Typesetting^{-1}

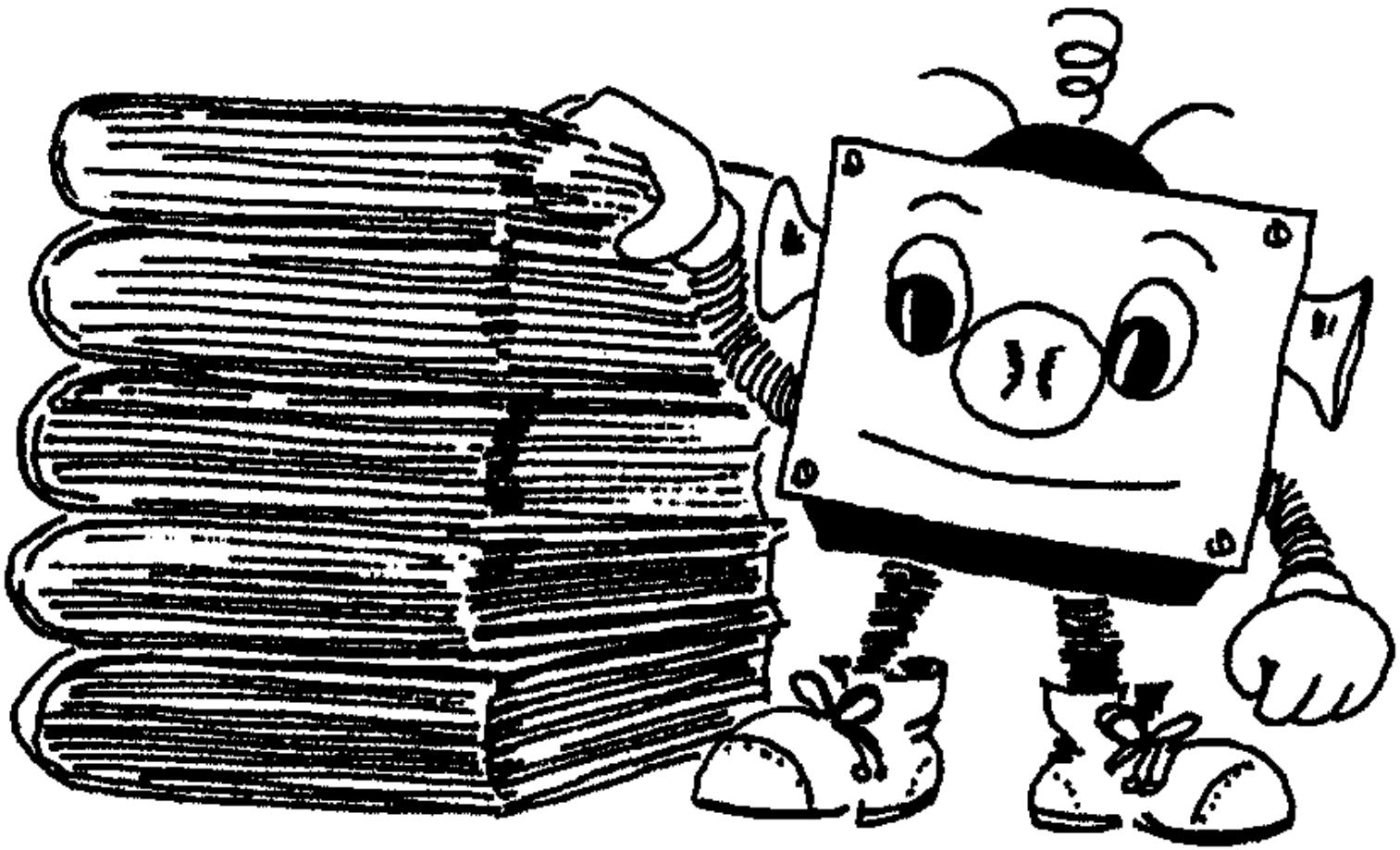
Petr Sojka

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<sojka@fi.muni.cz>

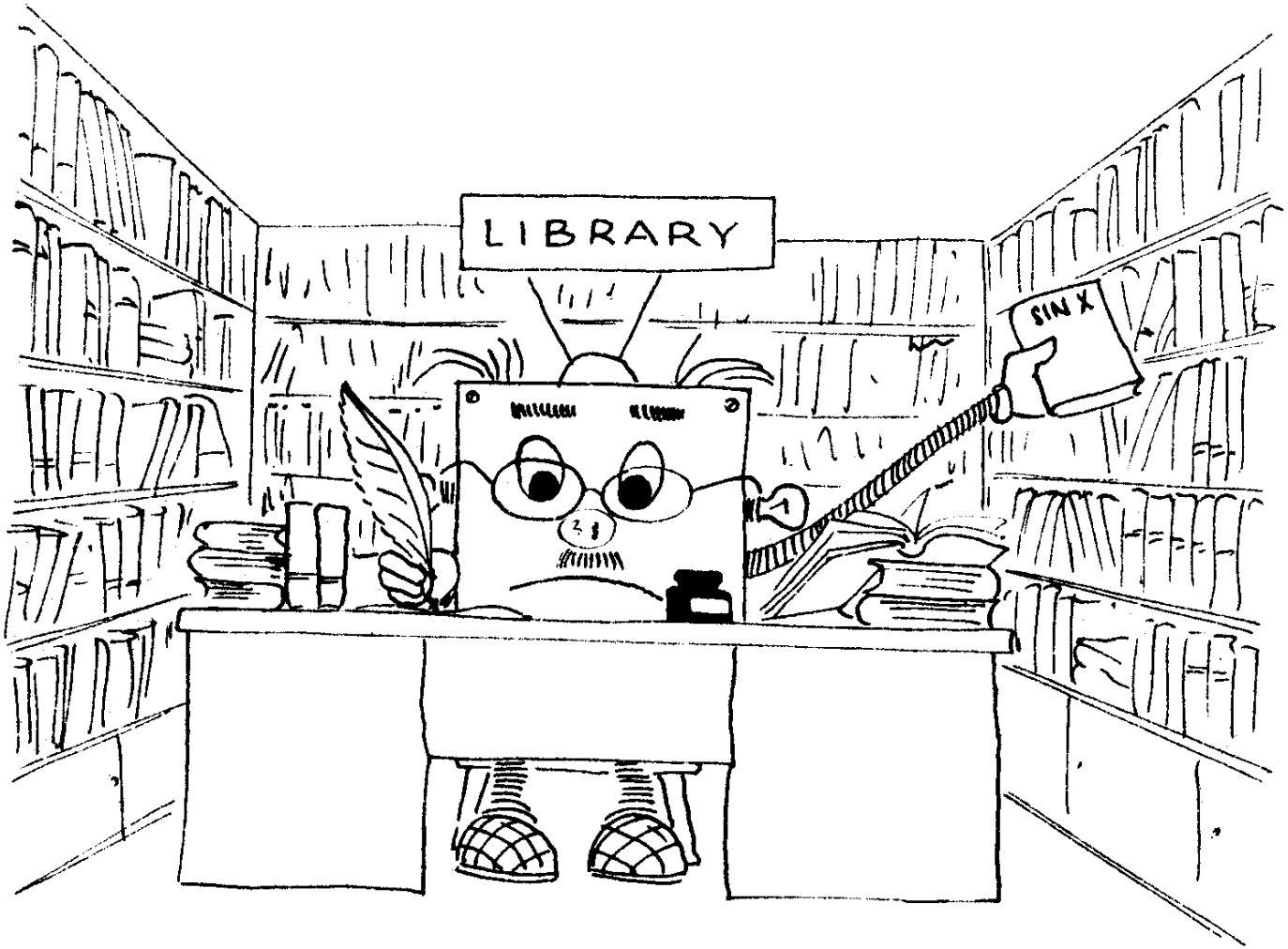
TEXperience 2013, Brejlov, Czech Republic
September 28th, 2013

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

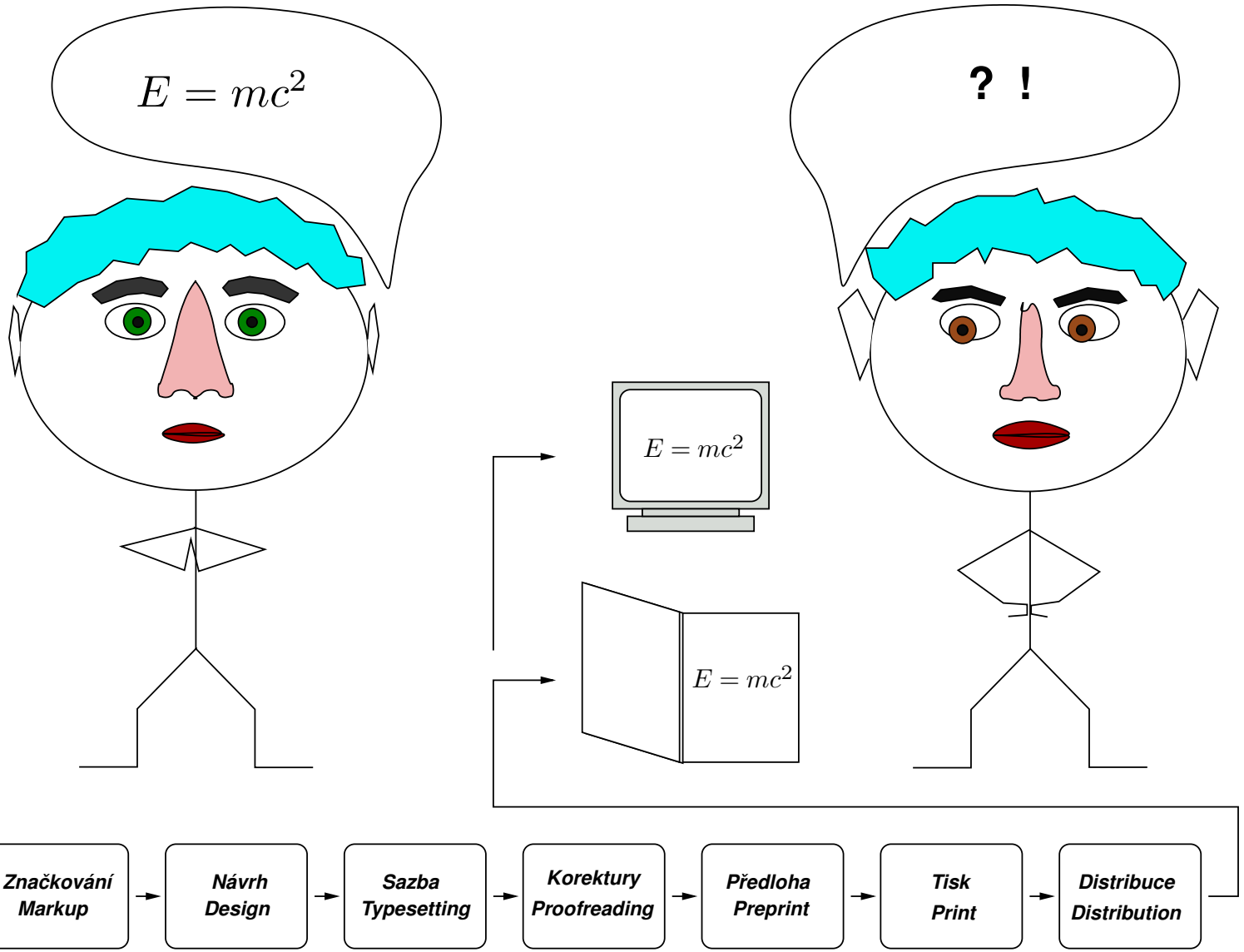
We typeset books, often scientific [with math]



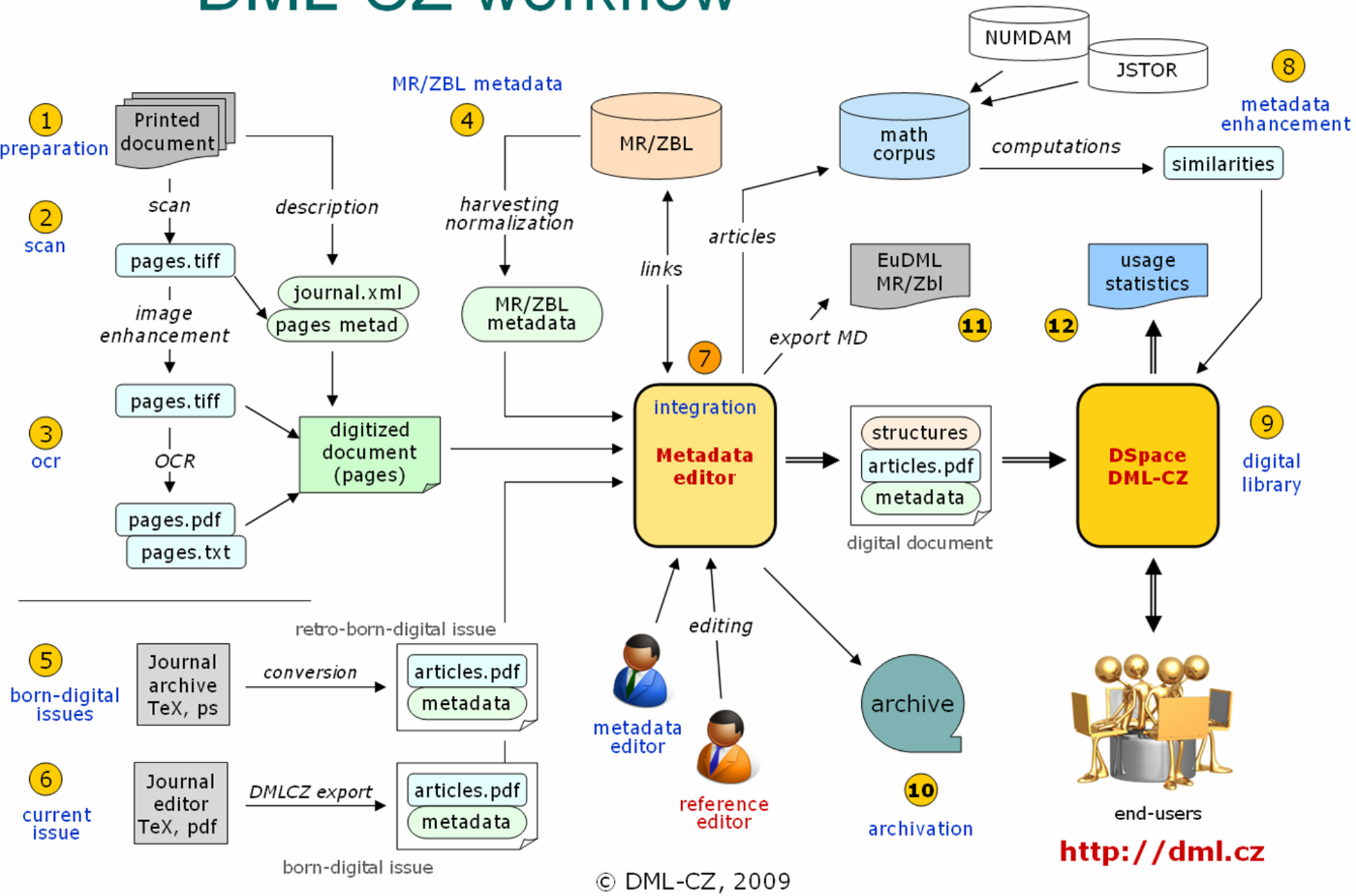
that end up in DML as The European Digital Mathematics Library: *EuDML*



Typesetting as part of pipe from the author to the reader



DML-CZ workflow



Take care! “God is in the details.” (Mies van der Rohe)

Data heterogenity, specificity: no free lunch to unify

Proof. Let $\hat{K}$ be a cube, $\hat{K} \subset \hat{G}$; put $K = \varphi^{-1}(\hat{K})$. According to theorem 50 we have $K \in \mathfrak{A}$ and it follows from theorem 24 that

$$P(K, v) = \int_K f(x)\,\mathrm{d}x\,. \tag{89}$$

The functional determinant T of the mapping $\psi = \varphi^{-1}$ fulfils the relation $T(\varphi(x)) \,.\, \det M(x) = 1$, so that

$$\int_K f(x)\,\mathrm{d}x = \int_{\hat{K}} f(\psi(y)) \,.\, |T(y)|\,\mathrm{d}y = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y\,. \tag{90}$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y$, which completes the proof.

Remark. The reader may compare this paper with [6].

REFERENCES

[1] *V. Jarník:* Diferenciální počet, Praha 1953.
[2] *V. Jarník:* Integrální počet II, Praha 1955.
[3] *J. Mařík:* Vrcholy jednotkové koule v prostoru funkcionál na daném polouspořádaném prostoru, Časopis pro pěst. mat., *79* (1954), 3—40.
[4] *Ян Маржик* (Jan Mařík): Представление функционала в виде интеграла, Чехословацкий мат. журнал, *5* (80), 1955, 467—487.
[5] *J. Mařík:* Plošný integrál, Časopis pro pěst. mat., *81* (1956), 79—82.
[6] *Ян Маржик* (Jan Mařík): Заметка к теории поверхностного интеграла, Чехословацкий мат. журнал, *6* (81), 1956, 387—400.
[7] *S. Saks:* Theory of the integral, New York.

Резюме

ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЖИК (Jan Mařík), Прага.

(Поступило в редакцию 10/X 1955 г.)

Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^{m} v_i^2(x) \leqq 1$ для всех $x \in A$. Пусть $\mathfrak{A}$ — система всех ограниченных измеримых множеств A, для которых $\|A\| < \infty$. Теорема 18 тогда утверждает: *Пусть $A \in \mathfrak{A}$; пусть D — граница множества A. Тогда на системе $\mathfrak{B}$ всех борелевских подмножеств множества D существует мера p и на*

ИОСИФ ВИССАРИОНОВИЧ СТАЛИН
1879—1953

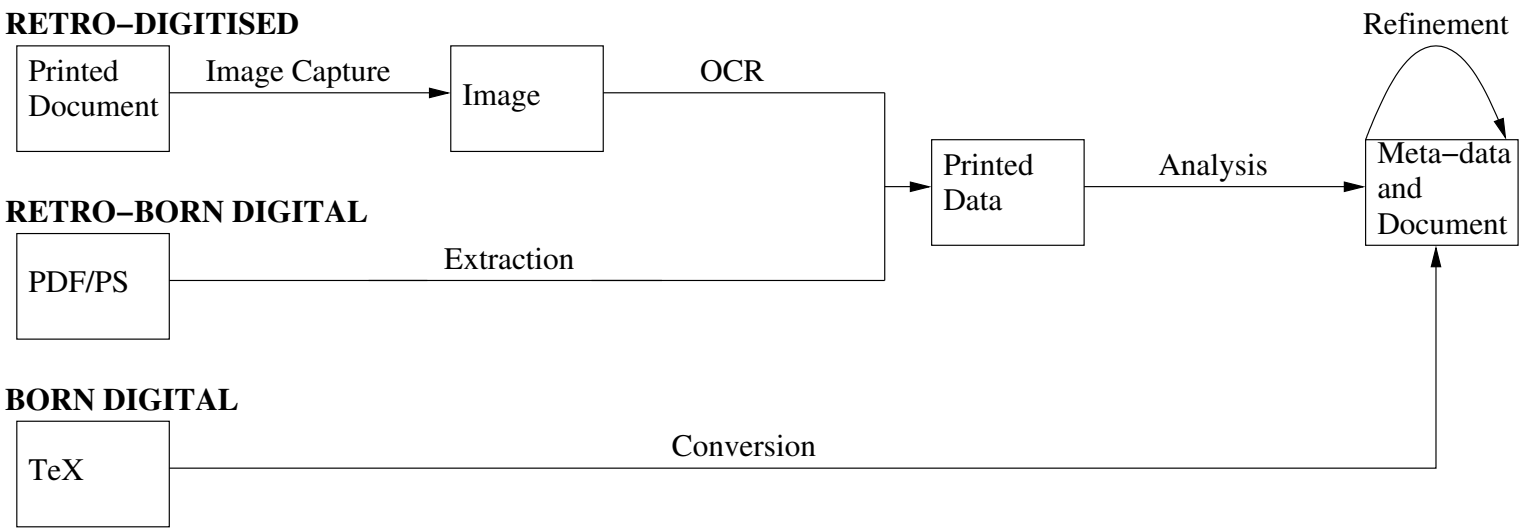
Document accessibility 4 DML processing challenges

Conversions (inversion of authoring+typesetting) needed from:

born-digital period: typesetting by TEX with export of [meta]data into digital library: maxTract

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), *two-layer PDF*

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution: finally Tesseract



From PDF to MathML (via LaTeX)

Most fulltext available as PDF only, often as low quality scanned volume pages. Aggregation via IP protected OAI-PMH, including the PDFs behing moving wall.

Workflow based in the case of:

born-digital PDFs: on maxTract, otherwise on PDFBox (plain text);

bitmap PDFs: on Infty, otherwise on Tesseract (no math).

Infty from Fukuoka

Run in parallel in Brno, Grenoble and Lisbon to speed up. Almost 200K papers (more than 1M pages).

Working with prof. Suzuki to improve further (automation, support for Russian, LaTeX driver,…).

Automated only, no time (and money) to fix OCR errors.

MathML output used for [internal] indexing and similarity computations only, not for metadata or export.

maxTract from Birmingham

```
\left(
\sum ^{ m }_{ i = 0 } a _{ i } x ^{ i }
\right)
```

$$r(x) = \sum_{i=0}^{p} c_i x^i.$$

$$[p(x)q(x)]r(x) = \left[\left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{i=0}^{n} b_i x^i\right)\right]\left(\sum_{i=0}^{p} c_i x^i\right)$$

$$= \left[\sum_{i=0}^{m+n}\left(\sum_{j=0}^{i} a_j b_{i-j}\right) x^i\right]\left(\sum_{i=0}^{p} c_i x^i\right)$$

```
open parenthesis
sum from i = zero to m of
a sub i x to the power of i
closing parenthesis
```

```
<math
  xmlns='http://www.w3.org/1998/Math/MathML'
  <mo>(</mo>
  <munderover>
    <mo>&Sum;</mo>
    <mrow>
      <mi>i</mi>
      <mo>=</mo>
      <mn>0</mn>
    </mrow>
    <mi>m</mi>
  </munderover>
  <msub>
    <mi>a</mi>
    <mi>i</mi>
  </msub>
  <msup>
    <mi>x</mi>
    <mi>i</mi>
  </msup>
  <mo>)</mo>
</math>
```

maxTract from Birmingham II: adding accessibility

Adding accessibility to mathematical documents on multiple levels:

- access to content for print impaired users, such as those with visual impairments, dyslexia or dyspraxia
- output compatible with web browsers, screen readers and tools such as copy and paste, which is achieved by enriching the regular text with mathematical markup. The output can also be used directly, within the limits of the presentation MathML produced, as machine readable mathematical input to software systems such as Mathematica or Maple.

On EuDML 10k+ fulltexts are served, mostly for reading in Chrome (HTML5 output) and/or Adobe Acrobat Reader (as multiple-layer PDFs, [no tagged PDFs yet]).

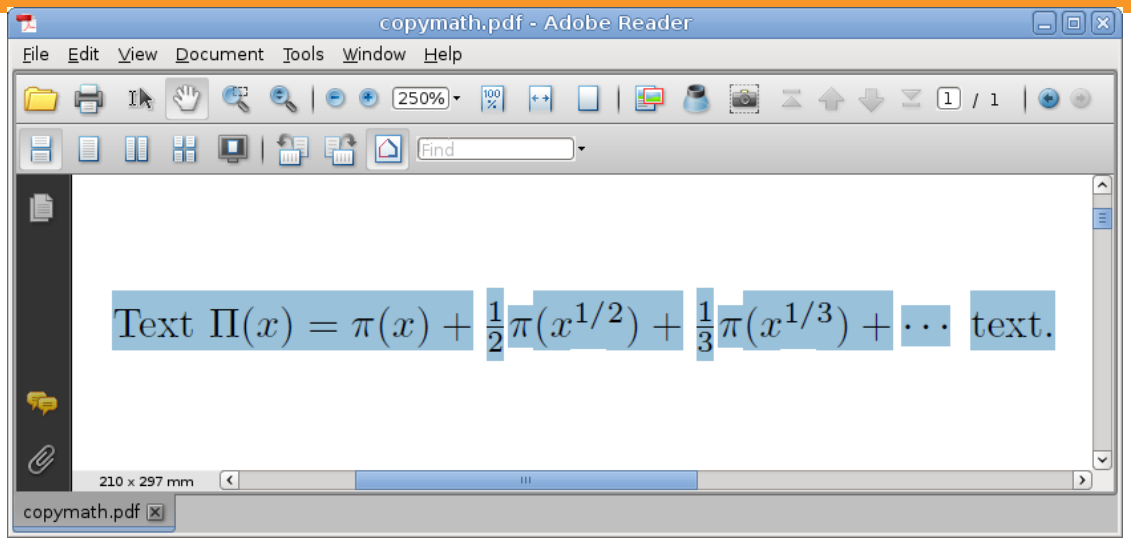
DML-CZ challenges and lessons learned

DML-CZ, the Czech Digital Mathematics Library, now serves more than *300,000 pages of more than 30,000 math papers*. Challenges were

- *migration of existing workflows (retro-digital, retro-digital and born-digital) into the repository*
- negotiations with Google Scholar towards better visibility
- math indexing and search
- ....

DML-CZ is according to The Ranking Web of World Repositories the best repository in CZ, 91. in EU and 203. in the world.

Math from standard PDF document



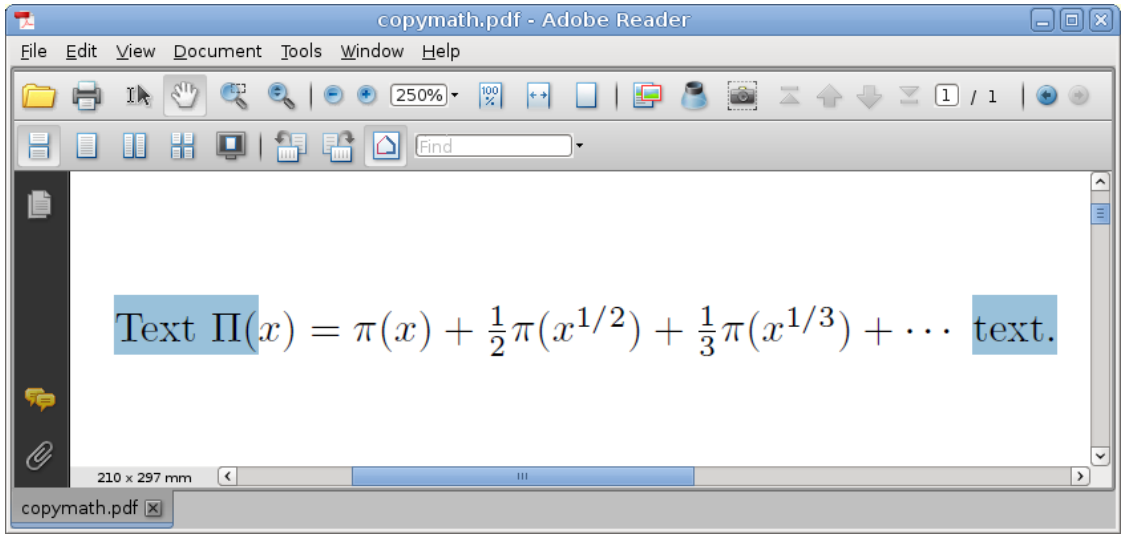
LaTeX source code:

```
Text $\Pi(x) = \pi(x) +
\frac{1}{2}\pi(x^{1/2}) +
\frac{1}{3}\pi(x^{1/3}) + \cdots$
text.
```

PDF code:

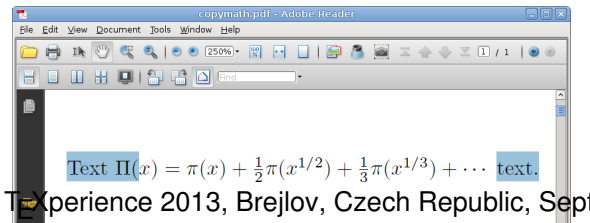
```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ/F17 9.9626 Tf 23.247 0 Td
[(\005\050)]TJ/F20 9.9626 Tf 11.346 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[\051=278(+)]TJ/F20 9.9626 Tf 17.158 0 Td [(\031)]TJ/F17 9.9626 Tf 6.036 0 Td
[(\050)]TJ/F20 9.9626 Tf 3.879 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
```

copymath-enabled PDF document



LaTeX source code:

```
Text $\Pi(x) = \pi(x) +
\frac{1}{2}\pi(x^{1/2}) +
\frac{1}{3}\pi(x^{1/3}) + \cdots$
text.
```



Implementation

- The `\ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not straightforward and requires nonstandard modifications of the LaTeX mathematical environments.

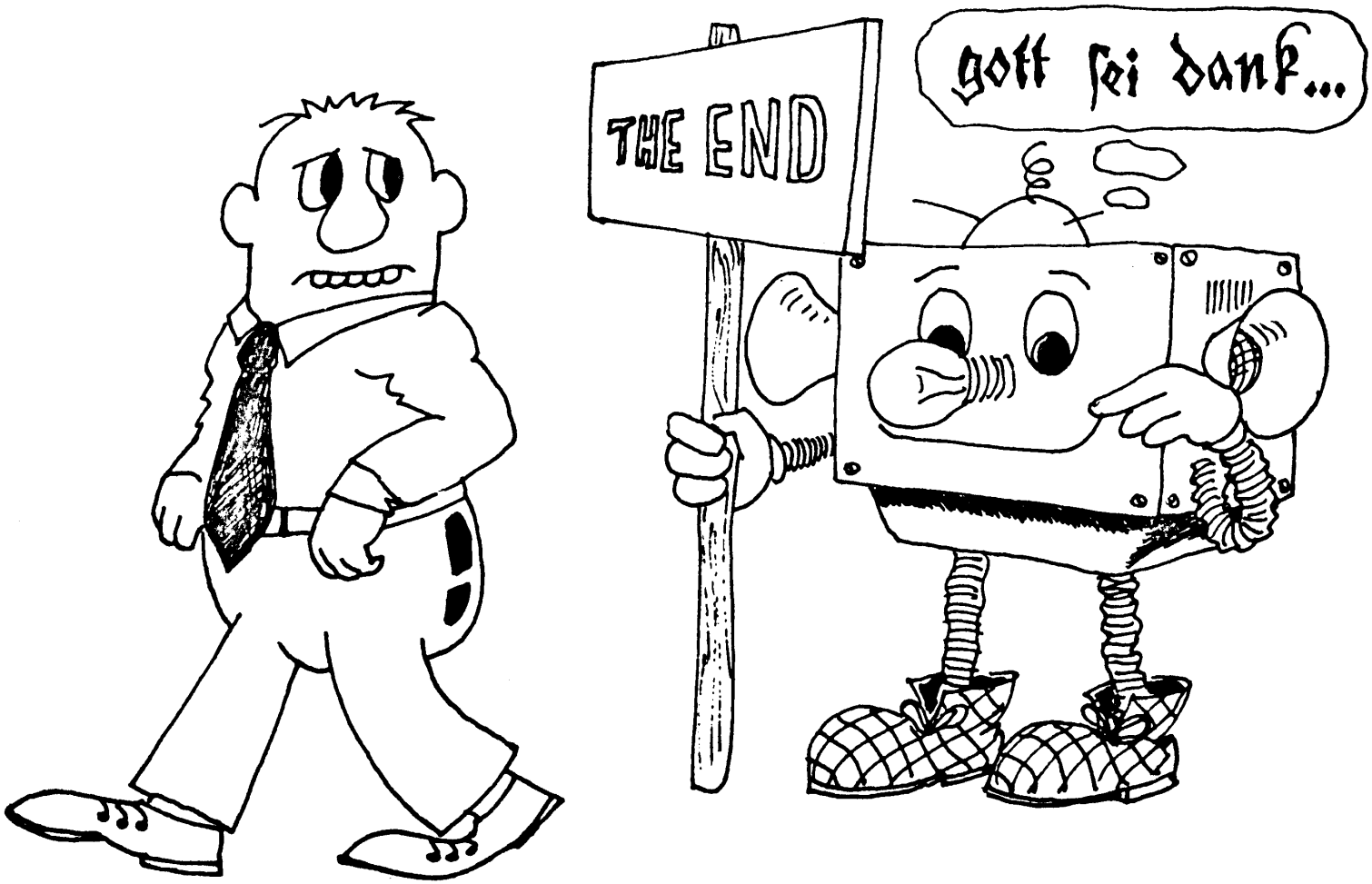
Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- It seems that not all PDF viewers respect contents of the `\ActualText` command.
- Adobe Reader ignores the "_" sign inside `\ActualText` provided another character is present.
- Possibility to be misused.

Future work

- Robust Math OCR is necessary
- Robust born-digital PDF2Math conversion is needed as well
- only then challenges as: multilingual math retrieval, MathML indexing and search, math common sense, mathematical document classification, document similarity could be possible

That's it!




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


Grimm, J.: Producing MathML with Tralics. In: Sojka [4], pp. 105–117, <http://dml.cz/dmlcz/702579>


MREC – Mathematical REtrieval Collection, <http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html>


Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <http://www.fi.muni.cz/ sojka/dml-2010-program.html>


Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Urban, J., Rabe, F., (eds.) Proceedings of CICM Conference 2011 (Calculemus/MKM). Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 228–243. Springer-Verlag, Berlin, Germany (Jul 2011), <http://dx.doi.org/10.1007/978-3-642-22673-1_16>


Líška, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math T ask. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task. pp. 686-691. NII, Tokyo, 2013. PDF


D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pp. 91–103, Aachen, 2012.


Sojka, Petr and Martin Líška. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>


Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: MathML-aware Article Conversion from LaTeX. In: Sojka, P. (ed.) Proceedings of DML 2009. pp. 109–120. Masaryk University, Grand Bend, Ontario, CA (Jul 2009), <http://dml.cz/dmlcz/702561>


Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science 3, 299–307 (2010), <http://dx.doi.org/10.1007/s11786-010-0024-7>


Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [4], pp. 11–24, <http://dml.cz/dmlcz/702569>


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://dml.cz/dmlcz/702604>.


Credits for LDA pictures goes to David M. Blei.


Credits for illustrations goes to Jiří Franek.