF - Stands for ***P**ortable **D**ocument **F**ormat*.

- PDF is a cross-platform file format developed in 1991 by Adobe Systems. PDF captures formatting information from a variety of desktop publishing applications, making it possible to have compressed, formatted documents appear on the recipient's monitor or printer precisely as they were intended. This is accomplished, in part, by having the requisite fonts embedded in the actual PDF file. The document essentially contains everything needed to display itself, with nothing assumed about the end user’s computer, except that he or she will have downloaded and installed the Adobe Acrobat *Reader*, a free application distributed by Adobe Systems. It is interesting to note that, according to information at the Seybold Conference Website, “…120 of the Federal government agencies have adopted PDF as a standard and in the regulatory agencies, this is having a very deep impact into the way that the government is going paperless. These agencies include the IRS, the SEC, patent office, Library of Congress many others that use PDF on a day-to-day basis to communicate information.”

Appendix–II: Conceptual Issues

The step-by-step instructions that are blocked out below pertain to a particular set of software tools and represent a best attempt to create an efficient workflow process within the constraints that existed at the time the process was devised. As new products come onto the market and/or as existing tools are upgraded, with new and different capabilities, this process may need to be updated. For this reason, it is vital to try and gain a clear understanding of the work at a conceptual level. This section attempts to illuminate some of the core concepts underlying and relevant to the ACIR digitization project. You will find it fruitful to also read Appendix-iv, which comments on the key capabilities of various software applications.

- **File Size**. File size is very important. In a .PDF file that is placed on the Internet, the file size determines how long the user has to wait to download the document. The first page is especially important, since, if it does not load quickly, the user may decide it is not worth the investment of time needed to view the document. In documents that have been scanned by an outside source, file size may be more difficult for us to control. In the documents that you scan, steps can be taken to greatly reduce file sizes, while maintaining the quality of the document within reasonable limits. There are two main ways this is achieved. First, for scanning we mainly use *OmniPage Pro*, a product whose native scanning driver has been highly optimized for scanning black & white text. (Therefore we begin with a small size, around 50 kilobytes or so in size.) Second, we use *Adobe PhotoShop* to crop away any black smudges that show up on the perimeters of our page scans. (Just eliminating these dense, black portions of the image considerably reduces file size. According to discussions on the ARL-Ereserve list in April 1999, this step can reduce file sizes by around 20%.).

- **Byte Serving.** This feature permits long .PDF documents to be downloaded and viewed page-by-page. Neither internal hotlinks nor bookmarks are adversely affected by byte serving. However, to benefit from this feature. As you create .PDF files, from .TIFFs, you are instructed to follow this sequence of actions:

  1) When saving a document use the *Save As* function.
  2) In the pop up box, make sure that *Optimize* is checked.
  3) After you click on *Save*, it may take a few moments for Adobe to reorder certain internal, hidden elements of the file.

  It is these three steps that ready the for byte serving. In addition, the Web server must support byte serving (which our server does) and the user must have options on the Adobe Acrobat Reader plug-in correctly configured. Details are set forth in technical documents at Adobe's support Web site: http://www.adobe.com/support/techguides/acrobat/byteserve/byteservmain.html and also available at the UNT Libraries' online help document at: http://www.library.unt.edu/gpo/acir/technicaldoc.htm .

Appendix–III: ADA Compliance

One of our goals is to make the documents that we are placing on the web accessible to the visually impaired. Aside from being an inherently valuable thing to do, it is required by the *Americans with Disabilities Act of 1990*. We have an adaptive computer lab on campus that currently runs the speech software "Jaws." So far as we aware, this is the only product currently available that can read a .PDF file. This it can do, provided that the image has text "in the background". (There are many parts of the .PDF file that are not visible. Words are available for searching and/or for cutting and pasting provided that the text "capture" process has been run.)

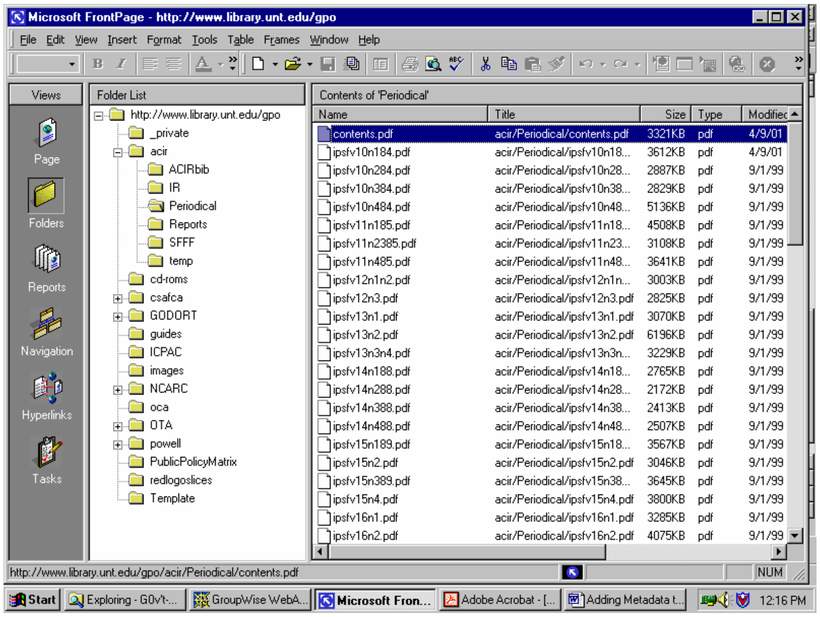
In order to get the document in a format that the speech software can understand, follow this process (Appendix B, Section 1):

1. After a document has been through the capture process, under the Edit menu, select *Copy File to Clipboard* (Ctrl. + Shift + K).
2. Open a program that can read the RTF file format (e.g., MSWord, WordPad, etc.)
3. Paste into this program (Ctrl. + V). The text from the document should appear on the screen.
4. Clean up any obviously bad areas.
5. Save this as ASCII text under the same name as the PDF on the shared directory.
   (There would then be two files created, "filename.pdf" and "filename.txt".)

A service that analyses Web pages and reports on their ADA compliance may be found at http://www.cast.org/bobby/

Appendix-IV: Adding Metadata to ACIR files

This section, in continuation with Section 5.3 of this document, provides information on how to add Metadata to ACIR PDF files. As can be seen from the above screen print, the ACIR Web directory contains a number of sub directories. Most of the scanned documents are reside in “Periodical” and “SFFF” directories of the Web server. The following sample indicates the metadataizing process of the “Intergovernmental Perspective” and “Significant Features of Fiscal Federalism” documents of the ACIR publications.



Contents of 'Periodical'

Name	Title	Size	Type	Modifiec
contents.pdf	acir/Periodical/contents.pdf	3321KB	pdf	4/9/01
ipsfv10n184.pdf	acir/Periodical/ipsfv10n18...	3612KB	pdf	4/9/01
ipsfv10n284.pdf	acir/Periodical/ipsfv10n28...	2887KB	pdf	9/1/99
ipsfv10n384.pdf	acir/Periodical/ipsfv10n38...	2829KB	pdf	9/1/99
ipsfv10n484.pdf	acir/Periodical/ipsfv10n48...	5136KB	pdf	9/1/99
ipsfv11n185.pdf	acir/Periodical/ipsfv11n18...	4508KB	pdf	9/1/99
ipsfv11n2385.pdf	acir/Periodical/ipsfv11n23...	3108KB	pdf	9/1/99
ipsfv11n485.pdf	acir/Periodical/ipsfv11n48...	3641KB	pdf	9/1/99
ipsfv12n1n2.pdf	acir/Periodical/ipsfv12n1n...	3003KB	pdf	9/1/99
ipsfv12n3.pdf	acir/Periodical/ipsfv12n3.pdf	2825KB	pdf	9/1/99
ipsfv13n1.pdf	acir/Periodical/ipsfv13n1.pdf	3070KB	pdf	9/1/99
ipsfv13n2.pdf	acir/Periodical/ipsfv13n2.pdf	6196KB	pdf	9/1/99
ipsfv13n3n4.pdf	acir/Periodical/ipsfv13n3n...	3229KB	pdf	9/1/99
ipsfv14n188.pdf	acir/Periodical/ipsfv14n18...	2765KB	pdf	9/1/99
ipsfv14n288.pdf	acir/Periodical/ipsfv14n28...	2172KB	pdf	9/1/99
ipsfv14n388.pdf	acir/Periodical/ipsfv14n38...	2413KB	pdf	9/1/99
ipsfv14n488.pdf	acir/Periodical/ipsfv14n48...	2507KB	pdf	9/1/99
ipsfv15n189.pdf	acir/Periodical/ipsfv15n18...	3567KB	pdf	9/1/99
ipsfv15n2.pdf	acir/Periodical/ipsfv15n2.pdf	3046KB	pdf	9/1/99
ipsfv15n389.pdf	acir/Periodical/ipsfv15n38...	3645KB	pdf	9/1/99
ipsfv15n4.pdf	acir/Periodical/ipsfv15n4.pdf	3800KB	pdf	9/1/99
ipsfv16n1.pdf	acir/Periodical/ipsfv16n1.pdf	3285KB	pdf	9/1/99
ipsfv16n2.pdf	acir/Periodical/ipsfv16n2.pdf	4075KB	pdf	9/1/99

Figure-8: ACIR Directory

To create metadata for PDF document, open the PDF document in Adobe Acrobat, choose the *File* menu, select *Document Info*, and choose *General*. Then you will find a metadata creation tool and then enter the following information (for this specific document category) under the Document Information in the corresponding fields:

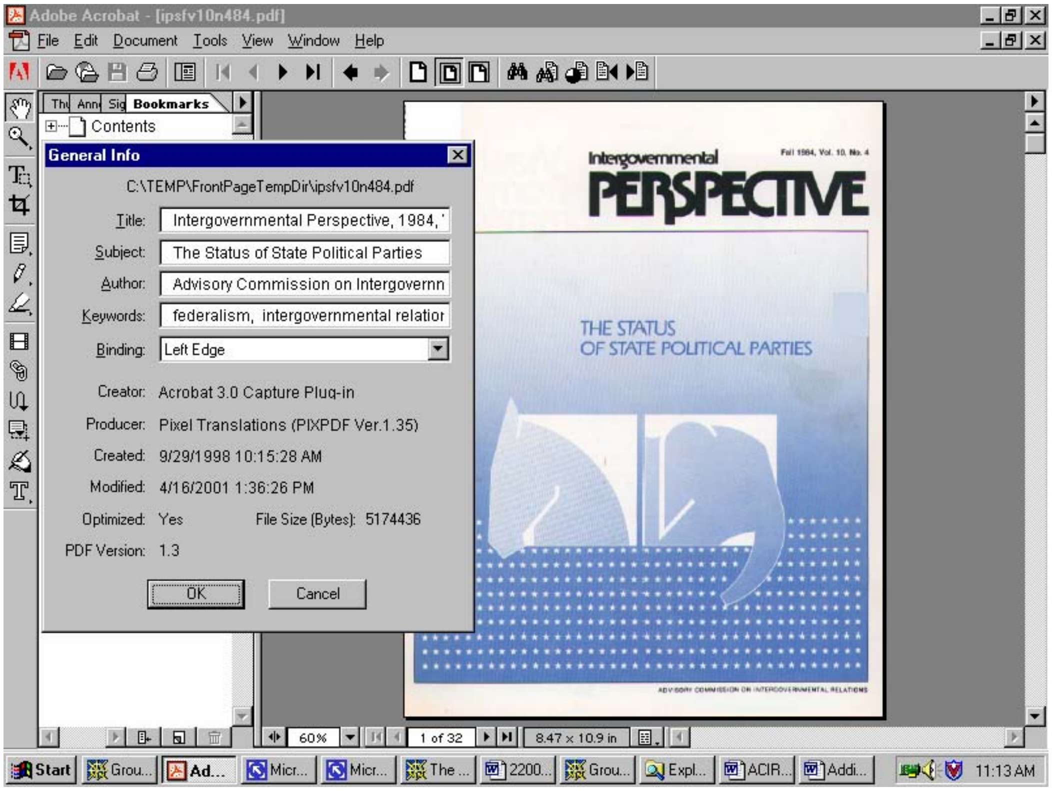
Title: Intergovernmental Perspective, 19##, V.##, No.#

Subject: *[sub-title of the issue]*

Author: Advisory Commission on Intergovernmental Relations (ACIR)

[*]**Keywords:** federalism,intergovernmental relations,United States of America, *[up to five other specific key words as needed]*.

Binding: Left Edge



General Info

C:\TEMP\FrontPageTempDir\ipsfv10n484.pdf

Title: Intergovernmental Perspective, 1984, '

Subject: The Status of State Political Parties

Author: Advisory Commission on Intergovernn

Keywords: federalism, intergovernmental relatior

Binding: Left Edge

Creator: Acrobat 3.0 Capture Plug-in

Producer: Pixel Translations (PIXPDF Ver.1.35)

Created: 9/29/1998 10:15:28 AM

Modified: 4/16/2001 1:36:26 PM

Optimized: Yes File Size (Bytes): 5174436

PDF Version: 1.3

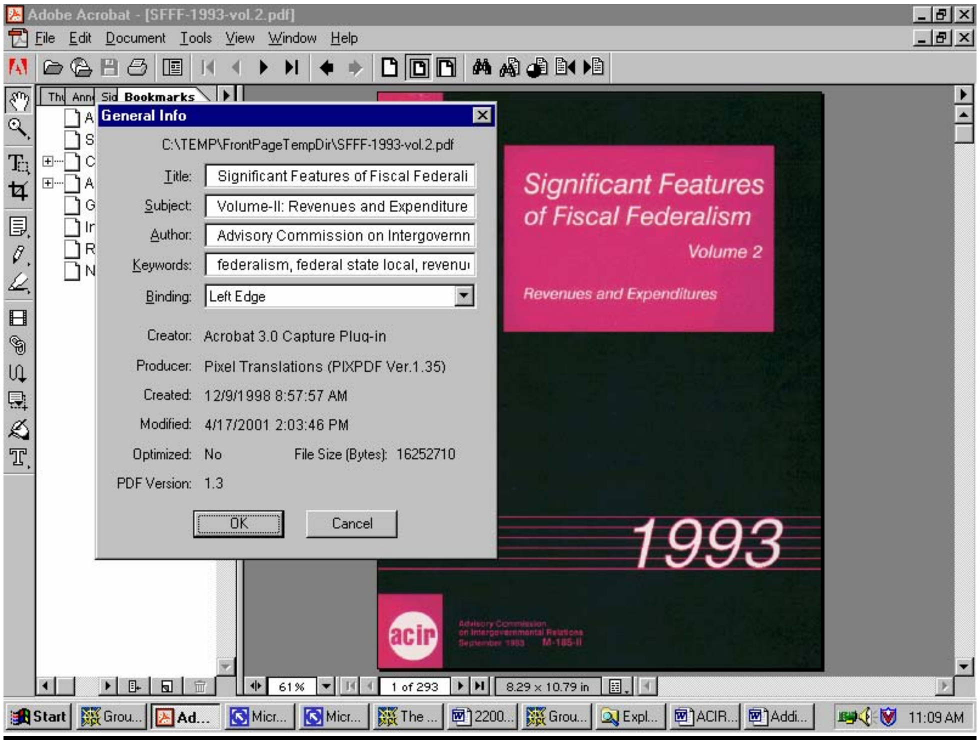
OK Cancel

Figure-9: Creating Metadata for ACIR Publications-I

[*] Depending on the subject of the specific issue, in addition to the above common keywords, up to 5 main words can be added, [such as budget, revenue, recession, state, urban, cities,

Similarly, use the same procedure for “Significant Features of Fiscal Federalism”.

Title:	Significant Features of Fiscal Federalism *yyyy* to *yyyy Edition*.
Subject:	*[sub-title of the issue]*
Author:	Advisory Commission on Intergovernmental Relations (ACIR)
Keywords:	federalism,federal state,revenues,debt,expenditures,fiscal issues,United States of America,*[up to five other specific key words as needed]*.
Binding:	Left Edge


Adobe Acrobat - [SFFF-1993-vol.2.pdf]
File Edit Document Tools View Window Help
Bookmarks
Significant Features of Fiscal Federalism
Volume 2
Revenues and Expenditures
1993
acir
61%
1 of 293
8.29 x 10.79 in
Start
11:09 AM

General Info

C:\TEMP\FrontPageTempDir\SFFF-1993-vol.2.pdf

Title: Significant Features of Fiscal Federali

Subject: Volume-II: Revenues and Expenditure

Author: Advisory Commission on Intergovernn

Keywords: federalism, federal state local, revenu

Binding: Left Edge

Creator: Acrobat 3.0 Capture Plug-in

Producer: Pixel Translations (PIXPDF Ver.1.35)

Created: 12/9/1998 8:57:57 AM

Modified: 4/17/2001 2:03:46 PM

Optimized: No File Size (Bytes): 16252710

PDF Version: 1.3

OK Cancel

Figure-10: Creating Metadata for ACIR Publications-II

In addition to enhancing the accessibilities, adding metadata for old PDF files and saving or resaving them as optimized PDF files with the newer available hardware/software versions, would also contribute to the preservation strategies by refreshing the old files to be accessed at least by the next versions/generations technologies.

Appendix–V: Notes on Software Applications

This appendix contains notes about various software applications that have been used in digitizing ACIR documents. It attempts to explain the underlying rationale for why we have chosen to use this particular set of tools. Undoubtedly, as new products and new versions of existing products come onto the market, the choice of which tools to use in which instances will, necessarily, change. Ideally, in future it will be possible to achieve a more efficient process by employing fewer tools and steps to achieve the desired result.

- **Adobe Acrobat (Version 4.0):** This main functionality offered by this program is that it enables us to bind separate *.PDF files into one master document and, by using the free *Capture* plug-in, it enables one to index any recognizable text that has been scanned. (Note that there is also a standalone product called *Adobe Capture,* which is intended for use by anyone who is capturing a large number of pages. The plug-in is not really meant for large jobs; however, it is serviceable.) Another capability that is highly important to this work is that Exchange enables us to set up internal hotlinks within the bound .PDF file. It also permits one to set up bookmarks that display in a frame along the left side of the body of the document, in a manner similar to Microsoft Word’s document map view.
- **Adobe Acrobat Reader:** This free plug-in is available from Adobe in many versions, including versions that support Windows, Apple and Unix systems. The reader extends the capabilities of Web browsers such as Netscape Communicator, Microsoft Internet Explorer and Opera so that they are able to open and display .PDF files.
- **Adobe PhotoShop:** This tool is used to capture any color pages, since *OmniPage Pro* does not do a good job capturing these. Note that the files thereby created are rendered as black & white images, according to the step-by-step instructions supplied above. It is also used to crop .TIFF files since, unlike, say, Adobe Acrobat, it actually reduces the size of the file, and does not merely mask the black smudges, which often show up at the perimeter of file scans. Note that,

while PhotoShop can in theory be used to scan black & white pages, in practice we find that the previous version of PhotoShop tends to generate bloated files which are too large for fast downloading, even when a great deal of effort goes in to tweaking the scan driver for optimal results. (Small file sizes are defined, for present purposes, as being in the neighborhood of 50 or so kilobytes.) However the recent version has significantly improved and, since September 2000, we have been scanning all (black and white and color) pages by using Photoshop software (version 5.5 or the latest).

- **OmniPage Pro (Version 9.0):** This program is designed to do essentially one thing, namely, to scan text-based documents. Omni Page Pro uses Optical Character Recognition (OCR) to scan printed pages into a readable and editable format. Its great strength is that it creates a very small file, typically smaller than 50 kilobytes. To do this, it must be able to communicate with the scanner using its native TWAIN driver, and not the TWAIN driver of the scanner's manufacturer. A list of supported drivers appears at the Web site http://www.caere.com . Given sufficient RAM, it is also very fast. While OmniPage does offer a text recognition feature, we do ***not*** use this capability, because Adobe Acrobat does not know how to work with such files. Thus, you must be certain that this option is marked *"defer until later"*. Another excellent feature of OmniPage Pro is that provides a deskewing mechanism. This is a highly useful feature, since it compensates for the skewing of text that naturally occurs when a book is laid face down upon a flatbed scanner. In addition to creating a more visually appealing file, it makes machine text recognition (i.e., the Adobe "capture" process) far more accurate. More information on Omni Page Pro, available at: http://www.caere.com/products
- **Microsoft Word:** You can save a Word file as a PDF with the use of a plug-in. The plug-in introduces a box on the toolbar that will show a PDF symbol and give its status. To save a Word file to a PDF, simply click this button. (This capability is not directly relevant to the ACIR project, but is a useful thing to know in general.) Of course, this document is originally created as a word file, and saved as a PDF file by using the aforementioned method.

Appendix–VI: Known Problems

- **Unknown Error While Saving. Cannot Save [9] or [10] (or similar error message).** This error usually occurs if you have made a lot of modifications to a document without saving using just the Ctrl. + S save. It will occur when trying to use the *Save As* feature during the final pass of Adobe's *optimize* cycle. After it happens you will not be able to save using either Save or *Save As.* The greatest likelihood is that your document was in fact saved. The error arises in the final stage of the *optimize* cycle, as the program is checking for errors. Simply close and re-launch Adobe Acrobat. Open your document and then, once again, invoke the *Save As* command from the relevant menu. This should enable you to recover the document.
- **Page Contains Image Plus Text Cannot Capture.** This is a strange error that occurs spontaneously and seemingly without explanation. Other times it may occur because you have already captured this page, or the page has images that have been placed on it. The best course is perhaps simply to skip the offending page or pages and resume capturing text from successive pages. Of course, if the number of pages so affected is significant, other measures, including re-scanning the document, may be indicated.
- **Document is not byteserving when tested on the Internet.** It may be that the document was not properly optimized. You can check to see if a document is optimized by selecting *Document Info* under the *File* menu and then choosing *General*. At the bottom of the pop-up box it will say whether the document is optimized. If it is not, pull it from the shared directory and do a *Save As* with *optimize* checked. Another problem can be that the client does not have the proper configuration selected for byteserving. Instructions regarding this are at: http://www.library.unt.edu/gpo/acir/technicaldoc.htm

Appendix–VII: Useful Sites and Email Lists

➧ **Adobe**: http://www.adobe.com/homepage.shtml

This site is also a great resource. The site is searchable and this is usually my first stop for information regarding specific error messages. There are many technical and white papers. There are also error message reports, all of which are called up by the search of the entire Adobe site.

➧ **Archive Builders**: http://www.archivebuilders.com/

This site provides useful information on document management, document imaging systems and digital libraries in general.

➧ **Association of Research Libraries, Electronic Reserves Email List (ARL-EReserve):** http://www.cni.org/Hforums/arl-ereserve/

A forum for discussion of issues surrounding management of electronic reserve within libraries. Encourages discussion of hardware and software selection, policy development, copyright concerns, project reports and case studies are welcome. A searchable archive of previous postings to the list is available. The librarians on this list are often quite knowledgeable about .PDF technology and the various Adobe products are frequently under discussion.

➧ **BlueWorld:** http://www.blueworld.com/blueworld/lists/acrobat.html

A discussion list for Adobe products, including a searchable archive of messages previously posted to the list.

➧ **Digital Library**: http://www.dlib.org/

This site provides useful information on global Digital Library Researches. It provides access to ***D-Lib Magazine*** and information on other D-Lib activities.

- **PDF Research Companion:** http://www.performancegraphics.com/
  This site has some problems with their frames. You will need to break the frames using the right mouse button to *Open in a New Window*. The site contains a wealth of information.

- **PDF Zone:** http://www.pdfzone.com/
  Especially useful for finding plug-ins that are PDF specific. Also a good source for answering some up-to-date questions regarding the latest releases of Adobe Acrobat/Acrobat related software. A related mailing list, with discussion of pertaining to the creation of .PDF archives such as ours, is available by filling out the form located at http://www.pdfzone.com/cgi-bin/wilma.cgi/pdf. A searchable archive of past questions and answers is maintained.

- **Planet PDF:** http://www.planetpdf.com/
  Although this is a commercial, marketing-oriented site, there are many useful features here, such as archives of tips and white papers.
- **Scanning tips:** http://www.scantips.com/
  This site (by Wayne Fulton) provides fundamentals and other basic scanning information to help you get the most from your scanner.

== // ==