Accessibility in Digital Mathematical Libraries — Current Status and Future Prospects

Alan P. Sexton
School of Computer Science
University of Birmingham
www.cs.bham.ac.uk/~aps

Abstract

Digital Mathematical Libraries (DMLs) are finally becoming more than just a dream with numerous prototypes and early systems graduating to become mainstream tools for mathematicians, scientists and engineers. Accessibility in DMLs for the visually impaired is, however, still a research issue, with very little deployed support in real DMLs. Nonetheless, progress is being made in this research, and prototype support for accessibility is now being built and deployed in some DMLs.

1 Introduction

The digital revolution has brought great improvements of convenience to mathematicians. Papers can be written to high standards of formatting, distributed widely, found easily and downloaded from sites all around the world without the need to wait even for the postal system. Publishers have woken to the dangers to their business models of electronic publishing and developed fee charging web interfaces to their publications, and have industriously worked to monetise their back catalogs by adding them online. Libraries, often starved of funding, have struggled to stretch themselves beyond the physical book and journal repository model of old and can often do little more than provide portals to publishers websites. A few have made serious attempts to curate digital collections, most notably of documents and newspapers that are rare, historically valuable, or of significant importance with respect to cultural heritage. Mostly this has been with a view to make the documents more easily and cheaply available to a wider community of users without exposing the original physical items to further stress that may damage them. Building digital collections of mathematical documents has been a niche task within this specialist field.

The concept of a digital mathematical library (DML) as providing significant added value beyond that of the physical collection is not very old. While the ArXiv[1] was founded in 1991, and JSTOR[2] in 1995, possibly the first dedicated DML was Project Euclid[3], founded in 1999, followed by NUMDAM[4], founded just before 2000. It can be argued that it is really only since 2001 that previously speculative discussions about DMLs solidified into wider serious consideration [4].

If DMLs are a niche area in a specialist field, then accessibility for DMLs has been a poorly funded and underdeveloped sub-area thereof. Yet while practical accessibility options for plain text documents have existed for a long time, accessibility for mathematical documents has been limited to specialist research projects, e.g. T.V. Raman's ASTER system [6] and Fitzpatrick and Monaghan's TechRead [5], it is only recently with developments from Suzuki's Infty system [7], that a commercially available practical tool for making mathematical documents accessible to the visually impaired has appeared.

In this talk we review the state of the art of accessibility in Digital Mathematical Libraries, discuss some of the technical and practical reasons limiting accessibility in DMLs and suggest research and actions that we believe will help to improve this situation.

[1] http://arxiv.org/
[2] http://www.jstor.org/
[3] http://projecteuclid.org/
[4] http://www.numdam.org/

2 Digital Mathematical Library Contents

DMLs contain mathematical documents and metadata about the documents. Different DMLs provide searching and browsing services of varying degrees of sophistication. Making such services accessible is not particularly difficult: guidelines for web page and web application accessibility are well understood by anyone who cares to study them. The problem is the mathematical content of the documents themselves. Whatever specific formats the documents are stored in, within those formats the documents may be of one of three types:

1. Retro-digitised: Here the document has been scanned from a paper copy and the pages are stored as bitmap images. The images may be encoded within a format such as Postscript, PDF or DejaVu, but there is no information in the document as to what the characters are. In order to make such documents accessible, either an optical document analysis process must be applied, involving optical character and formula recognition using, for example, the Infty system, or the document must be re-keyed at some expense. The resulting document can be made accessible.

2. Retro-born digital: In this case the document was originally produced electronically from sources using a system such as TEX, Troff or Microsoft Word, and then formatted into a presentation format such as Postscript, PDF or DejaVu. However only the presentation format versions are now available but not the sources. Three options are possible to turn such documents into a form that can be made accessible: It can be rendered to a bitmap format and optical document analysis techniques can be used recognise it as for retro-digitised documents. Re-keying is also possible but still expensive, or the presentation version can be analysed to extract the precise character information as recorded in the format and analysing the contents using that information, for example by using [2]. This latter option avoids the difficult problems caused by the inherent inaccuracies of optical recognition of characters from bitmap images.

3. Born digital. Here the original sources for the document are available. In such cases, various transformations are usually possible to directly translate it into an accessible format, e.g. using Tralics[5], the TEX to XHTML plus MathML translator used by NUMDAM.

3 Mathematical Document Analysis

Scientific documents pose special challenges for optical recognition and the provision of accessibility support that has not yet been fully and successfully addressed. They contain equations, plots, charts, technical diagrams, data tables, photographs and other elements that standard OCR tools either do not handle well or do not handle at all. Even if such elements can be correctly recognised, there are limits on the technologies currently available to make them truely accessible to the sight impaired user.

There is much research currently in progress aimed at tackling these problems. However, the field of mathematical document analysis has still attracted only a tiny fraction of the research that has been devoted to plain text document analysis. Nonetheless impressive results have been demonstrated in research and, now, commercial systems — although still by relatively isolated groups. For the field to mature, more work integrating different results is necessary, more cross system evaluations should be carried out and more lessons should be learned and applied from such work. There is at least some work in progress on such evaluations [3].

One of the common issues faced by all researchers in the area is the lack of open, high quality data sets that can be used both as a challenge and a goal for the research and as a source for comparative

[5] `http://www-sop.inria.fr/apics/tralics/`

evaluations. Preparing such data sets is a complex and expensive task. Suzuki et al published a ground truth data collection of significant size and value for this purpose [8]. However, the documents that this ground truth database was based on are not freely available, which hampers work on comparative evaluations. More recently, we have produced a new, high quality scan of a document[6] that promises to serve as a useful addition to such datasets: Abramowitz and Stegun's Handbook of Mathematical Functions [1] contains a large range of mathematical expressions, many line plots and tables of scientific data. Moreover, as it was published by the United States National Bureau of Standards, it is copyright free and hence fully publically available for scanning, to be reported about and for results of document analysis to made freely available. Its history and the respect in which scientists have held the book make it an authoritative source for many types of expressions, diagrams and tables. The fact that it was printed in the pre-LaTeX days means that it can help researchers to avoid over-tuning their systems to the much more readily available TEX/LATEX sourced documents. Figure 1 shows some example elements scanned from this book.

4 Accessibility for Mathematical Documents

There are many commercial and free tools to assist in accessibility, from screen-readers, to text-to-speech engines, to refreshable braille displays or embossers. However there is little support for mathematics in such tools, and to date, none that we have found in free software systems. A number of approaches have been tried, including the following:

1. Obtain or produce the LATEX source for the document in question and allow users to read the raw LATEX code through braille or screen readers. There are variants of this option such as translating the LATEX code directly into braille.

2. Generate XHTML plus MathML from the document and depend on users having suitable tools to handle such formats (commercial tools are available such as Design Science's MathPlayer[7]).

3. Generate Daisy XML format for the document and use a (commercial) Daisy reader. Daisy (the Digital Accessible Information System[8]) is a consortium that is developing standards for digital accessiblity. They have adopted MathML for mathematics and SVG for diagrams. While Daisy XML is likely to become the accepted standard for accessibility in the future, tools for handling mathematical documents in Daisy XML format are still in an early stage of development.

4. Generate plain text from the document that corresponds to how a human reader would read out loud the mathematical expressions and use a text-to-speech system or screen reader to read the generated text.

A long term solution is likely to be the use of Daisy XML. But this solution is still immature. All the other solutions tend to have major problems, incompatibilities or strong limitations. Therefore it is not surprising that DMLs have not yet invested in accessibility in a serious way other than to depend on XHTML and MathML. A prime example of this approach is NUMDAM.

5 EuDML

The newest DML project is a European one: EuDML. This is a € 1.6 Million project to integrate many different European digital mathematical libraries. "European" is interpreted loosely; the Russian DML

[6] http://www.cs.bham.ac.uk/~aps/research/projects/as/
[7] http://www.dessci.com/en/products/mathplayer/
[8] http://www.daisy.org/

$$f(s)=\mathcal{L}\{F(t)\}=\int_0^\infty e^{-st}F(t)dt$$

(a) Equation 1

$$p(n)=\frac{1}{\pi\sqrt{2}}\sum_{k=1}^{\infty}\sqrt{k}A_k(n)\,\frac{d}{dn}\,\frac{\sinh\left\{\frac{\pi}{k}\sqrt{\frac{2}{3}}\sqrt{n-\frac{1}{24}}\right\}}{\sqrt{n-\frac{1}{24}}}$$

(b) Equation 2

$\operatorname{erf} z_n=0 \qquad z_n=x_n+iy_n$

n	x_n	y_n	n	x_n	y_n
1	1.45061 616	1.88094 300	6	4.15899 840	4.43557 144
2	2.24465 928	2.61657 514	7	4.51631 940	4.78044 764
3	2.83974 105	3.17562 810	8	4.84797 031	5.10158 804
4	3.33546 074	3.64617 438	9	5.15876 791	5.40333 264
5	3.76900 557	4.06069 723	10	5.45219 220	5.68883 744

$$\operatorname{erf} z_n=\operatorname{erf}(-z_n)=\operatorname{erf}\bar{z}_n=\operatorname{erf}(-\bar{z}_n)=0$$

(c) Table

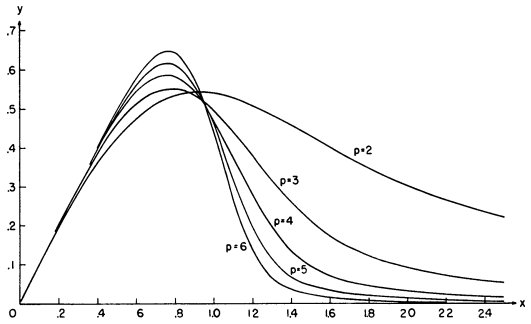

FIGURE 7.2. $y=e^{-x^p}\int_0^x e^{t^p}dt$.

$p=2(1)6$

(d) Line Plot 1

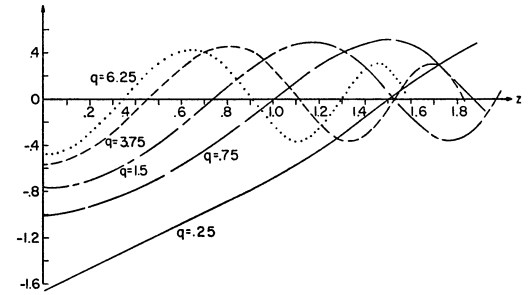

FIGURE 20.13. *Radial Mathieu Function of the Second Kind.*

(e) Line Plot 2

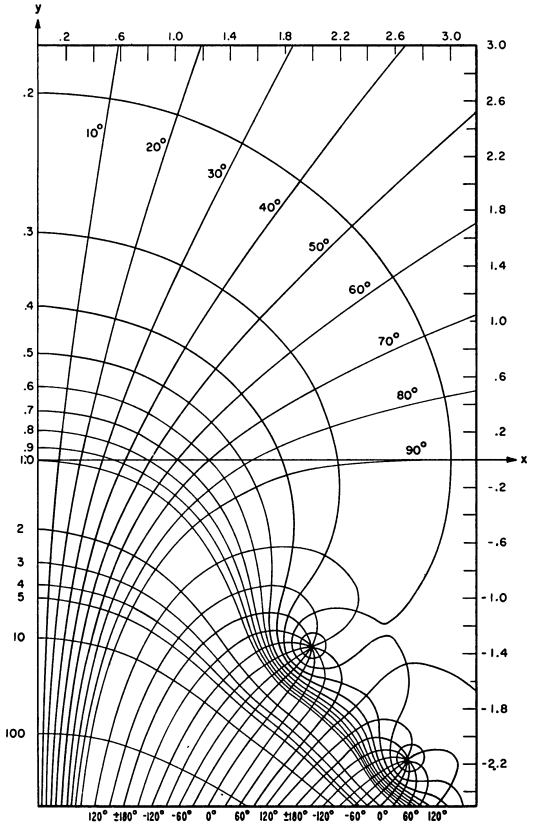

FIGURE 7.3. *Altitude Chart of $w(z)$.*

(f) Chart

Figure 1: Examples from Abramowitz and Stegun

RusDML is included. An interesting aspect of the project is the commitment to provide support for accessibility. Currently it uses Suzuki's Infty to handle Retro digital documents, Tralics to handle born digital documents and Baker, Sexton and Sorge's Maxtract to handle Retro-born digital documents. In the latter case, three different formats are generated for formulae: english spoken text, LaTeX and MathML.

An interesting issue, which is more political than technical, has recently become obvious. This is that most publishers secure their documents by, for example, encoding them in a PDF format with security features enabled. This has the effect of disallowing some kinds of analyses of such documents. This

will remain an issue for users who have visual impairments unless some special access rights can be negotiated with publishers.

6 Conclusions

Dedicated Digital Mathematical Libraries are less than 12 years old. It is not surprising that accessibility issues for such libraries, considering the complex problems involved, have not yet been adequately resolved. A great deal of work is necessary before the visually impaired will have effective access to DMLs. Some of this work is political in nature. Much of it requires building a community of researchers and developers that extend beyond their local groups and collaborate to share their knowledge and experience in a more integrated way than they have before.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. National Bureau of Standards, Washington D.C., 1972. 10th printing, December 1972, with corrections.

[2] Josef Baker, Alan Sexton, and Volker Sorge. Faithful mathematical formula recognition from pdf documents. In *9th IAPR International Workshop on Document Analysis Systems, Extended Abstracts*, pages 485–492, Boston, USA, 2010. ACM Press.

[3] Josef B. Baker, Alan P. Sexton, Volker Sorge, and Masakazu Suzuki. Comparing approaches to mathematical document analysis from PDF. In *Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, pages 463–467, September 2011.

[4] John Ewing. Twenty centuries of mathematics: Digitizing and disseminating the past mathematical literature. *Notices of the AMS*, 49(7):771–777, 2002.

[5] Donal Fitzpatrick and Alex Monaghan. TechRead: A system for deriving braille and spoken output from LaTeX documents. In *Proc. of ICCHP98*, pages 316–323, 1998.

[6] T. V. Raman. ASTER — Towards Modality-Independent Electronic Documents. *Multimedia Tools Appl.*, 6(2):141–151, 1998.

[7] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. Kanahori. Infty — an integrated OCR system for mathematical documents. In *Proceedings of ACM Symposium on Document Engineering*, pages 95–104. ACM Press, 2003.

[8] M. Suzuki, S. Uchida, and A. Nomura. A ground-truthed mathematical character and symbol image database. In *Proc. of ICDAR*, pages 675–679. IEEE Computer Society, 2005.