Some Best Practices for Retrodigitization

Committee on Electronic Information and Communication

June 29, 2005

1 Preface

Mathematical research has been revolutionized by the availability of much literature on the internet. It is now commonplace to find papers published within the last decade as digital images that may be read or printed using an ordinary web browser. Older papers that were born in print are harder to obtain, but they may be made available by creating digital images from the printed pages. This process of retrodigitization can potentially make the entire printed history of mathematics available to anyone with an internet connection.

Having a scanned image of a paper available is helpful, but it is only a small first step in making the digitization useful. There must be some ability to recognize the text (and perhaps some of the mathematics) in the paper so that searches may be done efficiently.

In addition, there must be some method for the paper to make itself known on the internet. This implies that there must be some recognized methods of communicating with the internet search engines. It is important that different digital libraries do this with compatibility.

There is no doubt that digital libraries will emerge. Hopefully they will work together harmoniously and present a seamless entity to the user. This best-practices document presents some technical parameters that may be used to fulfill this aim. It is one is a series of similar documents produced by the Committee on Electronic Information and Communication (CEIC) of the International Mathematical Union (IMU) at the request of various individuals and groups involved in digital initiatives. The World Digitial Mathematics Library (WDML) is part of this initiative (http://www.wdml.org/).

2 Overview

Any retrodigitization project must proceed in several stages. First, the material must be scanned to produce a digital image. This image is sometimes called the *graphics plane*. It is this file that is normally presented to the viewer when the paper is requested using a web browser. The importance of a high-quality scan is hard to overstate. This is not only because it is what the user sees, but also because it is used to extract the text. Generally speaking there will be

only one opportunity to scan the material (especially if the scanning process is destructive to the material being scanned), so it must be done correctly the first time.

The quality of the graphics plane must be sufficient to proceed to the next stage: to allow an optical character reader (OCR) to recognize and extract the text from it. The extracted material is called the *text plane*. It is used, for example, to search for phrases within the document. Accurate recognition of widest range of text and mathematical characters is an obvious goal.

In addition, there must be a description of the identifying characteristics of the document being digitized: the names of the authors, the title, the year and place of publication, and so forth. These descriptions are collectively called the metadata. It is vital that the metadata be accurate, since it is used to identify the document to the internet.

In the next three sections we make specific suggestions regarding these three component steps.

3 Scanning pages

- **scope:** It is recommended that all pages including covers and advertisements be scanned.
- **resolution:** The scanned images must be capable of resolving small mathematical and text characters (subscripts of subscripts, for example). Five point type should be resolvable. As such, a minimum of 600dpi should be used. Higher resolutions may be used effectively (as for example with high-resolution graphics viewers).
- **colour:** Although most mathematical papers usually have little coloured material, it is important that any such material be reproduced accurately. For pages containing coloured matter, it is recommended that 24 bit colour be used.
- **vertical alignment:** If the scanned image is out of vertical alignment, the accuracy of the OCR will degrade substantially. It is recommended that all documents be no more than two degrees out of vertical alignment.
- **cropping:** Pages should not be cropped in the initial scanning since, for example, the margins occasionally contain interesting information.
- **compression:** Any compression should be lossless. A standard format (such as TIFF or lossless JPEG) should be used for storing the images.
- **archiving:** the un-retouched original scans should be archived. Future technological advances may be easy to apply to the original scanned images; rescanning will be impossible in many instances.
- **grouping of images:** images should be grouped into logical units (article, chapter, etc.) for efficient downloading.

4 Optical Character Recognition

It is recognized that the text plane may be primarily used locally and considerable flexibility may be allowed in the format.

- **character set:** mathematical papers contain many mathematical and foreign language characters. It is recommended the Unicode encoding of characters be used to allow non-Latin letters to be part of the text plane. If possible, recognizable mathematical characters should be included.
- **coordination of text and image planes:** it is recommended that the text be keyed to the image, that is, the coordinate positions of the characters be included in the text plane to allow accurate feedback to user searches.
- **unrecognizable material:** much mathematical material is unrecognizable by current OCR technology. It is recommended that this material be tagged so that it will be easy to revisit when future technologies make it advantageous to do so.

5 Metadata

- **format:** the metadata should be stored as an XML document with a publicly available DTD. Examples and further information may be found (http://www.numdam.org/OAI/minidml.xsd) at the NUMDAM project.
- **dissemination:** the metadata should be available easily to appropriate internet harvesters. OAI-PMH compliance is urged. For further information see (http://www.openarchives.org/OAI/2.0/guidelines-repository.htm)

6 Document delivery

Shorter documents should be delivered as a single file. Larger ones should be divided into logical subunits. Formats for document delivery files should be openly available. Hyperlinks should be included whenever possible. The use of Portable Document Format (PDF) or Deja Vu (DjVu) files (or both) is recommended.

7 Conclusion

These recommendations reflect the currently available technology (Summer, 2005). They are given with the recognition that some local variations may be necessary, and that there may be significant changes in technology in the near future. Indeed, even now there are several different methods that may be used for retrodigitization. Nonetheless, it is hoped that those undertaking

digitization projects will strive to meet these recommendations since the value of the project will only be realized fully if interoperability and integration is assured.