Toolset for image and text processing and metadata enhacement - Final release

Petr Sojka, Krzysztof Wojciechowski, Nicolas Houillon, Michal Ružicka, Maciej Koluda, Radim Hatlapatka, Thierry Bouche, Franck Lontin, Vlastimil Krejcír, Gilberto Pedrosa, et al.

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

D7.4: Toolset for Image and Text Processing and Metadata Enhancements — Final Release

Revision: 1.0 as of 9th February 2013

Authors:

Petr Sojka	Masaryk University, MU
Krzyś Wojciechowski	ICM Warsaw, ICM
Nicolas Houillon	UJF/CMD Grenoble

Contributors:

Michal Růžička	Masaryk University, MU
Maciej Kołuda	ICM Warsaw, ICM
Radim Hatlapatka	Masaryk University, MU
Thierry Bouche	UJF/CMD Grenoble
Franck Lontin	CNRS/CMD Grenoble
Vlastimil Krejčíř	Masaryk University, MU
Gilberto Pedrosa	IST Lisbon, IST
Miroslav Hrdina	Masaryk University, MU
Jiří Sochor	Masaryk University, MU
Pavel Rychlý	Masaryk University, MU
Aleš Horák	Masaryk University, MU
Alan Sexton	University of Birmingham, UB

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	Dec 1st, 2012	Petr Sojka	MU	First version with structure of the deliverable.
0.2	Dec 6th, 2012	Maciej Kołuda	ICM	Added 'Deduplication and merging' section.
0.3	Dec 7th, 2012	Krzyś Wojciechowski	ICM	'Deduplication and merging' section rewriten.
0.4	Jan 30th, 2013	Petr Sojka	MU	First coherent version, still with TODOs.
0.5	Feb 4th, 2013	Thierry Bouche Frank Lontin	UJF	Infty at Mathdoc.
0.6	Feb 7th, 2013	Michal Růžička	MU	References to D7.3.
0.7	Feb 7th, 2013	Petr Sojka	MU	Intro, TODO partial closing.
0.8	Feb 8th, 2013	Petr Sojka Gilberto Pedrosa	MU IST	Conclusion, index. OCT at IST.
0.9	Feb 8th, 2013	Petr Sojka	MU	Cleanup. Version for internal review.
0.91	Feb 8th, 2013	Alan Sexton	UB	Internal review and some cleanup
0.92	Feb 8th, 2013	Petr Sojka	MU	Comments from Internal review reflected
1.0	Feb 9th, 2013	Petr Sojka	MU	Final release.

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Contents

Executive Summary

This demonstration description presents tools and partial workflow results produced by EuDML [partners] and either integrated and used in core EuDML processing and/or made available as standalone tool or as demonstrations. Enhancement workflow and tools whose functionality should find, check, merge, correct and enhance metadata and text or PDF document full text of items in the EuDML collection are described. Demonstration web pages allow testing and evaluation of these tools, in addition to the project site itself, where enhanced data are projected.

Demonstrator URLs

`http://nlp.fi.muni.cz/projekty/eudml/eudmldemo.php`

`http://www.zentralblatt-math.org/eudml/lookup?zblid=1163.57016`

`http://www.zentralblatt-math.org/eudml/match?query=B24, Inhomogeneous fixed point ensembles, 1811`

`http://eudml.mathdoc.fr/eudml-demo/EnhanceNLMTeXwMathML`

`http://eudml.mathdoc.fr/eudml-demo/TeX2NLM`

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

`http://server.bd2.inesc-id.pt/inftyEudml/`

`https://project.eudml.org/EuDmlAnalysisDemo/`

"The only thing standing between you and your goal is the bullshit story you keep telling yourself as to why you can't achieve it."
Jordan Belfort

1 Introduction

Data and metadata ingested in EuDML from data providers are sometimes suboptimal, incomplete, or in formats not directly usable for EuDML services. Some items are duplicated, and need to be discarded, some have to be enhanced from versions from verified sources, some have to be converted into different format and some have to be edited on content provider site manually. This demonstration deliverable describes the final versions of tools that have been used in the project for these purposes. The tools needed were identified in D7.1 [22], an initial version of the toolset has been described in D7.2 [24] and a value release with work-flows in D7.3 [27]. This demonstration presents the final status of tools developed, integrated and used on the EuDML data from data providers. It is not a full description—to avoid the repetition of information from previous deliverables, we simply refer to them. For the tools not described previously, we describe functionality, usability, scalability, and use in a comprehensive workflow integrated in the EuDML enhancement subsystems.

1.1 The Toolset Structure

Let us recall the toolset structure, with *subsystems* represented as edges in Figure 1.

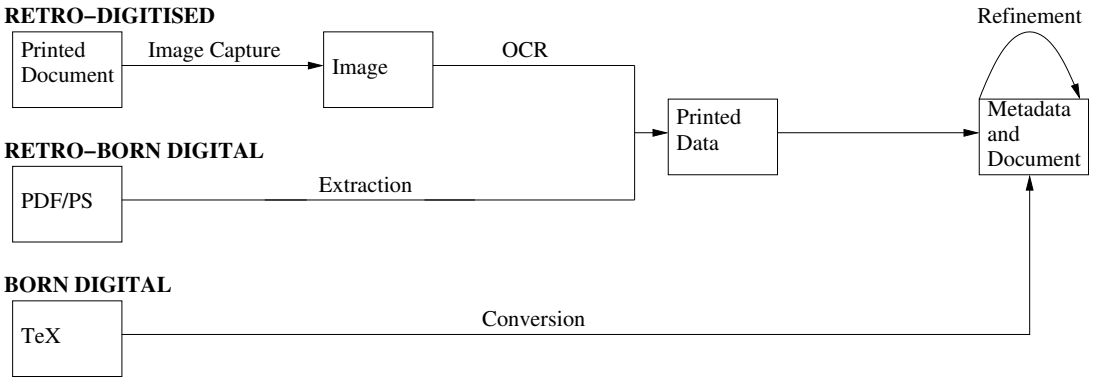

Figure 1: Metadata processing

The toolset consists of five subsystems: OCR, Extraction, Conversion and Refinement, and a set of external tools offered to data providers (External). Subsystems are built based on the smaller bricks of software—*eutools*—as defined in [22, Section 1.3]. Eutools were developed and tested, usually at the technology providers' sites, with well defined interfaces allowing their integration into subsystems on the EuDML core system site. The toolset consists of the eutools listed in Table 1 on the following page.

Interfaces used by eutools within EuDML were listed in D7.2 [24, Section 2].

Table 1: Eutools overview

Subsystem	Partner	Eutool	Functionality
OCR	MU	**PdfToTextViaOCR**	Basic plaintext extraction from bitmap images which are rendered from a PDF document.
OCR	IST	**PdfToText+MathViaOCR**	Plaintext and Math extraction from bitmap images which are rendered from a PDF document.
Extraction	ICM	**PDF Text Extractor**	Extracts plain text from a PDF document.
Extraction	UB	**MaxTract**	Analyses a PDF document and produces various accessible versions of the document including LaTeX, MathML and plain text.
Conversion	CMD	**TeX2NLM**	Identifies TeX formulas in a character string and replaces each with an NLM structure with both TeX and MathML alternatives.
Conversion	CMD	**EnhanceNLMTeXwMML**	Takes an NLM document and 1) adds a MathML alternative to each formula represented by the NLM formula structure that has a TeX version but not MathML, and 2) finds textual TeX formula and replaces each with an NLM formula structure containing both the original TeX and its MathML equivalent.
Refinement	FIZ	**ZBMath Lookup Service**	Get Zbl metadata for a publication given its Zentralblatt identifier.
Refinement	FIZ	**ZBMath Match Service**	Get ZblId and Zbl basic metadata for a publication given its bibliography reference as string.
Refinement	MU	**PdfJbIm**	Recompress bitmap streams in a PDF document with JBIG2.
Refinement	MU	**math_metadata_lookup**	MR/Zbl metadata search and fetch.
External	MU	**ME**	Metadata Editor—standalone editing for use at providers' sites.

1.2 Structure of the Demo Description

We start with following Section 2 by listing eutools that have been implemented, and refer to the eutool description together with the web pages that demonstrate their functionality.

Section 3 describes WP7 workflows—how eutools are integrated into a processing pipeline. Once the data are enhanced, it is time to find duplicates and to choose the best (richest) canonical representant of an EuDML item—how it is done is described in Section 4. Section 5 sums up demonstrated tools and other achievements of WP7.

2 Eutools

In this section basic information is provided for every tool, including their defined input and output interfaces, license information, programming language and evaluation, as well as the URL of the demonstration web site.

2.1 PdfToTextViaOCR

PdfToTextViaOCR is a tool written in Java which renders images from PDF and extracts text from them using an OCR engine. The OCR engine used is Tesseract [13], which is licensed under Apache License 2.0[1] and supports a wide variety of languages.

This tool is used as a fallback solution in cases where the use of **PdfToText+MathViaOCR** is not applicable.

Of 139,600 processed PDF files, 12,344 has been processed by **PdfToTextViaOCR** inside the EuDML system. Most of PDF files were processed using **PdfToText+MathViaOCR** (see Section 2.2), as this tool also produces MathML for indexing.

For more detailed information we refer to D7.3 [27], only minor updates has been done since D7.3.

A demonstration standalone version of this tool is available at `http://nlp.fi.muni.cz/projekty/eudml/eudmldemo.php`.

2.2 PdfToText+MathViaOCR

PdfToText+MathViaOCR is an integrated mathematical document reader system that uses the INFTY system [28, 14, 18] for OCR (Optical Character Recognition). The system recognizes PDFs with scanned page images of scientific documents, including mathematical expressions, and outputs the results in LaTeX, HTML, XML etc.

After harvesting the full text content of the digital records using REPOX, the system will identify those which contain only bitmap page images via **MaxTract**[2], and use the **PdfToText+MathViaOCR** to convert to rich full text representation (LaTeX or XHTML or plain text with math) and store the result in the EuDML storage system. The interface used for passing the results to the EuDML core is the HTTP protocol.

A further description of the tool can be found in D7.3 [27].

A demonstration version of this tool is available at `http://server.bd2.inesc-id.pt/inftyEudml/`.

As OCR recognition times are around ten seconds per page on a modern PC, to speed up developments it has been agreed that IST, UJF and MU will do the digitization in parallel and will synchronize the OCR workflow developments: UJF will prepare full texts of NUMDAM, MU will digitize DML-CZ and IST will do the rest.

2.2.1 OCR at MU

The Czech Digital Mathematical library (DML-CZ) contains around 300,000 pages of most important mathematical scientific works published within Czech or Slovak Republic. The DML-CZ data include 13 journals, 6 proceedings and 4 monographies in multiple languages (Czech, Slovak, English, German, Russian, French and more). The data is mostly retro-born digital, which means that their digital form were not existent before digitalization. For EuDML purposes the OCR part was done again using solely the INFTYREADER software. Most of the data was recognized using version 2.9.4, and only the first

1. `http://www.apache.org/licenses/LICENSE-2.0`
2. In previous deliverables this functionality was provided by a **MaxTract** predecessor tool called **PDF-Tester**

3 journals were done with 2.9.3. At the time of writing we already have version 2.9.5 with improved Russian support, as a result of visit to Brno in December 2012 by Prof. Suzuki, the main INFTY developer.

The total number of articles needed to be recognized was around 31,000. To date, full texts with math of 24,443 of these items have already been ingested into EuDML. The data used for recognition came as images in TIF format, which are mostly 600 DPI. The images were mostly bitonal, the rest was in grayscale. The recommendation for INFTYREADER is to use it with binary images, but the testing showed that the grayscale images performed reasonably well. The software can be run either from a GUI or via a command line interface. The latter was used in batch processing of the DML-CZ data.

The batch processing showed that the time needed for OCR of a single image is roughly 8.9 seconds. This would mean that the whole DML-CZ library would take roughly 742 hours to complete. The time was further reduced to a quarter by processing in 4 threads so that the recognition would take slightly over a week of continuous OCR to finish.

Two formats were considered as the recognition output: LaTeX, with additional conversion to MathML with Tralics software, and XHTML with MathML markup. During the recognition, issues were found with both outputs. The output files often contained errors, which prevented the output from being used in further processing in the EuDML workflow. Some of the errors were solved by modification of the `ConvertTable.tbl` configuration file, which controls the generated output of INFTYREADER. The rest were reported to the `ConvertTable.tbl` creators in Japan which resulted in the release of a new updated version of the program. Unfortunately some errors remained and post-processing scripts were used on XHTML output in order to have at least one error-free output. In this way the XHTML with MathML output format became viable option. The LaTeX output format still had errors whose removal were not so simple—there were mostly issues with math mode not being correctly terminated or not present in the right place.

For error testing a set of 50 articles in multiple languages and image preferences was selected. The LaTeX output had 33 TeX errors in total. Most of the errors were caused by `Cyrillic_be` error (21), then there were 6 errors related to misuse of math mode, 5 double superscript errors and 1 invalid character error. The XHTML output produced many more errors, but all of them were resolved with post processing. In total XHTML output contained 306 errors. Approximately, half of them (153) were caused by mispelling the `&Prime;` as `&Prim;`. The second big group were errors caused by images created during recognition (118). Then there were 26 occurrences of `&tprim;`, instead of the correct `&tprime;`, element 'u' undefined (6) and finally 3 errors caused by invalid characters.

We plan to redo the recognition once we check that latest INFTY version fixes all encountered problems. After automation via the Windows scripting language *AutoIt* and other tools it takes about one week of continuous running on a dedicated virtual machine.

2.2.2 OCR at UJF

NUMDAM contains around 700,000 pages of mathematical works published generally in France. The NUMDAM data sets include 34 journals, 29 seminar proceedings and 3 series of memoirs in multiple languages (the most frequent being: French, English, and

Italian). The data is mostly retro-digitised. Each individual article has been cleaned in such a way that if two articles are printed as a sequence on the same page, then the part of the page with text from the other article is erased, so that the full-text pertains to articles as opposed to pages. Each article has already been OCRed at production time mostly with FineReader, and the full text is hidden in the delivered PDFs. It was harvested by EuDML and used for indexing.

For EuDML purposes, a license of the math-aware OCR software InftyReader was acquired by Université Joseph-Fourier, including a FineReader runtime to improve text-level multilingual recognition, as provided to us by Science Accessibility Net.

We set up a virtual Windows machine with Samba drive and Cygwin/SSH extensions linked to a Unix machine so that batch preparation, copy, OCR, and transformation could be chained easily.

The first OCR run was launched over the Bourbaki seminar collection, with Infty version 2.9.2. A big portion of this serial is typewritten, with handwritten symbols, which yielded a very low quality output and a high error rate. After discussion with Masakazu Suzuki, he told us that we could not expect good results with a source not typeset in a more traditional manner, so we excluded all typewritten seminars from our subsequent Infty runs, as well as the very old journals such as *Annales de gergonne* published in early 19th century with a very nice although not very regular typography.

As a number of problems with version 2.9.2 appeared, we installed Infty version 2.9.3 and cleaning scripts, with a new conversion table to produce much better TEX and HTML output for 20 journals (including Gallica's JMPA that lacked even text OCR), one seminar (Bourbaki) and one memoirs series. The final run used version 2.9.4. We have started to regenerate new TEX full texts with the new IMLConv software starting from Infty's KML files (raw OCR result with character candidates and coordinates) to end user formats TEX and HTML. The interest in doing so was that the generated TEX files had less errors than in previous Infty version, so it would be possible to get better quality TEX files without the time needed to OCR again TIFF files (it took somewhere between 10–20 seconds to OCR a standard monochrome 600 DPI page).

We ended up OCRing 31,056 articles, 610,000 pages. Each generated TEX file was run through a cleaning script, a conversion to XML/MathML based on **TeX2NLM**, then a cleaning script to remove benign errors (such as errors for "double superscripts" which are obviously out of scope of OCR, and of secondary importance to EuDML math formulae handling). XHTML was also run through a very simple script removing benign systematic errors (such as a few illegal UTF-8 characters, and some non-HTML elements or misspelled entities—most of these errors were corrected in Infty's conversion table and output driver in the course of this project).

This generated 28,946 HTML files, 31,274 XML files (papers of more than 100 pages generate multiple files).

We then implemented two XSL transformations: the first one starts from XHTML, concatenates files pertaining to a single item, and carries out some clean-up; the second one starts from XML generated from TEX (with benign errors removed), concatenates files pertaining to a single item, removes all text tagging and keeps math as MathML

(using the XML schema defined for indexing of full texts with formulae in the context of the EuDML project).

In this way 28,711 XHTML and 26,529 XML files were obtained. They were contributed to the EuDML central system by serving these files through an HTTP server and adding links to them in the OAI-PMH records of the corresponding items.

2.2.3 OCR at IST

The process to run Infty on the several collections started with the acquisition of Infty+FineReader by the IST. A Windows virtual machine was installed in a local server on which the Infty software was installed. This provided a test environment to run simple validation tests to tune the quality and output formats.

To avoid download delays all the PDF files were retrieved from each collection and stored locally.

A script in Java was developed to obtain the language information from the metadata record in REPOXfor each PDF file, in order to identify the correct language to be used as input for Infty (these languages can be: English, German, French, Italian, Dutch, Polish, Spanish, Slovak, Swedish, Czech). The output formats chosen by the EuDML committee were: LaTeX, XHTML with MathML markup, and KML.

This process allows each collection's full-text content to be processed by Infty. Also, the script creates a report file in XML that describes the duration of the process, start and end time, global state, number of pages and files processed, time in seconds per page, and all the errors that occurred during the process (with detailed information about the error and which record it occurred in).

Finally the results of Infty were stored in the file system of our server and exposed using the same script through a web API based on a REST architecture. Moreover, the developed script also allows a partial Inftyre-processing by comparing the PDFs that returned with error with the Infty results stored in the server. This feature allows the re-running of Inftyon the records that previously had errors, or missing records.

The results obtained for each collection made available by the data providers were:

- IMI-BAS/BulDML: 715;
- CSIC+USC/DML-E: 6,296;
- FIZ/ElibM: 32,814;
- BNF/JMPA: 2,081;
- BNP/PM 1,347;
- ICM/PL-DML 15,149;
- SIMAI/BDIM 2,138;
- SUBGoe 76,107 – work still in processing;
- IU/HDML 3,633[3]

3. The quality of the rest of the full-text content from IU/HDML collection is too low to be processed by Infty.

2.3 PDF Text Extractor

The **PDF Text Extractor** is written in Java, and uses the Apache PDFBox library [20] to obtain plain text from a PDF document. PDFBox is open source software distributed under Apache License Version 2.0.

For more detailed information we refer to D7.2 [24].

A joint demonstration of text extraction from PDF documents, bibliographic reference extraction from plain text, and bibliographic reference parsing is available at `https://project.eudml.org/EuDmlAnalysisDemo/`.

2.4 MaxTract

MathML, LaTeX and accessible format extraction from PDF documents are handled by a tool written in OCaml that returns the full page text of a PDF document in various formats with mathematical expressions embedded and marked up as MathML. The software requires an uncompressed PDF file which is created via a call to PDFTK, which is licensed under the GNU General Public License Version 2.

Given an appropriate PDF file, the full page content of the whole document can be extracted. The tool works by extracting the fonts and content streams from a PDF which are then parsed by the **MaxTract**, producing a list of symbols and graphics for each page. These, in conjunction with a list of glyphs obtained via image analysis of the page images rendered from the PDF document, are used to split each page into a number of lines. Each line is parsed to create a parse tree, then processed by a driver that separates text from in-line math expressions and produces MathML markup for any formulae that occur on that line. The software is based upon the work described in [2, 3].

A limitation of the tool is that it can only work with PDF files making use of Type 1 fonts and embedded font encodings. This generally means that the file will have been generated from LaTeX, Troff, Scientific Word or a number of other document production systems, which does limit the number of potential sources. Issues identified in D7.3 [27] concerning compatibility, freezing and segmentation have been addressed in the latest version.

The URL of the demonstration web site is `http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/maxtract.php`.

2.5 TeX2NLM

TeX2NLM is a tool to convert TeX code to MathML. It was designed primarily to be used by **EnhanceNLMTeXwMML**, which adds MathML versions of TeX formulae in JATS documents (see Section 2.6 for more details). A formula search function uses this tool by converting TeX entered by the website user to MathML on the fly, and then using the EuDML tool for Mathematical Indexing and Search (MIaS) [25, 26, 17] to build a search query from the MathML.

For more detailed information we refer to D7.3 [27].

A demonstration prototype of this tools is available at `http://eudml.mathdoc.fr/eudml-demo/TeX2NLM`.

2.6 EnhanceNLMTeXwMML

EnhanceNLMTeXwMML is a tool to convert TEX code to MathML within JATS documents, either by adding a <mml:math> alternative to existing <inline-formula> or <disp-formula> formula elements, or by creating that formula structure containing the original TEX and the converted MathML version.

For more detailed information we refer to D7.3 [27].

This tool has allowed us to generate 66,808 NLM formulae from 10,363 article records in EuDML collections. It will now be run extensively on all EuDML content.

An interactive version of the tool is available for demonstration at `http://eudml.mathdoc.fr/eudml-demo/EnhanceNLMTeXwMathML`.

2.7 Plain Text Reference Segmenter

The **Plain Text Reference Segmenter** described in [27] and demonstrated at `https://project.eudml.org/EuDmlAnalysisDemo/` was finally not integrated to the EuDML systems.

2.8 Bibliographic Reference Parser

Implementation of the reference parser mechanism described in [27] and demonstrated at `https://project.eudml.org/EuDmlAnalysisDemo/` was withdrawn from the final version of the EuDML systems, because sufficient accuracy has not been reached and miskakes, for which there is no manpower to fix, would make bad impression in EuDML user interface.

2.9 ZBMath Metadata Lookup Service

The **ZBMath Lookup Service** is a special online interface to the Zentralblatt MATH (ZBMath) database [8] which, given a Zentralblatt MATH item identification number (ZbId, an ZBMath 'AN' field value), returns the metadata that are stored in the ZBMath database for this publication.

For more detailed information we refer to D7.3 [27].

The tool is demonstrated at `http://www.zentralblatt-math.org/eudml/lookup?zblid=1163.57016`.

The Zentralblatt identifier in the URL can be replaced by any valid Zentralblatt identifier.

2.10 ZBMath Reference Matching Service

This reference matching online **ZBMath Match Service** ('Match') receives a string and searches for a matching record in the ZBMath database. If it finds a best-matching record it returns it as an XML formatted answer. The Match service will not find an appropriate ZbId if there is insufficient information in the query, as its goal is to find exactly one result.

For more detailed information we refer to D7.3 [27].

The tool is demonstrated at `http://www.zentralblatt-math.org/eudml/match?query=B24, Inhomogeneous fixed point ensembles, 1811`. The reference string in the URL can be replaced by other reference strings.

2.11 PdfJbIm

PdfJbIm is a PDF enhancer written in Java which reduces the size of PDF documents containing bitonal images [12]. It takes advantage of the extremely high compression ratio of visually lossless JBIG2 compression [7].

PdfJbIm uses the external open-source encoder jbig2enc [16], developed by Adam Langley. The compression ratio of jbig2enc was improved by about ten percent by creating an additional comparison process for distinguishing representatives of symbols. For more information about **PdfJbIm** and the modification of jbig2enc see [11, 23]

A possible improvement to Jbig2enc by integrating it with an OCR engine was discarded as the achieved improvement is low when configuration settings are chosen that maintain a quality sufficient to support reliable OCR. Without that reliable level of OCR, recognition errors could cause loss in quality when compressing images in the documents. It also greatly increases the time to process each document as OCR recognition takes so long to process.

A version without OCR was tested on journals stored at DML-CZ where the size of the PDF documents originally compressed using Fax G4 was reduced on average by thirty percent. For more details see D7.2 and [23]. A further description of the tool can be found in D7.3 [27].

A demonstration version of **PdfJbIm** is available at `http://nlp.fi.muni.cz/projekty/eudml/eudmldemo.php`.

2.12 Pdfsizeopt

Mainly due to high demand on system resources and low priority for inclusion of this tool in the EuDML system, the **pdfsizeopt** [21] tool described in [27] and demonstrated at `http://nlp.fi.muni.cz/projekty/eudml/eudmldemo.php` has not yet been integrated to the current version of the EuDML systems. This tool can be integrated in the future.

2.13 math_metadata_lookup

The **math_metadata_lookup** program [15] is a small open-source (GNU LGPL [10]) command-line utility implemented in Ruby. It searches through mathematical reviews databases and fetches metadata. Supported reviews databases are Mathematical Reviews [1] and Zentralblatt MATH [8]. This tool is described in more detail in [27, Section 2.14].

A demonstration version of the tool is available at `http://nlp.fi.muni.cz/projekty/eudml/eudmldemo.php`.

2.14 Metadata Editor with Export to NLM

The Metadata Editor, **ME**, [19, 4] is an open-source (GNU GPL [9]) client-server web application implemented mainly in Ruby and Perl programming languages. It is designed

to manage, edit and validate metadata and full texts of digital publications prior to their integration into a digital library. Description of the tool can be found in [27, Section 2.15].

The core of the NLM conversion process—a set of XSLT stylesheets—was expanded in functionality to cover all publication types and metadata elements used by **ME** mainly thanks to Vlastimil Krejčíř.

The Metadata Editor is a stand-alone application that enables EuDML data providers, who do not have their own solution, to organize and annotate their digitized publications and prepare their data for the EuDML project. IMI-BAS representatives went to Brno to prepare for use of the Metadata Editor on 20th December 2012.

3 Integration of Eutools into the EuDML Enhancement Workflow

After the individual testing of tools at the technology provider sites and consequently in the context of the EuDML system, selected tools mentioned in the previous section *were integrated* at the EuDML central site at ICM in Warsaw, with the exception of **PdfToText+MathViaOCR**, **math_metadata_lookup** and **ME**, which provide enhancements at the providers' sites.

3.1 Processing Nodes

Most tools described in previous section are so-called *processing nodes*, which can be chained together into so-called *processes*.

The initial node in a process typically generates or otherwise obtains chunks of data which are consecutively processed by the following nodes. A node typically enhances the chunks that it receives on its input and sends the enhanced chunks to its output, possibly with side effects such as indexing the contents of the chunk. The final node in a process typically stores the enhanced chunk in storage or discards it. There is a processing framework written in Java which orchestrates the flow of data chunks between nodes. Therefore, an author of an individual tool only needs to implement a processing node with well-defined inputs and outputs.

The WP7 enhancement tools are used in two places in the system. One workflow takes place during the ingestion phase as services invoked by REPOX—called the *ingestion workflow*. The second core *enhancement workflow* takes place iteratively over validated ingested metadata and PDFs for every EuDML publication. The tools are integrated one by one by adding them into both workflows described in following sections.

3.2 Ingestion Workflow

The ingestion workflow takes part in the REPOX 2.0 part of EuDML. Even though the primary goal of the metadata ingestion REPOX phase is checking and validation of input data, some enhancements already take place, namely **PdfToText+MathViaOCR** is called on EuDML items from some data providers not contributing rich full texts. The result of this workflow is validated metadata in EuDML NLM format. For details of this part we refer to deliverable [5, Section 10].

3.3 Enhancement Workflow

The enhancement workflow is shown in Figure 2, which indicates the main procedural steps taken on the latest versions of the data ingested, taking into account time stamps from the ingestion process and previous enhancement loops.

Full text processing is shown in detail in Figure 3 on the next page.

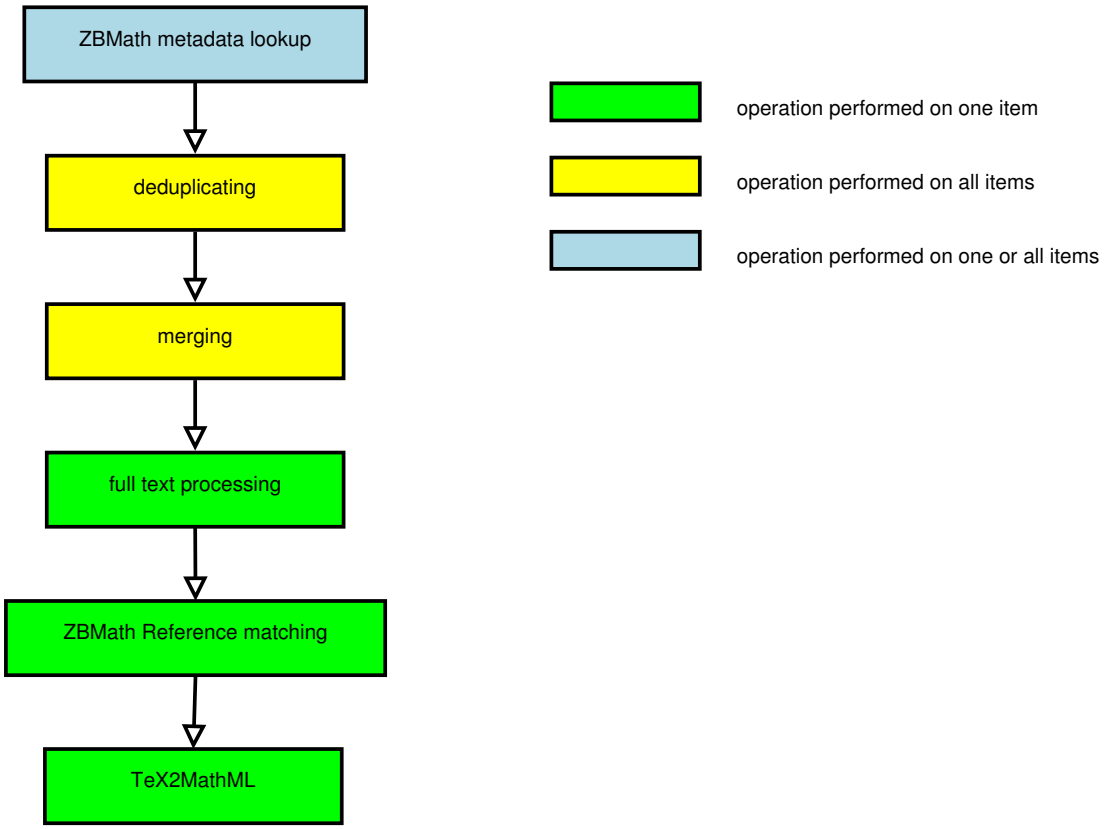

Figure 2: Enhancement workflow

1. Starting the enhancement process, ZBMath lookup is carried out. In the case that the Zentrallblat identifier is known, metadata from Zbl are fetched and added to those already collected from elsewhere.
2. Deduplication is performed, which identifies all duplicates not yet detected at Id assignment level.
3. Merging: records with multiple source records (including ZB math) are combined into single EuDML NLM records.

At this point records are in EuDML NLM format and are ready for enhancement. For some documents there is PDF with fulltext for the plain text content already, but not fulltext with mathematical formulae for indexing.

4. In the next steps, full texts are processed in the complex workflow shown in Figure 3 on the following page. At the time of writing 12,302 PDF files were processed by **MaxTract**.

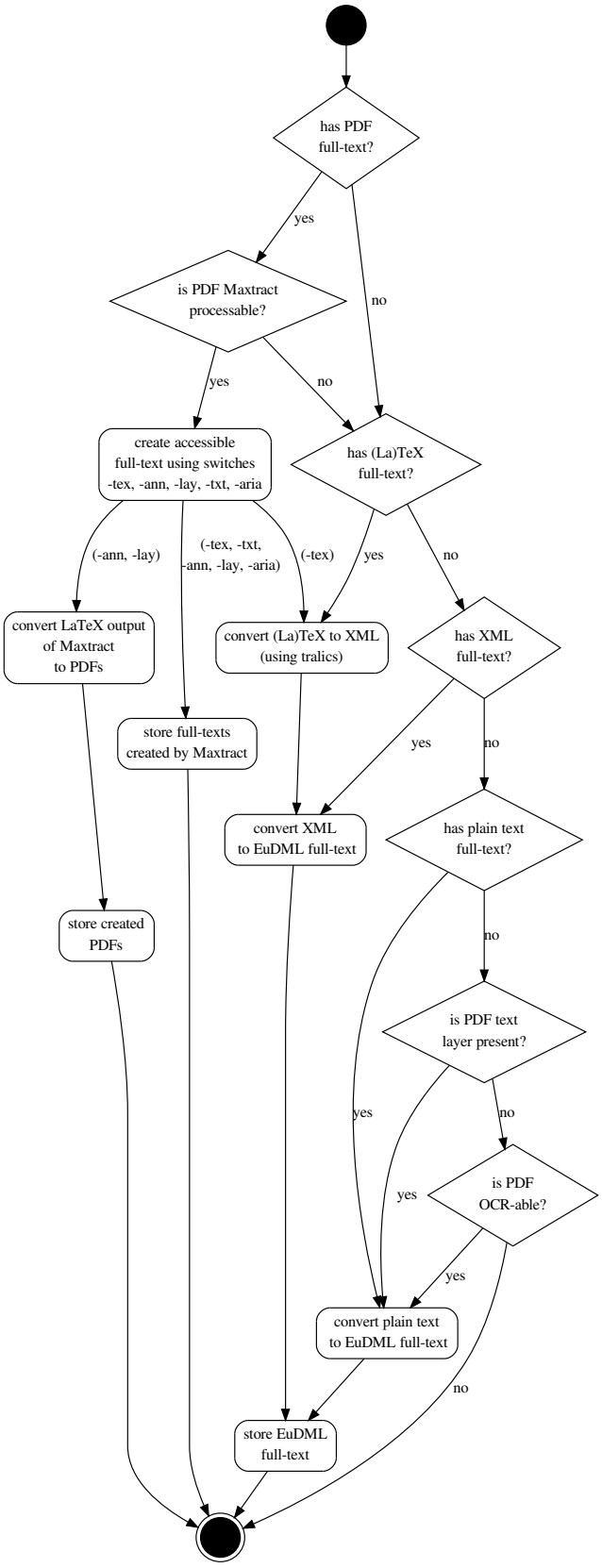

Figure 3: Full text processing workflow

5. Finally all interesting fields (abstract, title, plain text, references, etc.) will be processed by tool that converts TEX formulas to MathML format.

The result of this workflow is that fully enhanced metadata are present.

3.4 Enhancements by Matching

EuDML records not containing Zbl and MR IDs are enhanced by them as the principle input for the de-duplication algorithm. There are two tools available for this purpose—**ZBMath Reference Matching Service** (see 2.10) and **math_metadata_lookup** (see 2.13). **ZBMath Reference Matching Service** available at `http://www.zentralblatt-math.org/eudml/match?query=` is finally used for this purpose.

ZBMath Reference Matching Service supposes use of unstructured 'real-life' reference strings. But in the EuDML system, there is the possibility that we have articles described in NLM format but no 'real-life' citation of it. In order to use the **ZBMath Reference Matching Service**, we create the reference string in some way.

4 Deduplication and Merging

One document (article, book etc.) may be described by multiple metadata files, each delivered by different data providers. Taking this fact into account there is a need to deduplicate metadata and then create a representation of the duplicated document through merging within the EuDML system.

4.1 Deduplication

In the deduplication process, every document is represented by a *deduplication object* which has following attributes:

- publication title,
- publication date,
- identifiers set,
- authors data.

At the beginning every *deduplication object* is classified to a group based on publication year[4]. Afterwards every object is compared to the other objects in the same group and a similarity measure is calculated. If similarity measure is higher than a determined threshold, ranging from 0.0 (no equality) to 1.0 (full equality), then two documents are treated as a found duplicate. Calculating the similarity measure is performed in at most 4 steps (if any step fails further ones are not executed):

1. External identifier comparison. If the *deduplication objects* share identical identifiers (type and value) then full equality is returned, and further steps are omitted. If the *deduplication objects* have all identifiers of identical types but different values, service returns no equality. If no common identifier can be found or some of external identifiers differ—go to the next step.

4. This means that duplicates whose metadata have an error in their publication year would not be identified.

2. Compare publication year. If the difference is higher than one year then no equality is returned. Go to the next step otherwise[5].
3. Author names comparison. This is an implementation of the algorithm described in [6]. The idea behind this algorithm is to simplify author names to their initials and compare the resulting sets of letters. If the similarity measure in this step is below a fixed threshold then no equality is return. Go to the next step otherwise.
4. Publication title comparison. The proposed algorithm is similar to calculating Levenshtein distance, but is actually much more complex. The algorithm operates on words instead of letters. Each word has an assigned class and every class an importance level. For example "XXI" is classified as a "Roman Number", "January" is classified as a "English Month" and "Integral" is classified as a "Plain Text" word (other possible classes include "Special String", "Year Word", "Ordinal Word", and "Short Ordinal Word"). The assigned importance level for words influences the distance calculated between titles (for example similarity between "one." and "one" is high, and "XX" and "XXI" should be and is low).

4.2 Merging

Initially we assume that data providers are ordered with a priority inside EuDML. This is an arbitrary ordering based on heuristically evaluated quality of metadata provided. The current priority list is as follows: CEDRAM >EDPS >BDIM >DML-CZ>NUMDAM >PM >Bul-DML >DML-E >DML-PL >HDML >ElibM >GDZ.

According to this priority one metadata file is chosen as a *base metadata file*. During merging, missing information will be filled into this file.

The next step is to identify the ZBMATH identifier. If the `zbl-item-id` link is not present in the *base metadata file*, then it is searched for in the other metadata files (according to priority ordering). If found then it is added to the *base metadata file*. If not, the process tries to figure out the `zbl-item-id` by using the bibliographic reference matching service provided by ZBMATH. At the end of this step, if `zdl-item-id` has been found, the ZBMATH metadata is downloaded from the ZBMATH lookup service and stored internally.

After collecting the ZBMATH metadata, merging is executed as follows: First, the *Base metadata file* is filled with the ZBMATH data:

- journal volume (if missing),
- journal issue (if missing),
- pages (if missing),
- ISSN (if missing),
- MSC classification (always),
- keywords (always),
- language (if missing).

Next, other metadata are analyzed (again in priority order) and added:

- any id (always),
- trans-title (if missing for given language),

5. In case documents are preselected by publication year, this step always return equality, but this can be useful if the preselection method is changed.

- language (if missing),
- ext-link (always),
- self-uri (always),
- fpage (if missing),
- lpage (if missing),
- abstract (if missing),
- trans-abstract (if missing for given language),
- kwd-group (always),
- provider name (always),
- ref-list (if missing).

Note: After adding metadata from each provider (including ZBMATH), the *base metadata file* is updated, so adding a missing property is performed only once.

Deduplication and merging is also discussed in D4.4 [5, Section 9].

By the end of January 2013, the algorithms described in this section resulted in the detection of 3,617 pairs of duplicates, mostly in the journals Mathematica Bohemica Commentationes Mathematicae Universitatis Carolinae and Archivum Mathematicum. Deduplicateid item `https://eudml.org/doc/247421` has been created from `https://eudml.org/doc/118334` and `https://eudml.org/doc/21860`.

5 Summary, Conclusions

A summary of the eutools implemented is shown in Tables 2 and 3. Most enhancers are used internally in the main EuDML architecture, with the exception of **ME**, which is being offered to and used by potential partners (data providers) to edit the metadata to comply with EuDML data specification.

The tools were tested by the partner providing the tool, unit tested, and with the exception of **ME** and **math_metadata_lookup** they were integrated into the EuDML toolset merged into larger components and workflows and run on the central EuDML site on real data from providers.

Enhanced PDFs might be offered to partners via agreed interfaces. Otherwise they are used for EuDML internal purposes, mainly by WP7–WP10 toolsets, or exposed to EuDML users (e.g. accessibility enhancements created by **MaxTract** in WP10).

Eutools were merged into bigger components (one enhancer per subsystem) and tested on the central EuDML site on data collected from providers.

Eutools demonstrated in this demo and piped into workflows do deliver fully-fledged enhanced content for the EuDML system.

References

[1] American Mathematical Society. Mathematical Reviews, 2013. `http://www.ams.org/mr-database/`.

[2] Josef B. Baker, Alan P. Sexton, and Volker Sorge. Extracting Precise Data on the Mathematical Content of PDF Documents. In Petr Sojka, editor, *DML 2008: Proceedings of Towards*

Table 2: WP7 Eutools integration summary—OCR, Extraction, Conversion

Tool name	Input	Output	Main benefit for EuDML
		OCR subsystem	
PdfToTextViaOCR	PDF with images as input stream or file (content/<file name>)	Set of plaintext/text files	OCR-ed text suitable for indexing
PdfToText+MathViaOCR	PDF with images as input stream or file (content/<file name>)	Set of XHTML files with mathematics in MathML	OCR-ed text with MathML suitable for indexing
		Extraction subsystem	
PDFBox	Article PDF	Article fulltext	Extracts text from born-digital PDF (without using OCR)
MaxTract	Article PDF	List of MathML fragments for each formula within the file	Indexable, accessible MathML from a standard PDF
		Conversion	
TeX2NLM	UTF-8 encoded character string	Java DOM Element with TEX formulas converted to NLM structure with TEX and MathML alternatives	Upgrades untagged TEX formulas in a character string to full TEX+MathML NLM structure. Enables MathML based knowledge managment and information retrieval
EnhanceNLMTeXwMML	Java DOM Document	Java DOM Document with all recognized TEX formulas represented by the NLM formula structure, containing both TEX and MathML version.	Enhances TEX formulas in NLM documents, either already in a NLM formula structure or not, to all be in such a structure and have a MathML alternative.

Table 3: WP7 Eutools integration summary—Refinement and External

Refinement subsystem			
ZBMath Lookup Service	Zbl identifier	Metadata in Zbl's XML	Provides checked and rich article metadata given ZblId
ZBMath Match Service	bibliography reference string	Metadata in Zbl's XML	Provides checked and rich article metadata given reference string
PdfJbIm	PDF as file or input stream	Re-compressed PDF as file or in output stream	Reduces size of PDF files containing images by about 30%
math_metadata_lookup	Search options (title/author/year plain text strings)	Metadata found in mathematical reviews databases in desired format (plain text/ HTML/XML/JSON/Ruby/YAML)	Stand-alone tool able to search through mathematical reviews databases and fetch metadata according to given search options
External/standalone tools			
ME	Unorganized collection of scanned pages of documents	PDFs of digital publications organized to collections of journals/volumes/articles etc. with metadata description	Stand-alone tool for organization and management of digitized publications able to prepare EuDML-ready full texts and metadata description that enables data providers without their own solution to participate in the EuDML project

Digital Mathematics Library, pages 75–79, Birmingham, UK, July 2008. Masaryk University. http://dml.cz/handle/10338.dmlcz/702535.

[3] 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. Bertinoro, Italy, July 20–21st, 2011*, pages 65–75. Masaryk University, July 2011. http://hdl.handle.net/10338.dmlcz/702603.

[4] Miroslav Bartošek, Petr Kovář, Martin Šárfy, and Michal Růžička. Metadata Editor, 2010. http://is.muni.cz/publication/927548?lang=en.

[5] José Borbinha, Aleksander Nowiński, Gilberto Pedrosa, Krzysztof Wojciechowski, and José Delgado. The EuDML Information Life Cycle Process, May 2012. Deliverable D4.4, rev. 1.0, of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[6] Eduardo N. Borges, Moisés G. de Carvalho, Renata Galante, Marcos André Gonçalves, and Alberto H. F. Laender. An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Inf. Process. Manage.*, 47(5):706–718, September 2011.

[7] JBIG Committee. *14492 FCD*. ISO/IEC JTC 1/SC 29/WG 1, 1999. http://www.jpeg.org/public/fcd14492.pdf.

[8] FIZ Karlsruhe. Zentralblatt MATH – ZBMATH Online Database, 2013. http://www.zentralblatt-math.org/zmath/.

[9] Free Software Foundation. GNU General Public License, 2013. http://www.gnu.org/licenses/gpl.html.

[10] Free Software Foundation. GNU Lesser General Public License, 2013. http://www.gnu.org/licenses/lgpl.html.

[11] Radim Hatlapatka and Petr Sojka. PDF Enhancements Tools for a Digital Library: pdfJbIm and pdfsign. In Petr Sojka, editor, *Proceedings of DML 2010*, pages 45–55, Paris, France, July 2010. Masaryk University. http://is.muni.cz/publication/891674/.

[12] Radim Hatlapatka and Petr Sojka. Recompression of Bitmaps in PDF using JBIG2 format, December 2010. http://is.muni.cz/publication/927601?lang=en.

[13] Google HP. Tesseract. http://code.google.com/p/tesseract-ocr/.

[14] Toshihiro Kanahori and Masakazu Suzuki. Refinement of digitized documents through recognition of mathematical formulae. In *Proceedings of the 2nd International Workshop on Document Image Analysis for libraries*, pages 95–104, Lyon, France, April 2006.

[15] Petr Kovář. math_metadata_lookup, 2013. https://github.com/pejuko/math_metadata_lookup.

[16] Adam Langley. Homepage of jbig2enc encoder. [online], 2013. http://github.com/agl/jbig2enc.

[17] Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec. Web Interface and Collection for Mathematical Retrieval. In Petr Sojka and Thierry Bouche, editors, *Proceedings of DML 2011*, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University. http://hdl.handle.net/10338.dmlcz/702604.

[18] Masakazu Suzuki et al. Infty Project, 2013. http://www.inftyproject.org/.

[19] Digitization Metadata Editor, version 2.0 (Jan 30, 2012), January 2012. http://dme.svn.sourceforge.net/.

[20] PDFBox.org. PDFBox. http://www.pdfbox.org/.

[21] Peter Szabó. Pdfsizeopt. http://code.google.com/p/pdfsizeopt/.

[22] Petr Sojka, Josef Baker, Alan Sexton, and Volker Sorge. A State of the Art Report on Augmenting Metadata Techniques and Technology, December 2010. Deliverable D7.1

of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.eu/sites/default/files/D7.1-v1.2.pdf.

[23] Petr Sojka and Radim Hatlapatka. Document Engineering for a Digital Library: PDF recompression using JBIG2 and other optimization of PDF documents. In *Proceedings of the ACM Conference on Document Engineering, DocEng 2010*, pages 3–12, Manchester, September 2010. Association of Computing Machinery. http://portal.acm.org/citation.cfm?id=1860563.

[24] Petr Sojka and Radim Hatlapatka. Toolset for Image and Text Processing and Metadata Editing – Initial Release, February 2011. Deliverable D7.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.eu/sites/default/files/D7.2-v1.0.pdf.

[25] Petr Sojka and Martin Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, and Florian Rabe, editors, *Intelligent Computer Mathematics. Proceedings of 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011*, volume 6824 of *Lecture Notes in Artificial Intelligence, LNAI*, pages 228–243, Berlin, Germany, July 2011. Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-22673-1_16.

[26] Petr Sojka and Martin Líška. The Art of Mathematics Retrieval. In *Proceedings of the ACM Conference on Document Engineering, DocEng 2011*, pages 57–60, Mountain View, CA, September 2011. Association of Computing Machinery. http://doi.acm.org/10.1145/2034691.2034703.

[27] Petr Sojka, Krzysztof Wojciechowski, Nicolas Houillon, Michal Růžička, and Radim Hatlapatka. Toolset for Image and Text Processing and Metadata Enhancements – Value Release, March 2012. Deliverable D7.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, https://project.eudml.org/sites/default/files/D7.3.pdf.

[28] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori. INFTY — An integrated OCR system for mathematical documents. In C. Vanoirbeek, C. Roisin, and E. Munson, editors, *Proceedings of ACM Symposium on Document Engineering 2003*, pages 95–104, Grenoble, France, 2003. ACM.

Index