PDF File Migration to PDF/A: Technical Considerations

Frank L. Walker, Marie E. Gallagher, and George R. Thoma; Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, USA

Abstract

The PDF/A specification for long-term preservation of electronic documents became an international standard in 2005. This standard seeks to guarantee the long-term visual appearance of an electronic document. For collections to be archived as PDF files, it makes sense to select the PDF/A file format, because this particular type of PDF file makes it easier to migrate to future file formats. However, in the years before the PDF/A specification became a standard, many organizations began creating archives of collections in PDF, but in formats not necessarily compatible with PDF/A. Because of its value to preservation that PDF/A offers, there is an advantage to migrating collections to PDF/A. Commercial software tools are becoming available, both for creating PDF/A files and for evaluating their compatibility with the PDF/A standard. One such tool was used to study PDF files culled from the Internet as well as from an in-house collection to determine the chances of success for migrating an archived collection of PDF documents to PDF/A. This study explores the types of problems posed by such a migration, and determines the circumstances in which a migration would be successful.

Background

The Adobe Portable Document Format (PDF) has been in use for more than fifteen years and has been widely adopted for electronic document use and distribution. Users have installed more than half a billion copies of the freely available Acrobat Reader® on a wide variety of computing platforms, to view local computer-based PDF files, as well as an estimated 200 million Internet-based PDF files (approximately ten percent of all Web documents). Over the years, successive versions of the PDF file format have become exceedingly more complex as new features appear with each new release, such as: embedded multimedia, document annotation, password protection, encryption, forms, and 3D capabilities. The continuing growth of PDF capabilities have led to a file format that, while feature-rich, is undesirable for specific applications. As a result, various subsets of PDF either have been adopted or are under development for specific uses: PDF/X for the publishing industry, PDF/E for engineering document workflow, PDF/UA for handicapped accessibility, PDF/H for health records, and PDF/A for electronic document preservation. After three years of work by the Association for Information and Image Management (AIIM), the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES), and many government agencies and private organizations, the proposed PDF/A standard was approved by the International Standards Organization (ISO) in September 2005. This new standard is designated ISO 19005-1:2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1) [1,2].

The PDF/A-1 standard is a subset of the specifications for PDF version 1.4 [3]. It supports two conformance levels: Levels A and B. Both conformance levels preserve the long-term visual appearance of an electronic document. A PDF/A-1 file conforming to Level B (also referred to as PDF/A-1b) provides the minimal requirements for ensuring a document's long-term visual appearance. This is accomplished by embedding all fonts within the file, using a device-independent color system, and including XMP metadata for describing the document [4]. This standard also eliminates several features of PDF 1.4: LZW compression, encryption, external references, transparency, audio or visual multimedia, JavaScript executable code, and embedded files. As a further requirement for maintaining semantic and structural information in the document, files conforming to Level A (also referred to as PDF/A-1a) must be "tagged" and contain Unicode character maps. A tagged file contains logical structure information that specifies the natural reading order of its contents. This not only facilitates migration to future file formats, but improves accessibility by permitting a user to read the document's contents in proper sequence. The Unicode character map provides semantic information about the characters, and facilitates text searching and copying, particularly for Asian languages. Tagging is most easily accomplished when a document is first created, such as with Microsoft Word, which allows a user to specify the document structure through heading levels, paragraphs, and table titles. When the Word document is converted to PDF, this information is used to create tags in the PDF file. It is also possible for Adobe Acrobat Professional to be used to tag an existing PDF file, but this can be a labor-intensive task.

Questions arise when an organization considers preservation using PDF/A. The use of PDF/A facilitates preservation, but proactive steps are required to guarantee it. These steps include periodic file replication (before media decays), the widespread adoption of the PDF/A standard through the creation and use of software tools designed to create and render PDF/A files, and the migration of other file formats to and from PDF/A. Due to its simplified format, migration of PDF/A to future file formats when necessary will be easier. Preservation is also facilitated because a PDF/A file is completely self-contained: all resources necessary to enable a PDF/A reader to display or print the electronic document are contained in the file. In addition, the file contains the metadata describing the document. During the first fifteen years of its existence, PDF has been used not only as a format for electronic document exchange, but also for preservation. In some instances, considerable resources have been invested to create document collections in the PDF format. Institutions with a preservation objective may need to address the following: should existing PDF collections be converted to PDF/A? Is this possible? What problems may be encountered? As shown in Figure 1, PDF/A is a subset of Version 1.4 (published in November 2001), which is a subset of version 1.5 (August 2003), which in turn is a subset of version 1.6 (November 2004), and this is a subset of version 1.7 (October 2006). Because each new version offers more capabilities than the previous one, will it become more difficult for

a PDF archive created using the latest version of PDF to be convertible to PDF/A? A two-part study considers these questions by examining the types of problems encountered during such a migration, and determining the degree of migration success.

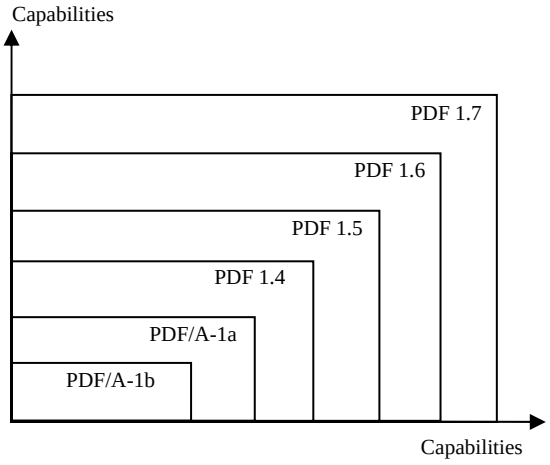

Figure 1. *PDF Version Capabilities*

Procedure

The study consists of two parts. The first part identifies the types of problems encountered in converting a general collection of Internet-based PDF files to PDF/A, and determines the potential degree of migration success. The second part of the study considers a specific example of an archived PDF collection at the United States National Library of Medicine (NLM) to determine whether it is a candidate for migration to PDF/A.

In part 1 of this study, we used samples of Internet-based PDF files from a wide variety of Web sites. The samples were taken from two time periods to see if the conversion to PDF/A at different times posed different problems. The first sample consists of 10,000 PDF files selected at random from several thousand Web sites between 2001 and 2003, with most of the files from 2002. We first located the files through Google searches. We then developed software to read the results of the Google searches, and automatically download these files. The second sample, assembled in September 2006, consists of 1,000 PDF files. For both samples, we compared all files to ensure that they are unique. We also performed quality checks on the files to ensure that they were downloaded properly without dropping bits. This PDF validation was accomplished by displaying each file using Acrobat Reader.

In 2006, new tools for creating and analyzing PDF/A files became commercially available. The tools generally fell into one of three functional categories: (1) for converting a non-PDF file to PDF/A; (2) for "preflighting" or analyzing a file for conformance with the PDF/A standard; and (3) for converting PDF files to PDF/A. At the time of this study the tools available could only manage conversion to Level B PDF/A files, not Level A files. One such tool that provided all three functions is PDF Appraiser, distributed by Apago, Inc. [5]. This was used to study the chances of success for converting the assembled PDF files to PDF/A. We used an evaluation version of this tool to analyze each of the 10,000 files in the first sample, and the 1,000 files in the second sample. PDF Appraiser determines whether a PDF file can be successfully converted to PDF/A. It lists all possible problems, grouping these into the ones that the tool can correct during conversion, and those that it cannot. Among the checks performed are an analysis of most objects in the file for syntax and consistency with the PDF/A standard, including the Info Dictionary, Catalog Dictionary, fonts, color spaces, ICC profiles, object streams, trailer dictionary, cross reference table, Unicode map, and XMP metadata. The results of each analysis was saved in an XML file, producing 10,000 XML file results for the first sample, and 1,000 for the second sample. We wrote software to read all the XML file analyses and produce an Excel-compatible spreadsheet that summarized the results, namely, the types of problems that may be encountered during conversion of a general collection of PDF files to PDF/A.

Part 2 of the study considers a specific collection of PDF files available on an NLM Web site, called Profiles in Science® (http://profiles.nlm.nih.gov/). Long-term preservation of this digital library has been a primary consideration from its inception. The purpose of Profiles in Science is to make available digital reproductions of historical items selected from the personal collections of prominent biomedical researchers and leaders in public health [6,7]. This Web site was launched in 1998 as a research project to expand access to these valuable collections and to promote the use of the Internet for research and teaching in the history of biomedical science. It features more than 20 collections containing published and unpublished items, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audiotapes, video clips, and other materials. In September 2006 there were 16,389 PDF files representing the paper-based portion of the collections. A few of these PDF files were created from color JPEG images, but over 16,370 PDF files were created using black and white images produced by scanning documents at 300 dots per inch resolution and storing them in a lossless compressed TIFF format. During the conversion from TIFF to PDF, the files were put through an OCR process to produce text-searchable PDF. The PDF files are available to the public through the Web site, and the original TIFF files have been archived off-line. We applied the PDF Appraiser analysis tool to approximately one percent of the PDF files in Profiles in Science. These randomly selected 172 files, representing samples from all collections, were analyzed and the results saved in separate XML files. Then our software read all 172 XML files, and produced a spreadsheet to summarize the analysis results. These results reveal the likelihood of successfully converting this specific collection to PDF/A.

Results – Part 1 of the Study

In part 1 of this study, we analyzed 10,000 PDF files randomly selected and downloaded from Internet Web sites over a three-year period (2001-2003), with the majority of the files downloaded in 2002. We found that the PDF Appraiser tool failed to process 274 of these files (2.74%): it either crashed or hanged, even after all files had passed a quality control check with Acrobat Reader. Of the remaining 9,726 files that the analysis tool successfully processed, we found that there were 332 unique producers of the files. A "producer" is basically a printer driver, such as Acrobat Distiller for Windows. Different versions of the same driver were counted as distinct producers. Our results showed that it would be possible to successfully convert 4,404

files to PDF/A, or 45.3% of the usable total. We spot-checked the reliability of conversion using the tool to convert a number of these files to PDF/A by using Adobe Acrobat Professional version 7 to "preflight" the resulting files. In every case the tool produced valid, displayable PDF/A files. The tool reported that the remaining 5,322 files (54.7% of the usable total) had problems that would prevent their conversion to PDF/A. Table 1 lists the ten most common producers and the percentage of PDF files they created that could not be converted to PDF/A. This table shows, as expected, that the most common PDF producers were various versions of Adobe Acrobat Distiller released during the period 1998 to 2002. There is no apparent trend in the conversion failure rates of these producers.

Table 1. Top Ten Producers and their Conversion Failure Rates: 2002 Sample

Producer	Percentage of all PDF Files in the Sample	Conversion Failure Rates: Percentage of Producer files that cannot be converted to PDF/A
Acrobat Distiller 4.05 (Windows)	9.9	47.4
Acrobat Distiller 4.0 (Windows)	9.6	62.0
Acrobat Distiller 4.0 for Macintosh	7.9	73.8
Acrobat Distiller 5.0 (Windows)	6.3	27.4
Acrobat Distiller 4.05 for Macintosh	5.5	69.5
Acrobat PDFWriter 3.02 (Windows)	5.0	48.8
Acrobat PDFWriter 4.0 (Windows)	3.6	50.9
Acrobat Distiller 3.01 (Windows)	3.5	70.5
Acrobat PDFWriter 4.05 for Windows NT	3.4	48.1
Acrobat PDFWriter 4.0 for Windows NT	3.3	43.9

We noticed similar results from the September 2006 sample of 1,000 Internet-based PDF files. In this sample, the analysis tool failed to process 50 files (5% of the total). Of the remaining 950 files, the tool found that 496 were convertible to PDF/A, or 52.2% of the usable files. This is nearly the same percentage as found in the earlier sample. The tool found a total of 148 producers in this sample. Table 2 lists the ten most common producers in this sample, and their conversion failure rates. The manufacturer released these producers during the period 2002 through 2005. It is interesting to note that the two oldest producers encountered, Acrobat Distiller 4.0 for Windows and Acrobat Distiller 4.05 for Windows) had significant increases in failure rates over the 2002 sample, but they were nearly the same failure rate as that of the second newest producer, Acrobat Distiller 7.0 for Windows.

Table 2. Top Ten Producers and their Conversion Failure Rates: 2006 Sample

Producer	Percentage of all PDF Files in the Sample	Conversion Failure Rates: Percentage of Producer files that cannot be converted to PDF/A
Acrobat Distiller 5.0.5 (Windows)	11.1	28.3
Acrobat Distiller 5.0 (Windows)	9.3	25.8
Acrobat Distiller 6.0 (Windows)	8.7	60.2
Acrobat Distiller 6.0.1 (Windows)	4.7	53.3
Acrobat Distiller 4.05 (Windows)	3.5	52.9
Acrobat Distiller 7.0.5 (Windows)	3.4	60.6
Acrobat Distiller 7.0 (Windows)	3.4	72.7
Acrobat PDFWriter 5.0 for Windows NT	3.3	50.0
Acrobat Distiller 4.0 (Windows)	3.0	79.3
Acrobat Distiller 4.05 for Macintosh	2.6	76.0

Table 3 shows the distribution of PDF file versions in the two samples, and their respective conversion failure rates. This reveals that, except for the small sample of files for PDF version 1.6, the conversion failure rate generally does not increase with newer versions of PDF. This indicates that the new features and capabilities offered by each new version of PDF do not appear to affect the ability to convert a file to PDF/A.

Table 3. Distribution of PDF Versions and Conversion Failure Rates

PDF Version	2002 Sample		2006 Sample	
	Number of files	Failure Rate	Number of files	Failure Rate
1	23	60.8%	1	0%
1.1	726	62.6	19	84.2
1.2	5409	54.1	222	55.8
1.3	3006	54.0	262	32.8
1.4	562	56.4	353	51.2
1.5	0	0	73	53.4
1.6	0	0	20	70.0

Table 4 lists the ten most common non-correctable problems identified by the tool for the 2002 sample of PDF files that could not be converted to PDF/A. These are all of a serious nature, and make migration impossible. The Frequency of Occurrence is the percentage of files in the sample experiencing the problem.

Table 4. Top Ten Non-Correctable Problems Preventing PDF/A Conversion: 2002 Sample

Problem	Description	Frequency of Occurrence
Font	Not embedded	37.2%
No matching CharSet entry	Missing value	11.9
Security	Invalid value	6.9
No matching glyph for CharCode	Missing value	3.5
BaseFont	Missing value	3.2
BG	Wrong type for object	.9
UCR	Wrong type for object	.9
Appearance	Missing value	.4
Action	Invalid value	.1
Incorrect ColorSpace	Invalid value	.1

Several of the most common non-correctable problems were problems with fonts:

- Font not embedded. The top problem is failure to embed fonts. Either an entire font or a subset of a font must be embedded within the PDF file. This type of error indicates it is not possible for the tool to embed the font in the file. The tool may not be able to embed the font either due to licensing restrictions, or perhaps it cannot find an appropriate font to embed.
- No matching CharSet entry. It is permissible to embed only a subset of a Type 1 font as long as all characters that are to be displayed are specified in the subset. This error indicates that an entry is missing in the CharSet element of the Font Descriptor.
- BaseFont missing value. This error is encountered if the PostScript name of the font is missing.
- No matching glyph for CharCode. A character code could not be matched to a glyph.

Another non-correctable problem was with security, or encryption. This occurred in 6.9 percent of all files in the sample. This indicates that the creator placed restrictions on file viewing, copying, modifying, or printing.

Table 5 lists the ten most common non-correctable problems in the 2006 file sample. The relative frequency of occurrence is nearly the same as that of the earlier sample.

Table 5. Top Ten Non-Correctable Problems Preventing PDF/A Conversion: 2006 Sample

Problem	Description	Frequency of Occurrence
Font	Not embedded	27.3%
No matching CharSet entry	Missing value	13.6
Security	Invalid value	6.4
No matching glyph for CharCode	Missing value	1.6
BaseFont	Missing value	1.5
Incorrect ColorSpace	Invalid Value	1.1
BM	Invalid value	.9
Count	Missing value	.5
Appearance	Missing value	.5
BG	Wrong type for object	.3

Table 6 gives the ten most common correctable problems that the tool encountered in the 2002 sample. These minor problems can be fixed during migration to PDF/A. Table 7 lists the same results for the 2006 sample. Among the most common correctable problems:

- Font not embedded. Although this is also listed as a non-correctable problem in Tables 4 and 5, it is correctable if the conversion software can embed the missing font in the PDF file, which would be most likely for the fourteen Postscript Type 1 fonts.
- DestOutputProfile missing value. This is an object that describes the ICC profile for device independent output color.
- XMP Metadata missing value. The metadata object is missing, and is required for the PDF/A specification.
- Colorspace Issues. There are a number of colorspace problems that the tool can correct while creating the PDF/A file.
- LZWDecode. The LZW compression algorithm is not permitted in PDF/A files, but images that are LZW-compressed are usually convertible to Zip or Group 4 compression.
- PDF/A tag not located. This indicates the file contained an XMP metadata object without elements for the PDF/A identification. This problem is easily fixed by transferring information from the Info object (e.g., producer, creation date, and subject).

Table 6. Top Ten Most Common Correctable Problems: 2002 Sample

Problem	Description	Frequency of Occurrence
DestOutputProfile	Missing value	93.0%
XMP Metadata	Missing value	81.3
Font	Not embedded	63.0
Colorspace Issues	Invalid value	58.5
TR	Forbidden object	28.7
LZWDecode	Encoded with invalid filter	23.2
Invalid Colorspace	Undefined	12.4
PDF/A tag not located	Missing value	11.6
Flags	Missing value	9.1
ID	Missing value	3.8

Table 7. Top Ten Most Common Correctable Problems: 2006 Sample

Problem	Description	Frequency of Occurrence
DestOutputProfile	Missing value	93.4%
PDF/A tag not located	Missing value	61.4
Font	Not embedded	48.4
XMP Metadata	Missing value	31.8
TR2	Forbidden object	23.2
Invalid Colorspace	Undefined	22.0
CIDSet	Missing value	19.4
CIDToGIDMap	Missing value	16.5
TR	Forbidden object	15.8
Colorspace Issues	Invalid value	15.7

We examined the two samples to determine the potential for conversion to PDF/A files with Level A compliance. While no tools were commercially available that specifically produced Level A files at the time of this study, we could estimate the potential for success of producing Level A files using our samples. In order to accomplish this, a PDF file would have to be convertible to Level B, but also be tagged, and have ToUnicode maps for its embedded fonts. In the 2002 sample there were 386 tagged files out of the 9,726 files that could be processed (3.9%). Of these, there were only 36 files containing embedded fonts with ToUnicode maps, with correctable problems. These are candidates for PDF/A Level A compliance. Unfortunately, this is only 0.37% of all files that the tool could process. This indicates that in a general collection of PDF files there is only a small percentage that could be migrated to a Level A compliant file (PDF/A-1a). It is interesting to note that in the more recent 1,000 file sample, there were 77 tagged files of the 950 files the tool processed (8.2%). Of these, there were only 15 correctable files containing embedded fonts with ToUnicode maps, making them candidates for Level A compliance (1.5% of the sample).

Because it is unlikely that an automatic process would be able to tag a PDF file accurately, then unless the file is already tagged, the likelihood of converting a non-tagged PDF file to PDF/A-1a is very small. In general we can conclude that if a PDF file can be converted to PDF/A, it is much more likely that it would be Level B compliant rather than Level A. This would be sufficient to maintain the long-term visual appearance of the document, but not enough to make it accessible to the handicapped or text searchable for some types of fonts.

One interesting aspect of this part of the study is that we found one type of PDF file to be almost always convertible to PDF/A with a high degree of success. This is an image-only or text-behind-image PDF file. In the 2006 sample, 31 files fell into this category. All other files in the sample population contained some form of visible text or text combined with images. Of these 31 files, 26 were convertible to PDF/A. There were 3 files that could not be converted due to encryption; had they not been encrypted they could have been converted. One file could not be converted due to a damaged color space, and another could not be converted because of a missing BaseFont. If we counted the three files that could not be converted due to a security lock placed on the files, then 29 of 31 files were convertible to PDF/A (94%). This leads into part 2 of the study, in which all archived PDF files fell into this category.

Results – Part 2 of the Study

Here we used the same procedure as in part 1 to analyze 172 PDF files (approximately a 1% sample) from the Profiles in Science collection at the National Library of Medicine. While this sample fell into the category of text-searchable image (text-behind-image) files, a few files had no searchable text because the original material was handwritten. The sample was indicative of the entire population of PDF files in the collection, as all fell into this category of PDF file. The analysis tool, which successfully processed all files in the sample, revealed that 100% of the sample population was convertible to Level B-compliant PDF/A. Three producers were used to create these samples: Acrobat PDFWriter 3.03 for Windows NT (89.5% of samples), Adobe PDFWriter 2.01 for Windows (8.7%), and Adobe PDF Library 4.0 (1.7%). The files were in PDF versions of 1.1, 1.2, or 1.3. All problems were correctable, with the most common ones being the following: missing XMP Metadata, missing value for the DestOutputProfile, and invalid compression (LZW). Since all PDF files in the Profiles in Science collection were created in the same manner, we can conclude that there is a high probability that all files in the collection are convertible to PDF/A-1b.

Conclusion

Organizations that have already archived files in the PDF format may consider migration to PDF/A, a new standard for long-term preservation of electronic documents. To determine whether a PDF collection is convertible to PDF/A, one of the emerging commercial tools may be used to analyze the collection. This study used one such tool (PDF Appraiser) to confirm that image-only PDF collections may be readily migrated to PDF/A Level B. Our investigation of Internet-based PDF files reveal that only about half of the PDF files available through Web sites can be converted to PDF/A-1b files, and that less than one percent is convertible to the more stringent PDF/A-1a. PDF files that are text-only or combine visible text with image pose a challenge to conversion tools. We found that new capabilities offered by recent versions of PDF do not appear to restrict the ability to convert a PDF file to PDF/A. Instead, most of the problems preventing

migration deal with incorrectly specified fonts, non-embedded fonts, encryption, and invalid color spaces. In order to achieve successful migration to PDF/A-1b, all non-standard fonts must be embedded in the file, all fonts and color spaces must be well defined, and no restrictions be placed on file use as governed through security settings.

Acknowledgement

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

[1] ISO 19005-1, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), available at http://www.iso.org/iso/en/ISOOnline.frontpage.

[2] PDF-Tools.com White Paper: PDF/A – The Basics. Version 1.0 February 1, 2006. Available at: http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf.

[3] Adobe Systems Incorporated, PDF Reference: Adobe Portable Document Format, Version 1.4, Addison-Wesley, Boston, 3rd edition (2001).

[4] Adobe Systems Incorporated, XMP Specification (2004).

[5] Apago, Inc. Web site: www.apagoinc.com .

[6] McCray, Alexa T., Marie E. Gallagher. "Principles for Digital Library Development." Communications of the ACM 44, no. 5 (May 2001): 48-54.

[7] Gallagher, Marie E., Christie Moffatt. "Surviving Change: The First Step toward Sustaining Your Digital Library." In: J. Trant and D. Bearman (eds.). Museums and the Web 2006: Proceedings, Toronto: Archives & Museum Informatics, published March 1, 2006 at http://www.archimuse.com/mw2006/papers/gallagher/gallagher.html

Author Biography

Frank L. Walker received his B.S. and M.S. degrees in electrical engineering from the University of Maryland. Since he joined the National Library of Medicine in 1979, he has designed, developed, performed research, and published a number of papers on computer systems utilizing electronic imaging, primarily for the purpose of electronic document storage, retrieval, transmission, and use. His current interest is in developing software tools for improving the communication and use of biomedical library information.

Marie E. Gallagher, a computer scientist in the National Library of Medicine's Lister Hill National Center for Biomedical Communications since 1990, is the project leader of the Digital Library Research and Development team. The team investigates systems and develops the software underlying Profiles in Science. Ms. Gallagher earned her B.S. degree in Computer Science and Mathematics from the College of William and Mary in Virginia.

George R. Thoma is a Branch Chief at an R&D division of the U.S. National Library of Medicine. He directs R&D programs in document image analysis, biomedical image processing, animated virtual books, and related areas. He earned a B.S. from Swarthmore College, and the M.S. and Ph.D. from the University of Pennsylvania, all in electrical engineering. Dr. Thoma is a Fellow of the SPIE, the International Society for Optical Engineering.