Accessibility Issues in Digital Mathematical Libraries

Petr Sojka, Michal Růžička, Maroš Kucbel, Martin Jarmar

Faculty of Informatics, Masaryk University, Brno, Czech Republic
✉ sojka@fi.muni.cz, {mruzicka, kocka, 172981}@mail.muni.cz

Keywords: accessibility of mathematics, canonicalization of mathematics, searchability of mathematics, digital mathematical libraries, MathML, Czech Braille

Abstract

The growing number of digital libraries does not serve only metadata of scientific or educational documents, but also the full-text data themselves. This brings new possibilities but also accessibility issues to cope with when designing user interfaces for exploratory search and accessing and reading the full-texts that are usually provided only in some version of PDF format.

We have participated in the design and solutions for the European Digital Mathematics Library (EuDML) and also participated in the preparation of primary data of the Czech Digital Mathematics Library (DML-CZ). In the paper we describe the developed technologies addressing Braille output of document content including mathematical formulae, document preprocessing and enhancement that increase accessibility, readability and exploration qualities (similarity of mathematical documents) of the documents in digital libraries of texts for tertiary education and research in STEM domain.

1 Introduction

The amount of digitally available and processable scientific texts and data grow at the unprecedented pace. Even small part of the scientific output that is peer reviewed and published in a controlled way (journals, proceedings, books) to be used for research and (tertiary) education is big and steadily growing. The chance that at least negligible amounts of this valuable scientific output are processed manually to be accessible for people with disabilitites is very low. Ensuring accessibility of wide range of scientific document types produced by wide range of the ICT-Information and Communications Technologies (EIT) is limited to small isles of literature used for educational purposes enforced by the Law enacting Section 508 in the USA. This does not cover research literature, blogs or Linked data, a must in today's research communities at universities. The remedy might be that publishing work-flows are adapted so that information is processed in as rich and as semantic forms as possible. Specifically challenging in this respect is the Science, Technology, Engineering, and Mathematics (STEM) domain full of mathematical formulae, diagrams and other two-dimensional structures appearing in STEM literature.

Validated publications and data are stored in the digital libraries. The growing number of digital libraries does not serve only metadata of scientific or educational documents but also the full-text data themselves. This brings new possibilities but also accessibility issues to cope with when designing user interfaces for exploratory search and for accessing and reading the full-texts that are usually provided only in some version of PDF format.

1.1 Accessibility in Digital Mathematical Libraries

We have participated in the design and solutions for the European Digital Mathematics Library (EuDML, https://eudml.org/) [18] and also participated in the preparation of primary data of the Czech Digital Mathematics Library (DML-CZ, http://dml.cz/) [4]. In both projects we have addresses several accessibility issues and designed the workflow so that available literature is as accessible, searchable and explorable as possible. In this paper we describe the developed technologies addressing Braille output of document content including mathematical formulae, document preprocessing and enhancement that increase accessibility, readability and exploration qualities (similarity of mathematical documents) of the documents in digital libraries of texts for tertiary education and research in STEM domain.

In the next section we briefly introduce mathematics processing in EuDML and necessity of canonicalization of math formulae. Section 3 refers about accessible formats and developed tools used in the EuDML workflow. Paper closes by conclusion and summary Section 4.

2 Mathematics Processing in EuDML

The EuDML is a Digital Mathematics Library that collects published scientific literature in a 'one stop shop' for math students and researchers. Metadata of related published items are collected via OAI-PMH protocol and most items are ingested including their full-texts in various formats as agreed with project data providers. There are papers that have more formulae than plain text, even in the metadata (titles, abstracts), and even leading edge NLP and machine learning tools are not designed to cope with this gracefully. Full texts are processed internally and enhanced [15] so that added value information could be computed by carefully designed automated workflow and presented to the users of EuDML portal.

2.1 Workflow

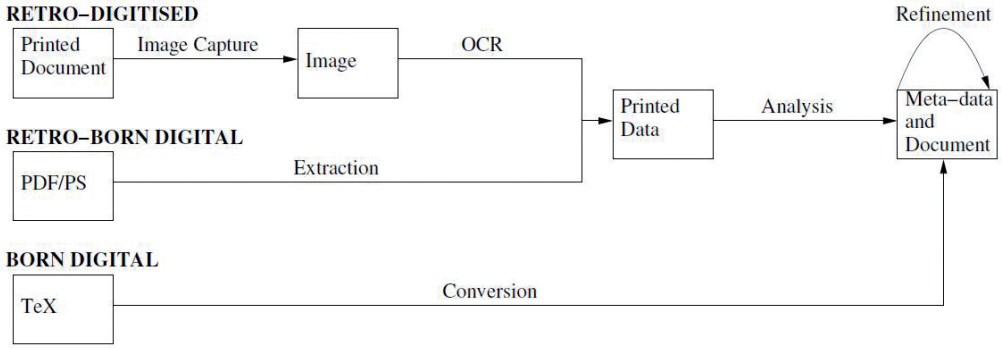
Top-level workflow is depicted on Figure 1. Ingested items are of various origin – some were digitized and available as scanned bitmaps and need Optical Character Recognition (OCR), some were created digitally by different publishing tools and provided in various forms of PDF, PostScript or even [X]HTML and need extraction. Some are in the primary (LaTeX) format and need conversion. Workflow tries to create all available data in one homogeneous format. Namely plain text with math formulae in W3C standard MathML: *plain math-text*.

To get plain math-text for further enhancements two frameworks are used:

- **Maxtract** for conversion of born-digital PDF to plain math-text. [3]
- **InftyReader** for OCR of other documents (usually rendered to series of page TIFFs or bitmapped PDF) to plain math-text. [17]

(For further information see Section 3 on page 92.)

Plain math-texts are generated from rich internal document representation that both tools create by sophisticated layout analysis algorithms. As this primary information is visual it is really hard to convert and disambiguate it into useful semantic markup (Content MathML [2]).



[Fig. 1] EuDML enhancement workflow

2.2 Canonicalization

Widely used language for encoding of mathematical content in digital mathematics libraries is MathML. [2] MathML encoding exists in two forms – Presentation and Content. In contrast to Content MathML encoding which grabs the meaning of formulae, Presentation MathML encodes appearance of the formulae. Thus, there is a lot of different possible encodings of mathematical statements with the same meaning using Presentation MathML. As Presentation MathML is widely used in the real world documents we have to cope with, this encoding has been primarily supported in our math-aware search engine [14]. The ambiguities that Presentation MathML bring, however, are issues not only for mathaware search engines but also for accessibility of mathematical documents. The same looking formulae may describe different mathematical notions, different content. For our processing, we have to pick up some canonical representation of formulae which should be as disambiguated as possible.

Our first attempt was use of UMCL (Universal Maths Conversion Library)[1]. [1] The main purpose of the UMCL toolset is the transcription of the MathML formulae to Braille national codes. However, part of the translation process is also canonicalization of input MathML intended to eliminate ambiguity of inputs and thus making it easier to translate it to the Braille national codes. The canonicalization module was implemented as a set of XSL transformations. These XSLT stylesheets were extracted from the UMCL toolset and we tried to use them for canonicalization of MathML inputs of our math-aware search engine.

Unfortunately, the use of the UMCL canonicalization module appeared to have severe deficiencies. Firstly, use of XSLT was not fast enough for processing of large amounts of data that are necessary for routine operation of math-aware search engine in digital mathematics library as large as EuDML. Secondly, and more importantly, the implementation of the UMCL module proved to change semantics of input formula during the transformation (see Listing 1).

1 http://inova.ufr-info-p6.jussieu.fr/maths/umcl

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
   <mfenced>
      <mrow>
         <mi> a </mi>
         <mo> + </mo>
         <mi> b </mi>
      </mrow>
   </mfenced>
</math>

<math xmlns="http://www.w3.org/1998/Math/MathML" id="formula0:1">
   <mrow id="formula0:2">
      <mrow id="formula0:3">
         <mi id="formula0:4">a</mi>
         <mo id="formula0:5">+</mo>
         <mi id="formula0:6">b</mi>
      </mrow>
      <mrow id="formula0:7">
         <mo id="formula0:8">+</mo>
         <mi id="formula0:9">b</mi>
      </mrow>
      <mi id="formula0:10">b</mi>
   </mrow>
</math>
```

[Listing 1] UMCL inputs MathML encoded formula ($a + b$) but outputs MathML encoded formula $a + b + bb$.

Taking into the account insufficient speed of the XSLT and complexity of the original solution we decide to abandon this solution and implement our own MathML normalization tool. [6] Canonicalization is going to deal with both Presentation and Content MathML. Almost all visual markup is deleted as it is not bearing semantic information. Also, different notations are unified into a canonical one. As outlined in [13] future version of the canonicalizer should take various *semantic annotations* in the context of formulae to disambiguate terms in a formulae[2] or even convert Presentation MathML to a Content one.

3 Accessibility Formats in EuDML

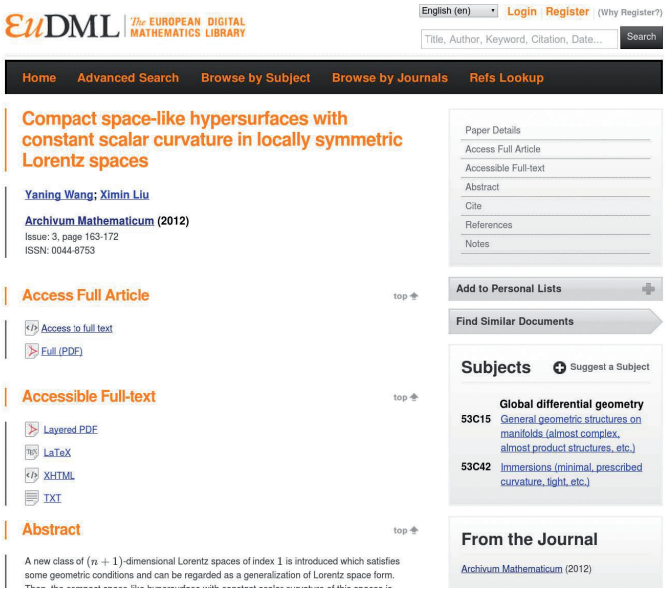
When evaluating accessibility options for EuDML, in addition to the plain text with math read as text, several math accessibility formats were identified [16, Section 3.1]: Braille with Math, DAISY XML, canonical MathML by UMCL and HRTeX (Human Readable TEX) by InftyReader. We have opted for supporting as many formats as possible given the process to create them is *fully automated* during internal enhancement processes [15].

2 Is f variable, function or functional?

Maxtract[3] is capable of preparation of different output formats:

- LaTeX for use with Tralics (a LaTeX to XHTML+MathML converter[4]) [8, 5].
- LaTeX for layered PDF with LaTeX and text layers.
- LaTeX for annotated PDF with LaTeX annotations.
- A simple text file.
- A text file with math in LaTeX.

These results are consequently available for internal purposes and for transformation to accessible formats (Layered PDF, LaTeX, XHTML, or TXT) for the EuDML users (see Figure 2).



[Fig. 2] Available accessible formats in EuDML item https://eudml.org/doc/247009

In the layered PDF the reader can switch between several copy & paste textual representations stored as separate plates in the PDF documents.

For successful analysis by Maxtract the PDF file must make sole use of Type 1 fonts with embedded encodings. This is not the case for the part of born-digital PDF documents in the digital library as well as for none of the scanned retro-digitized documents. In this case, InftyReader[5] OCR software can be used as this program has the unique feature of recognition of mathematical expressions in scanned documents. InftyReader accepts various bitmap image formats on the input (TIFF, BMP, GIF, PNG, PDF) and transforms them to LaTeX, XHTML+MathML and various XML formats. If the quality and resolution of input scans is sufficient, InftyReader produces reasonably good outputs that can be successfully used for preparation of accessible documents.

The widest possible use have text-based formats thanks to simplicity and support in screen-readers.

3 http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/maxtract.php

4 http://www-sop.inria.fr/marelle/tralics/

5 http://www.inftyproject.org/

3.1 CopyMath

pdfTEX users can enhance their PDF documents with various hidden text information. For this purpose, we can use the ability of pdfTEX to put into the output PDF document raw PDF commands. Including `/ActualText` to the code of the PDF enables us to provide users with different visual and text information.

```
\documentclass{article}
\begin{document}

Standard text.
\pdfliteral{/Span << /ActualText<\pdfescapehex{\detokenize{Copya-
  ble text.}}> >> BDC}
Printed text.
\pdfliteral{EMC}
Another standard text.

\end{document}
```

[Listing 2] Minimal example of use of ActualText PDF command use.

For example, PDF documents produced by pdfLATEX from the source code shown in Listing 2 visually presents the following text:

```
Standard text. Printed text. Another standard text.
```

However, following text is provided to the use by a PDF viewer when the copy&paste function is used or text search though the document is performed:

```
Standard text. Copyable text. Another standard text.
```

To allow users to use this technique as easily as possible for improvement of accessibility of mathematical contents of scientific documents we are experimenting with the CopyMath LATEX macro package. That package should allow user-friendly creation of layered PDF such that mathematical equations could be copy-pasted as original textual LATEX source representation. For details of the approach see [11].

3.2 Textual Output – Reading Formulae Aloud

We have developed MathML-to-text application[6] that processes XML files that contain one or more MathML blocks (plain math-texts) and converts each such block into plain text to be read aloud. This format can be useful as an input for speech synthesizer software, for computing similar articles [10] or for indexing and searching.

Input file is parsed using streaming API for XML. MathML block is then transformed into simplified DOM model. Based on MathML type, Presentation or Content, slightly different method is used. If both types are present Content MathML takes precedence as Presentation MathML can be ambiguous and unclear. Either way final result will be more or less the same.

Conversion mechanism is able to process most arithmetic operations, trigonometric functions, logical, set, and comparison operations. Many specific operations, for example

6 https://code.google.com/p/mathml-converter/

limit, summation, product, integral, etc., are also implemented. The same holds for well known mathematical constants, as well as widely used symbols. Based on input parameters, conversion of numbers is also possible. So far only output to English language is supported. Other languages may come in the future. Output file is a valid XML file with all occurrences of MathML substituted with converted plain text.

3.3 Braille Output

There is no simple language independent encoding of math in Braille – there are different versions for English, German, French or Czech. We have decided to support Czech version of Braille [7] by the development of the tool converting canonical MathML to Czech Braille [9]. The tool is implemented as one of the UMCL output drivers.

4 Conclusions, Future Work

We have presented our approaches and tools to increase the accessibility of mathematical texts in digital libraries. Our workflow adds math handling to the processing pipe and our tools are able to automate the Braille output generation, production of PDF suitable for copy&paste including math, . . . We hope that this work finds followers to the benefit of students and researchers in STEM domains.

Acknowledgement

This work has been partially supported by the European Union through its Competitiveness and Innovation Programme (Information and Communications Technologies Policy Support Programme, "Open access to scientific information", Grant Agreement No. 250503). We also thank for feedback we got from conference participants.

References

[1] Archambault, D.; Moço, V. Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In K. Miesenberger, J. Klaus, W. Zagler, A. Karshmer (Eds.) *Computers Helping People with Special Needs*, *Lecture Notes in Computer Science*. Volume 4061, pages 1191–1198. Springer: Berlin / Heidelberg, 2006. http://dx.doi.org/10.1007/11788713_172.

[2] Ausbrooks, R.; Buswell, S.; Carlisle, D.; Chavchanidze, G.; Dalmas, S.; Devitt, S.; Diaz, A.; Dooley, S.; Hunter, R.; Ion, P.; Kohlhase, M.; Lazrek, A.; Libbrecht, P.; Miller, B.; Miner, R.; Rowley, C.; Sargent, M.; Smith, B.; Soiffer, N.; Sutor, R.; Watt, S. *Mathematical Markup Language (MathML)* [online]. Version 3.0, 2010. W3C Recommendation 21 October 2010. Available in URL <http://www.w3.org/TR/2010/REC-MathML3-20101021/>.

[3] Baker, J. B.; Sexton, A. P.; Sorge, V. Towards Reverse Engineering of PDF Documents. In P. Sojka; T. Bouche (Eds.) *Towards a Digital Mathematics Library*. Bertinoro, Italy, July 20–21st, 2011. Brno: Masaryk University, July 2011. Pages 65–75. http://hdl.handle. net/10338.dmlcz/702603.

[4] Bartošek, M.; Lhoták, M.; Rákosník, J.; Sojka, P.; Šárfy, M. DML-CZ: The Objectives and the First Steps. In J. Borwein, E. M. Rocha, J. F. Rodrigues (Eds.) *CMDE 2006: Communicating Mathematics in the Digital Era*. A. K. Peters, MA, USA, 2008. pages 69–79.

[5] Bouche, T. CEDRICS: When CEDRAM Meets Tralics. In P. Sojka (Ed.) *Proceedings of DML 2008*. Birmingham, UK, July 2008. Brno: Masaryk University, 2008. Pages 153–165. http: //dml.cz/dmlcz/702544.

[6] Formánek, D.; Líška, M.; Růžička, M.; Sojka, P. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, P. Libbrecht (Eds.) *24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress*, number 921 in *CEUR Workshop Proceedings*. Aachen, 2012. Pages 91–103. http://ceur-ws.org/Vol-921/wip-05.pdf.

[7] Gonzúrová, W. *Handbook on Print Transcription Using the Dot Font Norm*. Prague, Czech Republic: K. E. Macan's Library and Publishing House for the Blind, 1997. Available (in Czech) on http://www.teiresias.muni.cz/czbraille/.

[8] Grimm, J. Producing MathML with Tralics. In Sojka [12]. Pages 105–117. http://dml.cz/dmlcz/702579.

[9] Jarmar, M. *Conversion of Mathematical Documents into Braille*. Master's thesis (advisor: Petr Sojka), Faculty of Informatics. Brno: Faculty of Informatics, Masaryk University, 2012. https://is.muni.cz/th/172981/fi_m/?lang=en.

[10] Lee, M.; Sojka, P.; Řehůřek, R. *Toolset for Entity and Semantic Associations – Value Release*. Deliverable D8.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, 2012. https://project.eudml.eu/sites/default/files/D8.3-v1.0.pdf.

[11] Růžička, M.; Sojka, P. Data Enhancements in a Digital Mathematics Library. In Sojka [12]. Pages 69–76. http://dml.cz/dmlcz/702575.

[12] Sojka, P. (Ed.). *Towards a Digital Mathematics Library*. Paris, France, July 2010. Brno: Masaryk University, 2010. http://www.fi.muni.cz/~sojka/dml-2010-program.html.

[13] Sojka, P. Exploiting Semantic Annotations in Math Information Retrieval. In J. Kamps, J. Karlgren, P. Mika, V. Murdock (Eds.) *Proceedings of ESAIR 2012 c/o CIKM 2012*. Maui, Hawaii, USA: Association for Computing Machinery, 2012. Pages 15–16. http://doi.acm.org/10.1145/2390148.2390157.

[14] Sojka, P.; Líška, M. The Art of Mathematics Retrieval. In *Proceedings of the ACM Conference on Document Engineering*, DocEng 2011. Mountain View, CA: Association of Computing Machinery, Sept. 2011. Pages 57–60. http://doi.acm.org/10.1145/2034691. 2034703.

[15] Sojka, P.; Wojciechowski, K.; Houillon, N.; Růžička, M.; Hatlapatka, R. *Toolset for Image and Text Processing and Metadata Enhancements – Value Release*. Deliverable D7.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, 2012. https://project.eudml.org/sites/default/files/D7.3.pdf.

[16] Sorge, V.; Lee, M.; Sojka, P.; Sexton, A. P. *State of the Art of Accessibility Tools.* Deliverable D10.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, 2011. https://project.eudml.eu/sites/default/files/D10.1.pdf.

[17] Suzuki, M.; Tamari, F.; Fukuda, R.; Uchida, S.; Kanahori, T. INFTY – An integrated OCR system for mathematical documents. In C. Vanoirbeek, C. Roisin, E. Munson (Eds.) *Proceedings of ACM Symposium on Document Engineering 2003.* Grenoble, France: ACM, 2003. Pages 95–104.

[18] Sylwestrzak, W.; Borbinha, J.; Bouche, T.; Nowiński, A.; Sojka, P. EuDML – Towards the European Digital Mathematics Library. In Sojka [12]. Pages 11–24. http://dml.cz/dmlcz/702569.