EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICTPSP

ICT POLICY SUPPORT PROGRAMME

part of the Competitiveness and Innovation Framework Programme CIP

DEMO

Project Acronym: EuDML

Grant Agreement number: 250503

Project Title: The European Digital Mathematics Library

D10.4: Enhancement and Deployment of Final Toolset

Revision: 1.0 as of 3rd December 2012

Authors:

Josef Baker	University of Birmingham
Alan P. Sexton	University of Birmingham
Volker Sorge	University of Birmingham

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	✓
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	20th November 2012	Alan P. Sexton	UB	Submitted first revision to SVN.
0.2	26th November 2012	Alan P. Sexton	UB	Completed dyslexia section.
0.3	28th November 2012	Volker Sorge	UB	Completed accessibility section.
0.4	29th November 2012	Volker Sorge	UB	Completed introduction and conclusions.
0.5	30th November 2012	Petr Sojka	MU	Just typo corrections during review reading.
1.0	2nd December 2012	Volker Sorge	UB	Final editing.

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Abstract
In this document we describe the final release of the accessibility toolkit for creating accessible mathematical documents in a wide variety of formats, designed to be compatible with a variety of software. We also give a brief overview of the processes and output of the toolkit.

Contents

List of Figures

1 Introduction

The goal of work package 10 is to enable accessibility to some of the mathematical documents within EuDML for print impaired users. This means users that have problems accessing regular print either due to visual impairments or learning disabilities such as dyslexia. An additional benefit of enhancing content for accessibility purposes is that all users can profit from advanced interaction possibilities, such as copying or searching.

In the last two deliverables of this workpackage (D10.2, D10.3) we primarily focused on accessibility support for the visually impaired by putting together an initial set of accessibility tools in D10.2 [3] and evaluating it with the help of a number of potential users — blind mathematicians and computer scientists, educators for blind students as well as accessibility support officers that prepare teaching material for blind students — in D10.3 [4]. In this report we present the final toolkit as integrated into EuDML that has been motivated by the outcome of the previous evaluation. In addition, we have included an enhancement providing one further format that is particularly geared towards web browser display and support for browser based text-to-speech support.

Beyond providing support for visually impaired users, EuDML attempts also to provide specialist support for dyslexic users. The state of the art was also reviewed in deliverable D10.1 [5]. This support has now been realised in the final toolkit by providing search term completion specifically for mathematical keywords that should aid dyslexic users.

In the next section we therefore first present the EuDML support for assisting in search term entry for dyslexics. General text-to-speech support for visually impaired or dyslexic users is then discussed in the subsequent section.

2 Support for Dyslexics

Dyslexia is a brain-based learning disability that interferes with the ability to learn to read. From the point of view of EuDML, there are two major aspects of support that is provided for users with this disability. For actual reading of EuDML web pages and papers, the text-to-speech support for visually impaired users is available. However, there remains the problem of entering correct search terms for users who suffer from the very common symptom among dyslexics of orthographic coding difficulties. This can be particularly challenging in mathematics given the technical vocabulary and the importance of mathematician's names for the identification of theorems, methods, concepts and bibliographical references.

2.1 Interface

The main user interface in EuDML is essentially a search engine, in which users enter terms of interest to their search. If the user is dyslexic, the terms they enter have an exceptionally high probability of being misspelled, thus hindering successful searching. In the final version of EuDML, text search fields prompt the user with phrase completions from a mathematical phrase list as soon as three or more characters have been typed in.

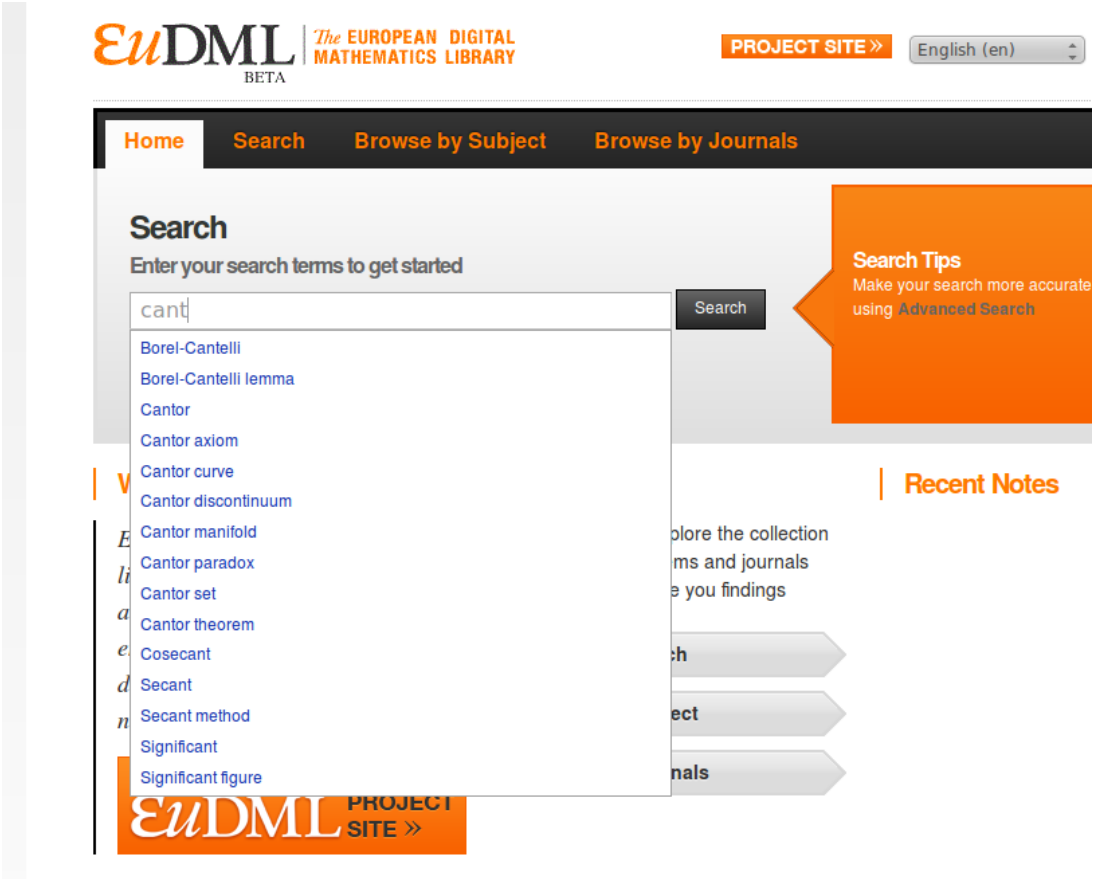

Figure 1: Search Completion for Dyslexic Support

In this way, users can select the (correctly spelled) phrase that most closely matches their search requirements without having to get their spelling perfectly correct.

2.2 Implementation

Phrase completion is provided by a standard AJAX completion mechanism that is already very familiar to users from, for example, Google search.

The completion mechanism is dependent on a lexicon of mathematical terms. To obtain this lexicon, titles of entries were extracted from the Encyclopedia of Mathematics [6]. This is an online, open access resource for the mathematics community. The original articles in the encyclopedia were from an online encyclopedia published by Kluwer Academic Publishers in 2002. Springer Verlag and the European Mathematical Society have collaborated on releasing the content to the public.

The titles of all entries in the encyclopedia were extracted and hand edited to remove duplications and phrases that were deemed to not be specifically mathematical in nature. The end result was a list of 8,160 phrases. A vocabulary list of 5,061 mathematical individual words and names were also extracted, but, following some experiments, it was decided that the phrase list was more useful to the purposes of EuDML.

Since the lexicon is purely a phrase list, it can be easily extended in the future by adding more phrases. However, increasing the size of the lexicon beyond a certain point becomes counterproductive because a very large number of suggested terms may obscure suitable choices more than it assists the dyslexic reader. Therefore EuDML will await user comments before deciding to further extend, or indeed, trim, the lexicon.

2.3 Demonstration

The dyslexia support can be easily explored via the current EuDML demo website at `http://eudml.org`. When inputting search terms into the search field, possible completions for mathematical search terms are computed instantly.

3 Accessibility Support

To support accessibility of full articles in EuDML we are able to serve a range of different formats that allow users access to content by a variety of means such as screen readers, caret browsing etc. These content formats are generated using the MaxTract system [1] from suitable articles.

As we have discussed both the different formats and the restrictions on our approach in previous reports [5, 3, 4], we will only summarise these points here. However, we have also developed an additional accessibility format as further enhancement to our accessibility toolkit. This format is particularly suited for screen readers and browser providing full ARIA support and we will discuss the format in more detail.

3.1 Accessibility Formats

Following the results of our evaluation of accessibility formats that was presented in deliverable D10.3 [4], we had decided to go with a limited set of formats that cater for particular needs and accessibility tools or screen readers. We will briefly recap these formats here.

Layered PDF These are PDF documents that contain, in addition to the regular mathematical article, two normally hidden layers. One containing normal ASCII text where all mathematical formulas and special notations are replaced by alternative text. A second layer provides the LaTeX code of every page. The main idea of providing this format is to cater for all users of the document: the regular document layer for sighted users, the alternative text layer for visually impaired users to be accessible with a screen reader and the code layer to enable copy and paste functionality for the formulas. The drawback of the format is that it uses specialist PDF markup that most non-Adobe viewers are incapable of interpreting correctly.

An example of a layered PDF produced by MaxTract is shown in Figure 2, containing screen-shots of the file when opened in Adobe Reader X. In the first screen-shot the reproduced display layer is open, in the second the LaTeX layer is open and the final screen-shot shows the plain text layer. Note that on the left hand side of each shot is the menu for choosing which layers are currently in use.

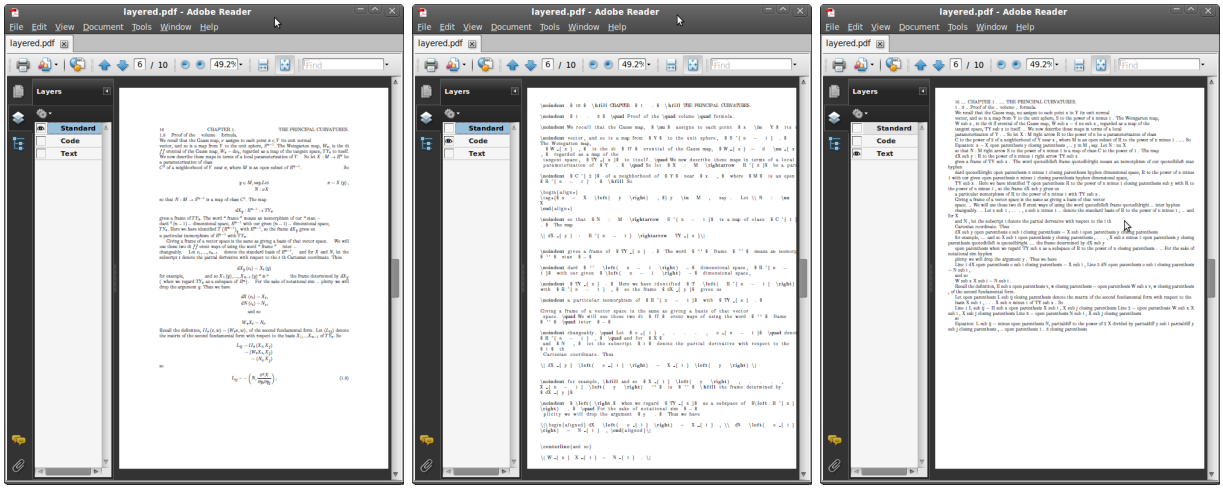

Figure 2: Viewing a multi-layered PDF file in Adobe Reader 9.

Annotated PDF A document of this form combines in a single layer both the normal content with markup for mathematical formulas. In fact, it closely resembles the original document, where mathematical formulas have additional tags, which lead to their representation in LaTeX markup. In our original evaluation we tested documents that contained annotations for both LaTeX and MathML markup. However, these generally led to documents that were difficult to read as the annotations overlayed some of the content. It was also felt that for accessibility purposes, visually impaired users would profit more from the LaTeX markup than from MathML. Similarly, if users were interested in copying formulas they would usually prefer LaTeX as well. Consequently in the final version we include documents with LaTeX annotations only.

Again a disadvantage of this format is that it can only be viewed properly in Adobe's Acrobat Reader but usually not in other viewers.

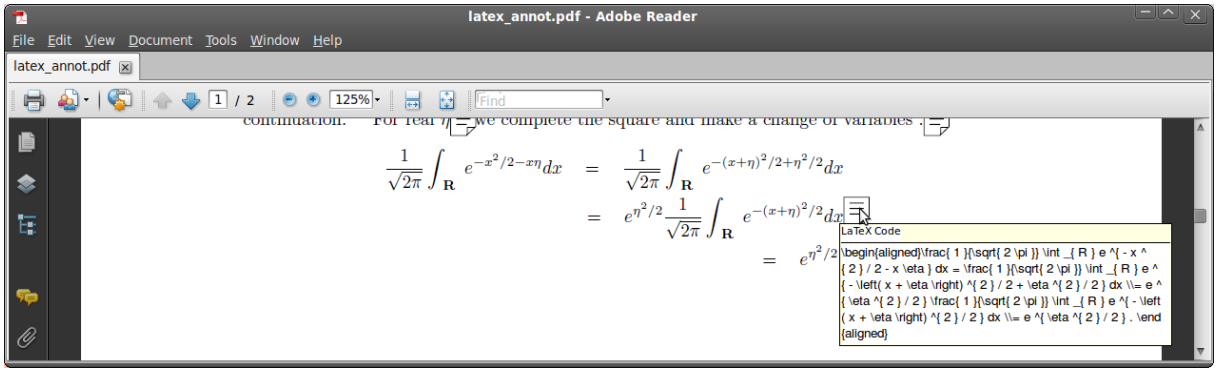

Figure 3: Viewing LaTeX annotations in Adobe Reader 9.

Examples of MaxTract-produced annotated PDF files open in Adobe Reader are shown in Figure 3. The underlying markup for a formula is viewed by hovering over or clicking upon the adjacent annotation symbol.

LaTeX This is an ASCII file containing full LaTeX markup. That is, it can be compiled easily into a PDF itself. While this is probably not a worthwhile goal, given that the original PDF with superior layout is available from EuDML, the idea is to provide a format that users familiar with LaTeX can read directly.

Text only This is similar to the alternative text layer in the layered PDF. The idea behind this format is to provide a simple ASCII file without any markup or hidden within layers. This enables a visually impaired reader to give the file directly to any text-to-speech engine of their choice without particular software restrictions.

3.2 Enhanced Accessibility Format

In addition to the accessible formats that MaxTract could produce previously and that were evaluated for deliverable D10.3 [4], we have now developed a further enhanced format that is provided via our accessibility toolkit. The idea of the format is to exploit *ARIA roles* to provide alternative text for mathematical content.

ARIA (Accessible Rich Internet Applications Suite) is a standard proposed by W3C's web accessibility initiative with the goal of supporting accessibility of web pages for disabled users (cf. `http://www.w3.org/WAI/intro/aria.php`). It not only focuses on support for screen readers but also on providing accessibility to features of web pages for users who can not use graphical input devices like mice. ARIA roles can be provided for particular of the segments of the HTML DOM that can be exploited by software tools like screenreaders to achieve certain effects. For example, input boxes can be pointed out to user, on dynamic webpages the user can be alerted to changes etc. The ARIA specification also provides a particular role in order to markup mathematics in web documents.

As an example we consider a simple expression like the quadratic formula

$$ax^2 + bx + c = 0.$$

Translating formula directly into MathML yields rather unwieldy markup, which is generally indigestible for current screen reading technology. However, the following HTML markup encapsulates the MathML into a `div` node containing a math ARIA role as well as a corresponding ARIA label. The latter contains an alternative text expression for the quadratic formula.

```
<div role="math"
     aria-label="a times x squared plus b times x plus c equals 0">
  <math xmlns='http://www.w3.org/1998/Math/MathML'>
    <mrow>
      <mrow>
        <mrow>
          <mi>a</mi>
          <mo> &InvisibleTimes; </mo>
          <msup>
            <mi>x</mi>
            <mn>2</mn>
          </msup>
        </mrow>
```

```
          <mo>+</mo>
          <mrow>
             <mi>b</mi>
             <mo> &InvisibleTimes; </mo>
             <mi>x</mi>
          </mrow>
          <mo>+</mo>
          <mi>c</mi>
       </mrow>
       <mo>=</mo>
       <mn>0</mn>
     </mrow>
  </math>
</div>
```

Although the possibility of adding ARIA math roles as markup does exist already for some time, our experience shows that authors of online mathematics have made very limited use of this feature. Mainly, because there is relatively little support for the feature and it costs considerable additional work for authors to provide meaningful alternative text. However, with both the ability of generating text and markup automatically as well as with anticipated future support for math ARIA roles by browsers and screen readers (for example support will be available in upcoming version of the Google Chrome browser), we consider it worthwhile to provide an accessible format exploiting this feature.

In our new format we now combine MaxTract's ability to generate both MathML and alternative text for every math expression it encounters. MathML content is simply wrapped in an ARIA role that contains a label with the textual description provided by MaxTract. In practice, MathML is generated from MaxTract's LaTeX output using Tralics [2]. The ARIA roles are then added retrospectively using a Python script.

3.3 Interface

The accessibility formats are reachable via the common EuDML interface. The idea is that for every document for which accessible formats exist, these are provided via links on the result page for a particular article. An example of this is displayed in Figure 4.

3.4 Demonstration

The necessary templates for the full integration of the accessibility format into EuDML are currently under preparation for the final release. Accessibility links will be automatically included and displayed if the necessary content is available. For a static demonstration of the display see for example `http://devel.eudml.org/unstable/doc/219377`.

4 Conclusions

This deliverable presented the final implementation of the enhanced accessibility toolkit for EuDML. In addition to the formats that have been previously identified to be beneficial for accessibility we have implemented an additional format that makes use of ARIA roles that can be supported by browser based screen readers. Furthermore, we have provided

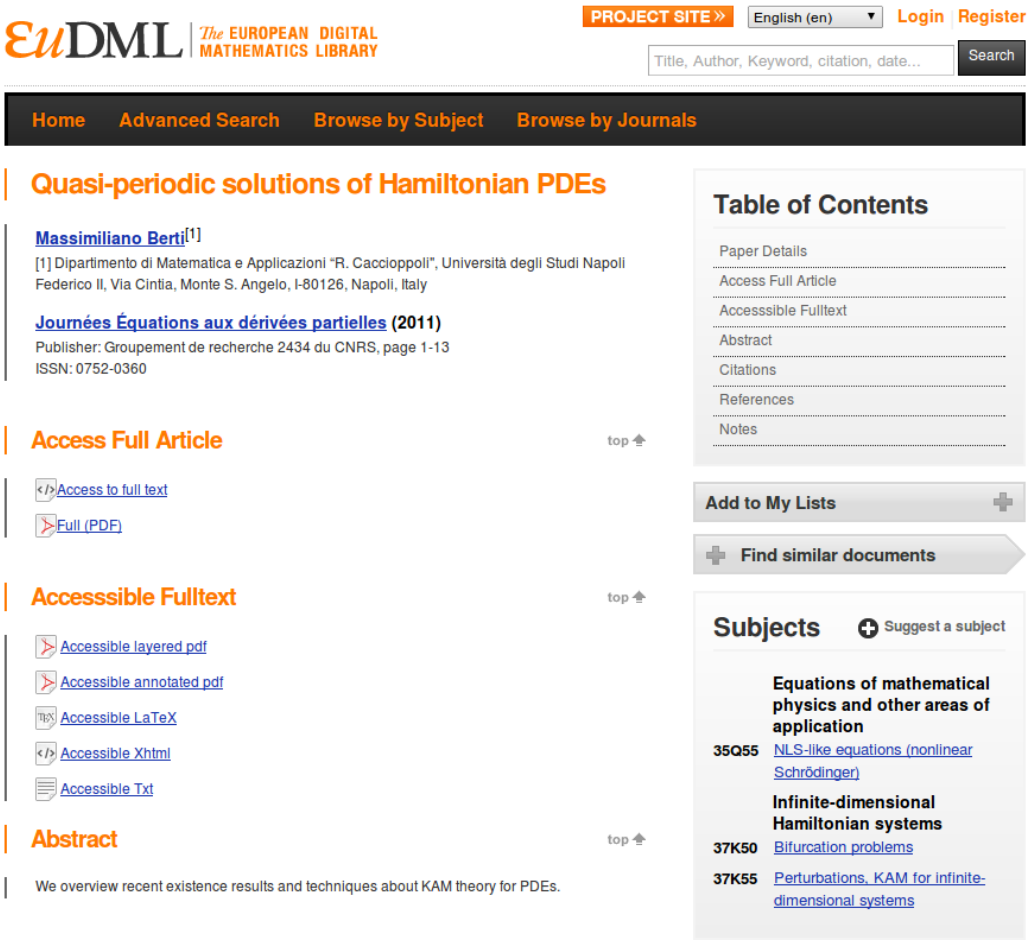

Figure 4: Provision of Accessibility Formats in EuDML

mathematics specific support for search term completion that aims at helping particularly dyslexic users. However, we believe that it is also of great value to regular users. Similarly, the accessible formats can be useful to any user, as they provide advanced support for accessing mathematics — for the purposes of searching, copying, etc. — on a scale that has never been achieved before.

The accessibility toolkit relies on the MaxTract tool to be able to extract content from existing PDF files. Thereby there exist restrictions on the format of these PDF files, in that they have to be digitally born, in PDF version 1.3 or above, and contain only Type 1 fonts. Furthermore, we can only provide accessible formats for content, for which there are not copyright restrictions with respect to the full text. As a consequence, at least of the time being, fully accessible formats can only be provided for a relatively small number of the EuDML collections. Currently, we have processed 1,100 documents from CEDRAM and DML-CZ, but we expect that by the end of the project to have also processed documents from other collections (such as PLDML, ELIBM, NUMDAM, and BULDML) and estimate the number of documents to be above 5,000. Such a large

number of mathematical documents have never before been made available in accessible formats in one place.

In addition, we can expect that new content added to EuDML in the future will be digitally born PDFs that can directly be made accessible with the existing workflow. Similarly, the quality of the accessible content will improve in the future as the MaxTract system improves. MaxTract is an ongoing research project that is continuously improved and refined and updated versions of MaxTract can and will be included in the EuDML workflow in the future.

Some of the future improvements we are working on are improved formula identification. For example, currently single vertical lines or words with large numbers of accented or diacritics are often mistaken for math formulas. We are also integrating a new histogrammatic line and formula separation technique that should improve error stemming from faulty separation of multi-line formulas. Other future improvements concern more context sensitive and domain specific translations of formulas into alternative text as well as broadening the scope of the textual translation of formulas to languages beyond English by exploiting i18n tools.

References

[1] Josef B. Baker, Alan P. Sexton, and Volker Sorge. Towards reverse engineering of PDF documents. In Petr Sojka and Thierry Bouche, editors, *Towards a Digital Mathematics Library, DML 2011*, pages 65–75, Bertinoro, Italy, July 2011. Masaryk University Press. http://hdl.handle.net/10338.dmlcz/702603.

[2] José Grimm. Tralics, a LaTeX to XML Translator. *TUGboat*, 24(3), 2003. http://www-sop.inria.fr/apics/tralics/.

[3] Volker Sorge and Josef Baker. Accessibility Toolset – Deployment in Live System, March 2012. Deliverable D10.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.eu/sites/default/files/D10.2.pdf.

[4] Volker Sorge and Josef Baker. Accessibility Toolset – Evaluation, August 2012. Deliverable D10.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.eu/sites/default/files/D10.3.pdf.

[5] Volker Sorge, Mark Lee, Petr Sojka, and Alan P. Sexton. State of the Art of Accessibility Tools, February 2011. Deliverable D10.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.eu/sites/default/files/D10.1.pdf.

[6] Springer Verlag and The European Mathematical Society. Encyclopedia of mathematics, 2012. http://www.encyclopediaofmath.org.