Tutorials

Publishing legacy documents on the Web

George Grätzer

Abstract

A great deal has been written recently about publishing LaTeX documents on the Web. But what happens if you are not lucky enough to have your document in LaTeX?

I am going to describe my adventures putting old documents on the Web. There are a few pitfalls on the way. If you follow this how-to guide, you can get your legacy articles on the Web in no time at all.

1 Introduction

I have made an effort in the past few years to make all my mathematical research articles available on the Web. If you check my Web site: `http://www.math.umanitoba.ca/homepages/gratzer/` you find all my articles, 165–202, in PDF format. The turning point was 1994, when David Carlisle's graphics package came into use. After 1994, I wrote all my articles in standard LaTeX and I included the diagrams — saved in EPS format — with the

```
\includegraphics
```

command of the graphics package. Now, seven or so years later, LaTeX is the same, EPS is the same, all the articles 165–202 typeset today as they did when they were written.

If you want to read about how to publish such documents on the Web, read Chapter 14 of my book [2]; if you want to read a whole book on the topic, read Michel Goossens and Sebastian Rahtz (with Eitan Gurari, Ross Moore, and Robert Sutor) [1]. The best book on the technical aspects of PDF is Thomas Merz [3].

I would like to thank R. Padmanabhan, Jacob Palme, and Thomas Merz for reading the manuscript and giving good advice.

2 Legacy documents

All my articles written BC (Before Carlisle) are legacy articles — with a few exceptions.

Here are the major categories of legacy documents:

1. Old documents written on a typewriter and then typeset in a printing shop. All my articles written before 1990 (1–132) fall into this category.

2. Documents written in an old word processor that is no longer available or in an old version of a word processor that went through too many changes. For instance, my papers that were written in Word 6.0/1995 (on the Mac) would need significant editing: many symbols appear wrong (was there a change in the encoding vector of the Symbol font?) and they are full of mysterious messages:

   Error! Bookmark not defined.

   You find lots of examples of this in the group 1990–99.
3. LaTeX papers whose source code has been misplaced; LaTeX papers that utilize packages or document classes that are no longer available or that are not compatible with today's LaTeX.
4. LaTeX papers with diagrams drawn with drawing programs that are no longer around. For instance, article 161.
5. LaTeX 2.09 and $\mathcal{A}\mathcal{M}\mathcal{S}$-LaTeX papers. Although Appendix G of [2] gives detailed and pretty straightforward instructions on how to convert such articles to LaTeX, carrying out such a conversion may be too large a task.

3 The quick and dirty solution

It became clear to me that a number of people going to my Web site are looking for older articles. So I decided that I will put all my articles on the Web.

This sounded really easy to do:

1. I scan the article.
2. I use Adobe Acrobat to turn the TIFF files created by the scanner to a PDF file.

It was indeed easy to do, but the result was terrible. Firstly, I came to realize that the edges of an old reprint were not necessarily cut parallel with the printed lines. So the scanned images were crooked. I also did not know how to optimally set my scanner. And when I set it to 300 dpi, black and white bitmap, the printed version came out 50% magnified, and as a result, really ugly (similar to a 200 dpi scan). Why is that? Go to the print dialog box of a PostScript printer, and choose the Acrobat pane; by default, it has a checkmarked box, `Fit to Page`. Uncheckmark the box, and the PDF file will print properly.

So you must warn the users, to leave the `Fit to Page` box unchecked. A better solution is in the next section.

4 The proper solution

We now use the following procedure to obtain the PDF files for legacy documents.

1. Scan each page of the document at 300 dpi, black and white bitmap.

Justification. The first decision to be made is at what dpi to scan. Since most printers these days are 1200 dpi printers, there is a temptation to scan at 1200 dpi.

Table 1 shows the size of a typical (large) printed page.

A 12 page paper at 300 dpi is about 1MB; it is more than 3MB at 600 dpi. At 1200 dpi, it is almost 6MB, obviously too large for most people to download.

Another way of increasing the file size is by using grayscale, instead of black and white bitmap, for the scanning. This increases the file size dramatically, and often decreases the quality of the PDF document. Avoid using grayscale unless there are grayscale illustrations in the document.

2. Open each page in Photoshop (or another similar application) and perform the following steps (stated specific to Photoshop):

Step 1. Change the page to grayscale
`Image>Mode>Grayscale`
keeping the `Image Ratio` at the default value 1.

Step 2. Enlarge the page and set the grid visible.
`View>Show>Grid`

Step 3. Use the eraser to get rid of dirt on the page.

Step 4. If necessary, rotate the picture in increments of 0.1 degrees to make the lines straight.
`Image>Rotate Canvas>Arbitrary...`

Step 5. When the lines are straight, change the page to black and white bitmap
`Image>Mode>Bitmap...`
keeping the `Output` at 300 pixels/inch.

Step 6. Change the canvas size to 8.5 inches by 11 inches; place the original image in the center (the default).
`Image>Canvas Size...`

Step 7. Save the image with `Save As...`, keep the TIFF format, and checkmark `LZW Compression`.

Justification. Step 1 is necessary, otherwise Step 4 cannot be done. Step 6 is useful, because then it no longer matters whether the `Fit to Page` box is checked.

Comment 1. Photoshop allows the creation of "buttons" to facilitate this process. The buttons are colored and carry a descriptive name to indicate the action that is carried out by one click on the button. I have five buttons for this procedure:

1. `To grayscale` It is colored gray. It does Steps 1 and 2.

Table 1: Size of a one-page scanned file

Scanning at	150 dpi	200 dpi	300 dpi	600 dpi
Size of TIFF file	172K	300K	668K	2.5MB
Size of PDF file	36K	52K	88K	212K
Quality of printed document	poor	poor	O.K.	excellent

2. `Turn clockwise` It turns clockwise by 0.1 degrees.
3. `Turn counterclockwise` It turns counterclockwise by 0.1 degrees.
4. `To bitmap` It is colored white. It does Step 5.
5. `Large canvas` It does Step 6.

Such a set of buttons can be saved. Then the set can be loaded as necessary.

Comment 2. J. Palme [4] raises the question how can you make a PDF document "international", that is, printable in North America in standard letter format, and outside of North America in A4 format. The summary of his advice is:

> If you are using U.S. Letter paper format, ensure that both the left and right margins are at least 21 mm (0.8 in).

Note that Step 6 ensures this in most circumstances, certainly with the 150 or so of my legacy articles.

Comment 3. You may get better PDF files if you make Step 4 slightly more sophisticated. Scan the document at 600 dpi. Then straighten the pages as follows:

- Select the Measure Tool (it hides behind the Eyedropper Tool).
- Drag it to draw a straight line across the page.
- Choose the menu item:
  `Image>Rotate Canvas>Arbitrary...`
  and a dialogue box comes up, showing the rotation necessary to straighten the page. Carry out the action by clicking `OK`.

Theoretically, this should give a better quality PDF document. In actual practice, I cannot see the improvement. Step 4, as recommended, has the advantage of simplicity. My papers were converted by a service bureau (Sri RAM Technicraft, e-mail: `sukanya8@mb.sympatico.ca`). Since they use untrained persons for such work, they appreciated the simplicity of the process.

3. Now take a look at how the pages are numbered. The scanner assigned them names such as

```
File1.xyz
File2.xyz
```

and so on. This will create problems if you have more than nine pages. So rename them

```
File01.xyz
File02.xyz
```

4. In Acrobat, choose `File>Import>Image`, choose all the pages in the document, and click on `Done`. The PDF file will be ready in a few seconds.

5. Acrobat will balk if you wish to make a PDF document of more than 50 pages. In this case, make more than one PDF document, and merge them as follows: open the first document in Acrobat. Choose `Document>Insert Pages...`; in the open dialog box, select the second PDF file. In the dialog box that comes up, for Location select `After`, and for page select `Last`. Clicking on `OK` will merge the two documents. Proceed thus until all the documents are merged.

Figure 1 shows a few lines of a scanned page on the Web. It is reminiscent of math papers printed on old 300 dpi laser printers.

5 Covers

Unfortunately, old reprints had covers containing vital information, maybe even the author and the title. As a rule, the journal information was on the cover only. So it is necessary in many instances to include the front cover with the article.

The covers pose special problems: they are often colored (say, medium blue) and somewhat the worse for the wear. If you scan them as in Section 4, the scanned image may be full of black spots — the image seems to be damaged beyond repair.

Here is what you have to do.

Step 1. Scan the cover at 600 dpi, grayscale.

Comment. Even the 600 dpi bitmap may be full of dirt or even completely black!

Step 2. In Photoshop, choose
`Select>Color Range...`.
In the dialogue box, pull down the `Select` menu, and choose `Shadows`. This selection consists of all the printed letters and logos (and a few dirt spots).

Step 3. Fill the selected area with black. This will change the grayscale print to black.

Step 4. Invert the selection with
`Select>Inverse`,
and fill the selection with white. This will create

(4) *implies* (2). Let $(a]$ be a principal ideal which has two different factorizations $(a] = \wedge P_\alpha = \wedge Q_\beta$. We choose an element $b > a$. Obviously, b is not element of all P_α and Q_β, e. g. let $b \notin P_1$ and $b \notin Q_1$. Combining condition (4) with Theorem V, we get that in L every prime ideal is maximal, hence $(b] \cup P_1 = (b] \cup Q_1 = L$. Since in a distributive lattice the relative complement is unique, we conclude $(b] \cap P_1 \neq (b] \cap Q_1$, furthermore $(b] \cap P_\alpha$ is a

Figure 1: 300 dpi scan on the Web.

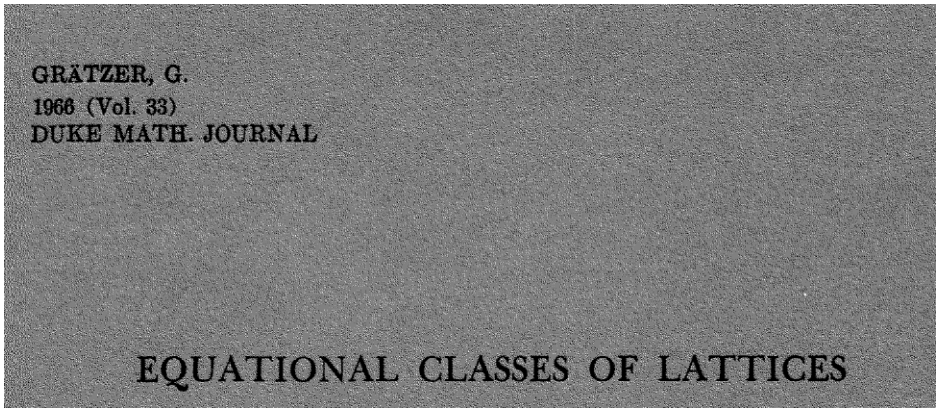

Figure 2: Cover scanned at 600 dpi.

a white background. Clean up the dirt, and if necessary, repeat Steps 3 and 4. Now you should have a beautiful black and white cover.

Figure 2 shows a portion of a cover scanned grayscale at 600 dpi. Figure 3 is the cleaned up version.

R. Padmanabhan suggested to cut from the cleaned up cover the information missing from the first page and paste it on the top of the first page. This saves storage and download time. However, the reader may be under the impression that the scanned pages are faithful representations of the original, which would not remain quite true with this scheme.

6 Future

Obviously, some years in the future, my Web site as set up today will appear obsolete. When most users will have very fast Web connections, when data storage will be measured in hundreds of gigabytes, all articles will be scanned at 2400 dpi.

More importantly, Adobe Acrobat introduced the capture command, which allows us to store a legacy document in two ways: as an image and as text (linked to the image). This has the advantage that we can view a legacy document as it looked originally, and the text layer allows complete searching capabilities. Unfortunately, I was completely unsuccessful in my attempts to use capture for my documents. Even PDF files converted from LaTeX showed a very high failure rate.

You can get a glimpse of the future at `http://www.jstor.org/`, the Web site of JSTOR, The Journal Storage, The Scholarly Journal Archive. The image files are created at 600 dpi and the text recognition is accomplished with proprietary software (and carefully proofread). The mathematics journals include the Proceedings of the American Mathematical Society and the Transactions of the American Mathematical Society.

The future will have all this and links: *internal links* in an article from "by Lemma 5" to Lemma 5, from "see [12]" to the citation [12]; and *externally*, from citation [12] to the actual article.

References

[1] Michel Goossens and Sebastian Rahtz (with Eitan Gurari, Ross Moore, and Robert Sutor), *The LaTeX Web Companion: Integrating TeX, HTML and XML*. Addison-Wesley, Reading, MA, 1999.

[2] George Grätzer, *Math into LaTeX*, third edition, Birkhäuser Verlag, Boston, Springer-Verlag, New York, 2000. xl+584 pp. ISBN 0-8176-4131-9, ISBN 3-7643-4131-9.

[3] Thomas Merz, *Web Publishing with Acrobat PDF*. Springer-Verlag New York, 1998.

GRÄTZER, G.
1966 (Vol. 33)
DUKE MATH. JOURNAL

EQUATIONAL CLASSES OF LATTICES

Figure 3: Cover cleaned up.

[4] Jacob Palme, *Making Postscript and PDF International*, Network Working Group, Request for Comments: 2346, Stockholm University, May 1998.
http://dsv.su.se/jpalme/ietf/jp-ietf-home.html\#anchor1470437,
http://www.ietf.cnri.reston.va.us/rfc/rfc2346.txt

◇ George Grätzer
Department of Mathematics
University of Manitoba
Winnipeg MN, R3T 2N2
Canada
gratzer@cc.umanitoba.ca
http://server.math.umanitoba.ca/homepages/gratzer/