Search for Maths

Michal Růžička

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<mruzicka@mail.muni.cz>

Practise for DTEDI Thesis proposal, FI MU, Brno, Czech Republic

4th April 2012

Outline

Motivation for Search for Maths (DML panel discussion)

Q: ‘What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?’

A: **‘Math formulae search.’**

Prof. James Davenport, CEIC member, MKM 2011 PC chair, on panel at DML 2011 workshop in Bertinoro as a reply

Existing Systems – Pros and Cons

- **EgoMath** and **EgoMath2**: based on full text web search system Egothor * presentation MathML for indexing * idea of formulae augmentation, α-equivalence algorithms and relevance calculation
- **MathDex**: formerly MathFind * seven digit figure NSF grant by Design Science (Robert Miner) * Lucene based, indexing n-grams of presentation MathML * pioneering conversion effort
- **LaTeXSearch**: MSE offered by Springer * closed source * only for LaTeX math string approximate match based on strings * no formulae structure matching * small database: 3 M formulae from 'random' sources (cf. 200 M in arXiv)
- **LeActiveMath**: indexing string tokens from OMDoc with OpenMath semantic notation * *only* for documents authored for LeActiveMath learning environment
- **DLMF**: *only* for documents authored for DLMF in special markup * equation search
- **MathWeb Search**: semantic approach – uses substitution trees – not based on full text searching * supports Content MathML and OpenMath * problem with acquiring semantic data

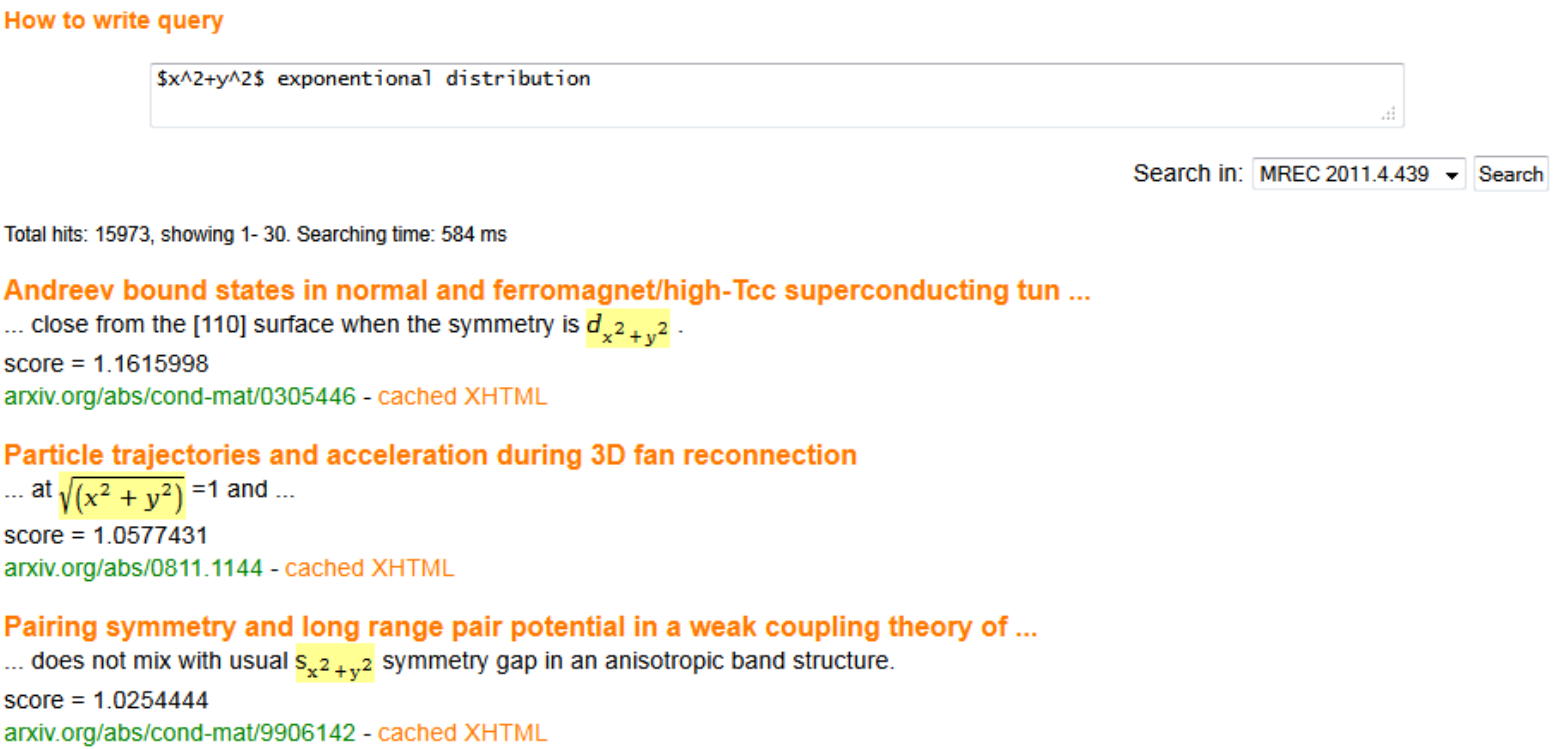
Math Search

Math Search

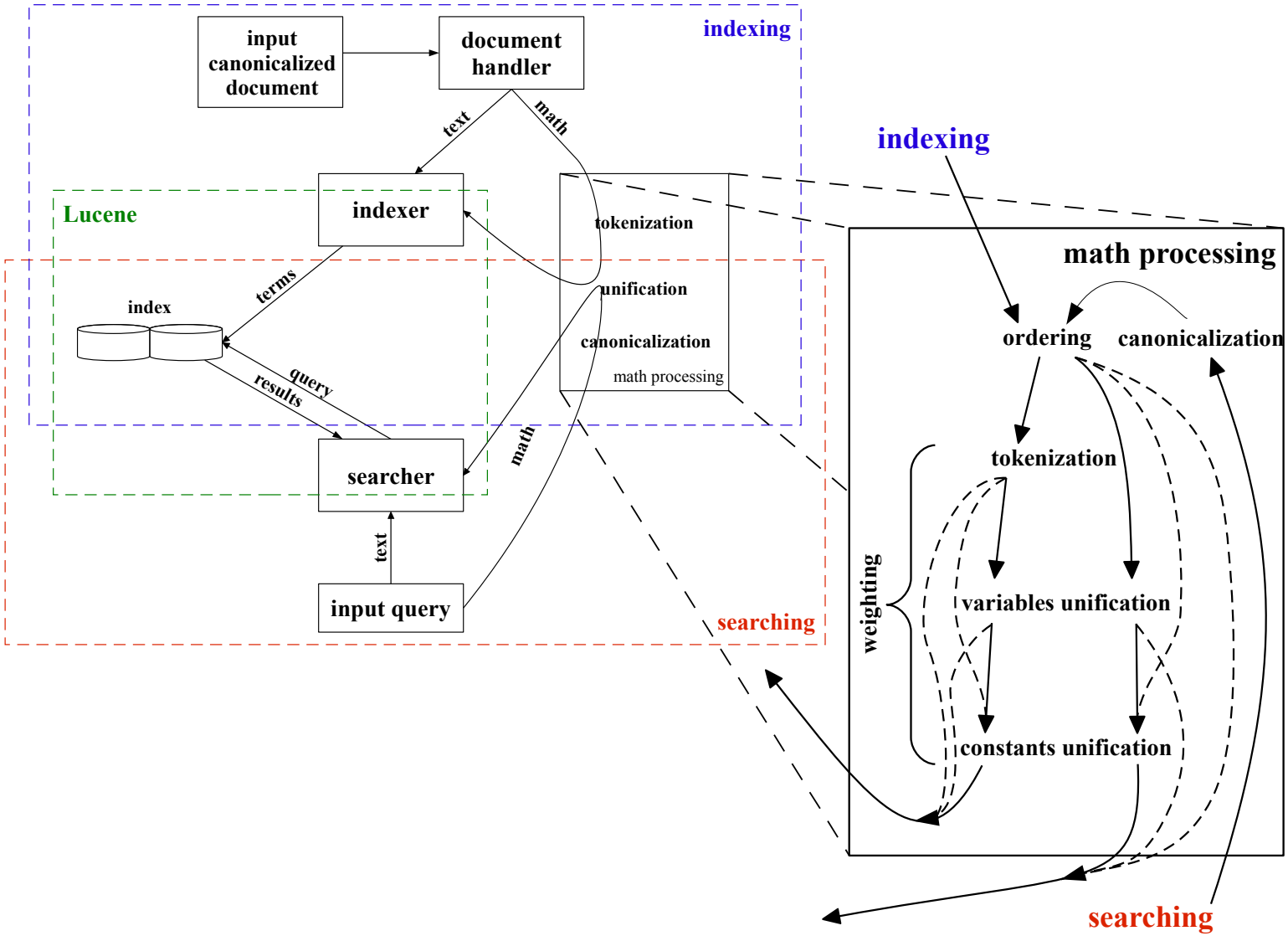
- Searching is crucial part of accessibility of scientific documents.
- Text and keyword based search seems to be well established.
- Mathematics formulae search seems to be problem.

MIaS – Math Indexer and Searcher

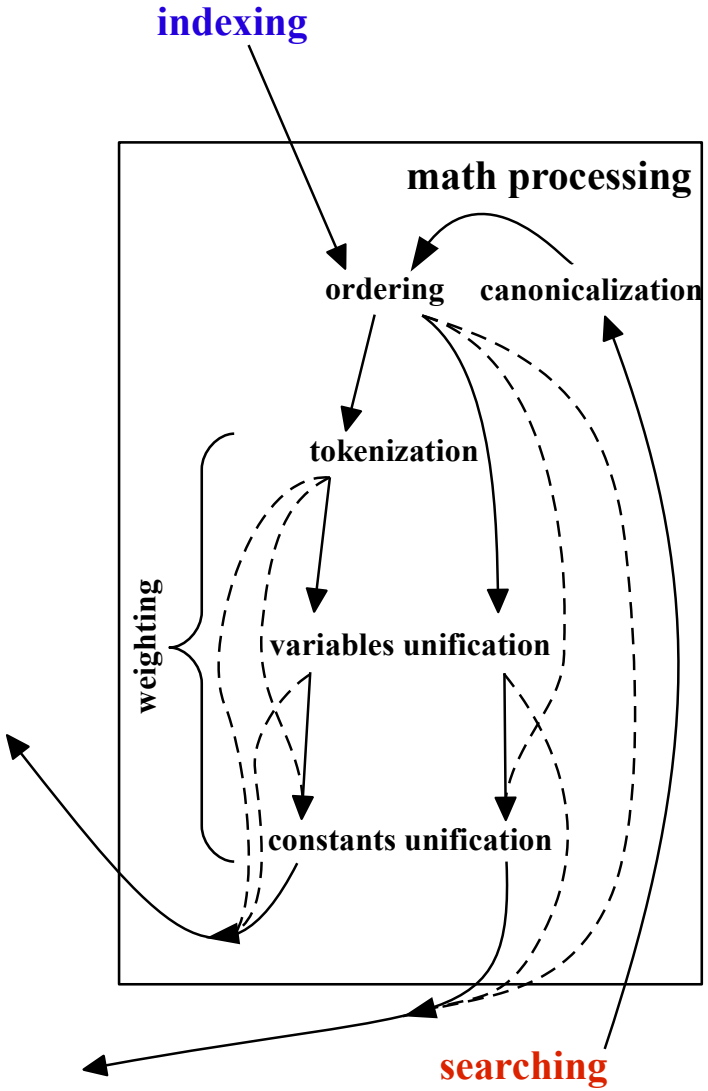
- Math-aware, full-text based search engine.
- Joins textual and mathematical querying.
- MathML or TEX input.

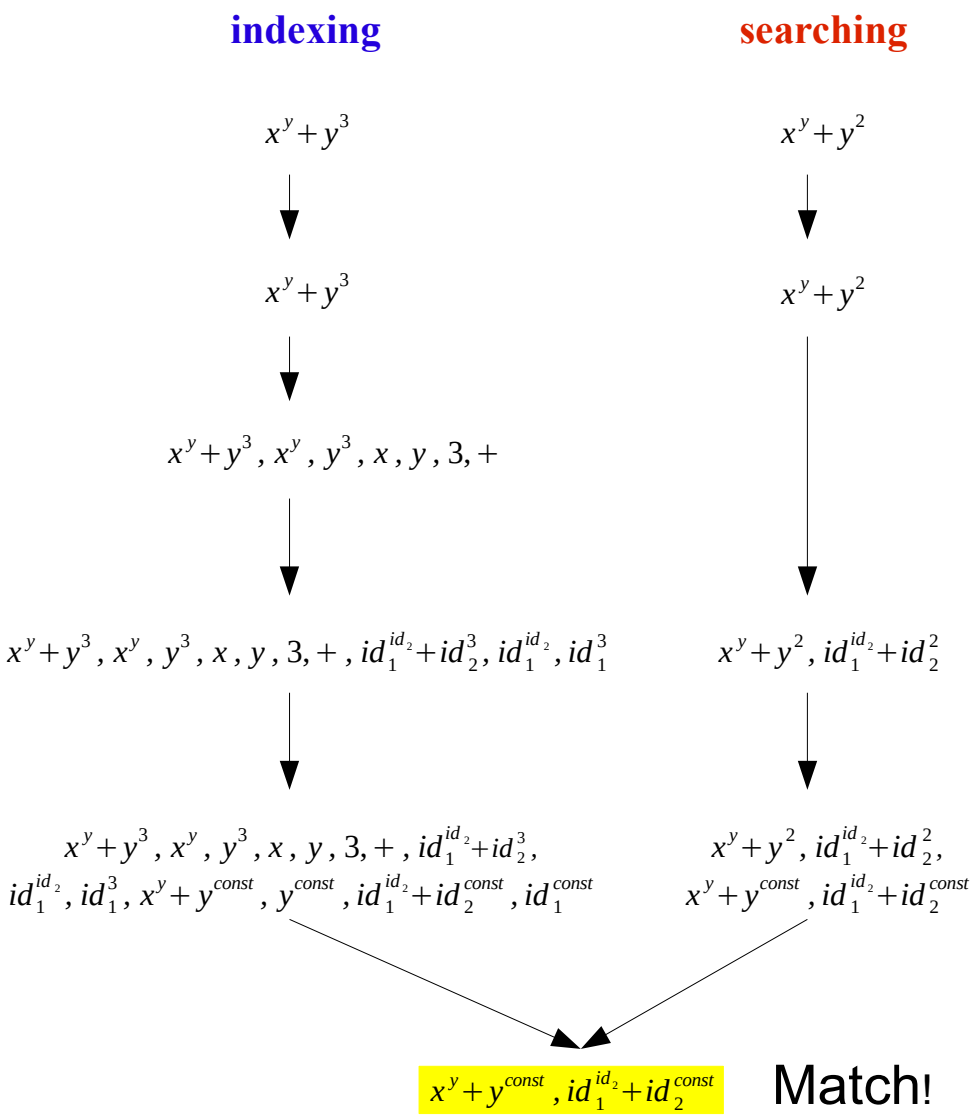


Math indexing design



Example


indexing
math processing
ordering
canonicalization
tokenization
weighting
variables unification
constants unification
searching


indexing
searching
x^y+y^3
x^y+y^3
$x^y+y^3\,,\,x^y\,,\,y^3\,,\,x\,,\,y\,,\,3,+$
$x^y+y^3\,,\,x^y\,,\,y^3\,,\,x\,,\,y\,,\,3,+\,,\,id_1^{id_2}+id_2^3,\,id_1^{id_2},\,id_1^3$
$x^y+y^3\,,\,x^y\,,\,y^3\,,\,x\,,\,y\,,\,3,+\,,\,id_1^{id_2}+id_2^3,$
$id_1^{id_2},\,id_1^3,\,x^y+y^{const}\,,\,y^{const}\,,\,id_1^{id_2}+id_2^{const}\,,\,id_1^{const}$
x^y+y^2
x^y+y^2
$x^y+y^2\,,\,id_1^{id_2}+id_2^2$
$x^y+y^2\,,\,id_1^{id_2}+id_2^2,$
$x^y+y^{const}\,,\,id_1^{id_2}+id_2^{const}$
$x^y+y^{const}\,,\,id_1^{id_2}+id_2^{const}$
Match!

MREC corpus

Mathematics REtrieval Corpus (MREC, version 2011.4.439).

- 439,423 documents (originated from arXMLiv [10], validated, enriched with metadata for snippet generation).
- Uncompressed size 124 GB, compressed 15 GB.
- 158 million input formulae, 2.9 billion subexpressions indexed (Lucene index size 47 GB).

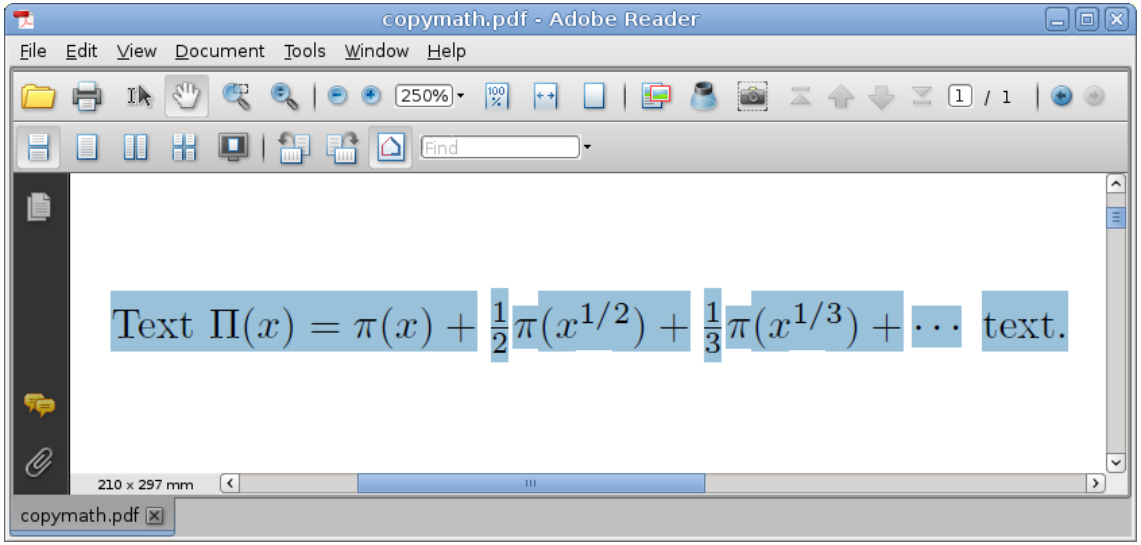
Selected Publications

- Sojka, Petr, Líška, Martin, Růžička Michal. Building Corpora of Technical Texts : Approaches and Tools. In Aleš Horák, Pavel Rychlý. *Fifth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2011.* Brno: Tribun EU, 2011, 71–82. ISBN 978-80-263-0077-9.
- Růžička, Michal, Sojka, Petr. *Redakční systém odborného časopisu s podporou exportu do digitální knihovny.* Zpravodaj CSTUG, Brno: CSTUG, 2011, 4–20. ISSN 1211-6661. 2011.
- Líška, Martin, Sojka, Petr, Růžička, Michal, Mravec, Peter. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. *DML 2011: Towards a Digital Mathematics Library.* Brno: Masaryk University, 2011, 77–84. ISBN 978-80-210-5542-1.
- Růžička, Michal a Sojka, Petr. Data Enhancements in a Digital Mathematical Library. In *DML 2010 Towards a Digital Mathematics Library.* First edition. Brno: Masaryk University Press, 2010, 69–76. ISBN 978-80-210-5242-0.
- Filej, Miha, Růžička, Michal, Šárfy, Martin, Sojka, Petr. Metadata Editing and Validation for a Digital Mathematics Library. In *DML 2010 Towards a Digital Mathematics Library.* First edition. Brno: Masaryk University Press, 2010. 57–62. ISBN 978-80-210-5242-0.
- Růžička, Michal. Automated Processing of TeX-Typeset Articles for a Digital Library. In *DML 2008: Towards Digital Mathematics Library.* First edition. Brno: Masaryk University, 2008. 167–176. ISBN 978-80-210-4658-0.

Future Works

- Award winning Gensim system – Topic Modelling for Humans by Radim Řehůřek.
- Image search à la MUFIN?
- Mathematical equivalence computation via symbolic algebra system?
- Weighting optimization for MIaS (by machine learning)?
- Addition of Content MathML tree indexing?
- `copypaste` LaTeX package?

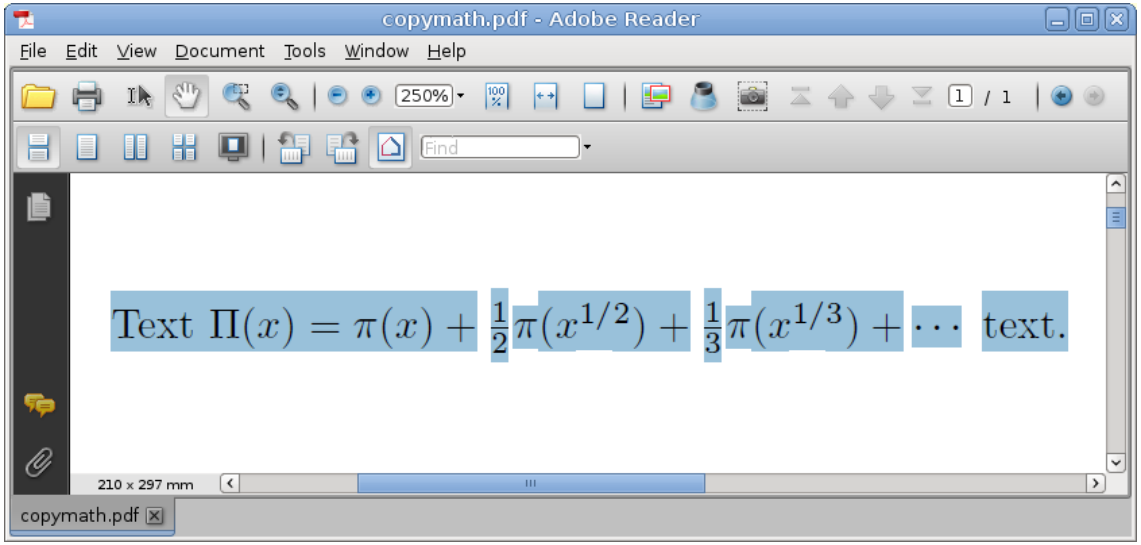
A Standard PDF Document



LaTeX source code:

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

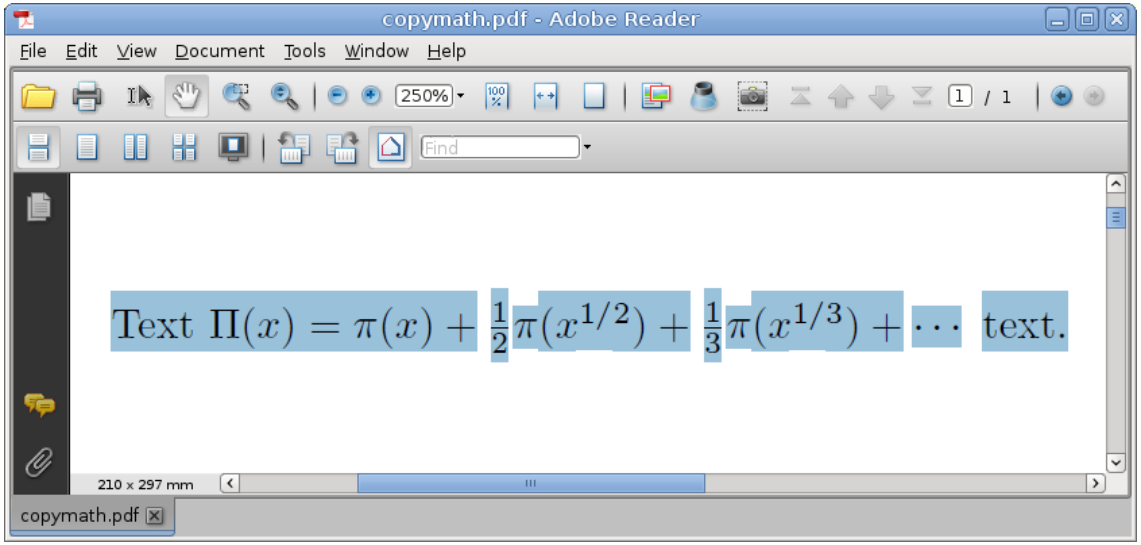
A Standard PDF Document



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 3.875 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923 Td [(1)]TJ
ET
```

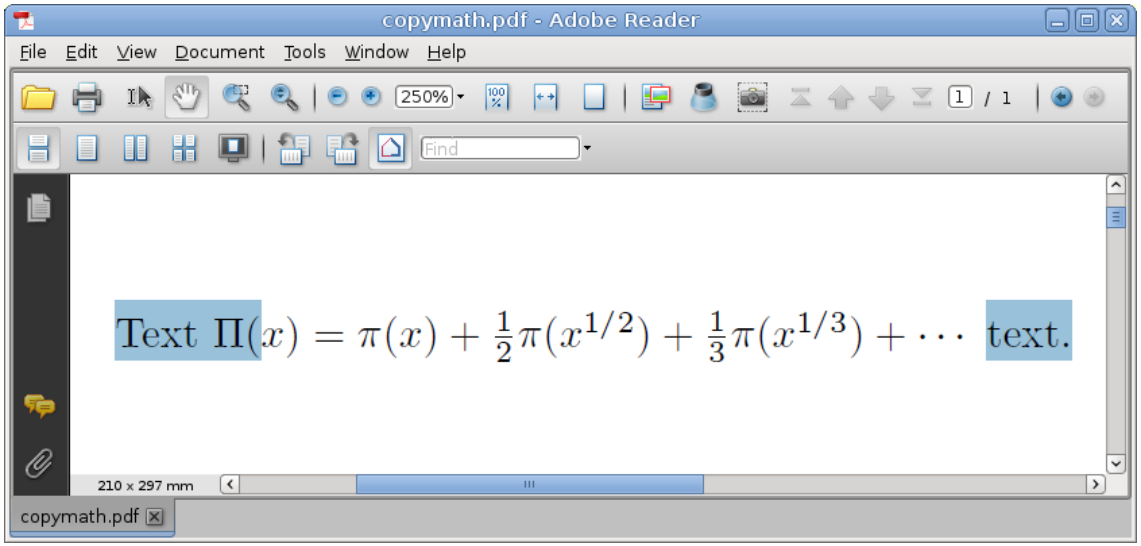
A Standard PDF Document



Text copied to clipboard:

```
Text   ( ) =   ( ) + 1
2 ( 1/2) + 1
3 ( 1/3) + · · · text.
```

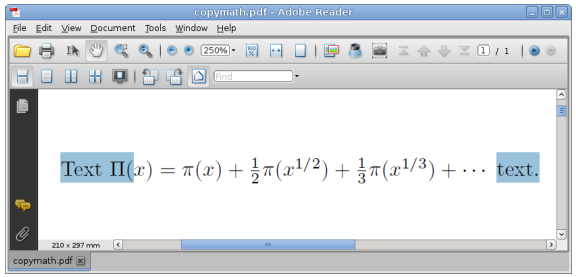
copypaste Enabled PDF Document



LaTeX source code:

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

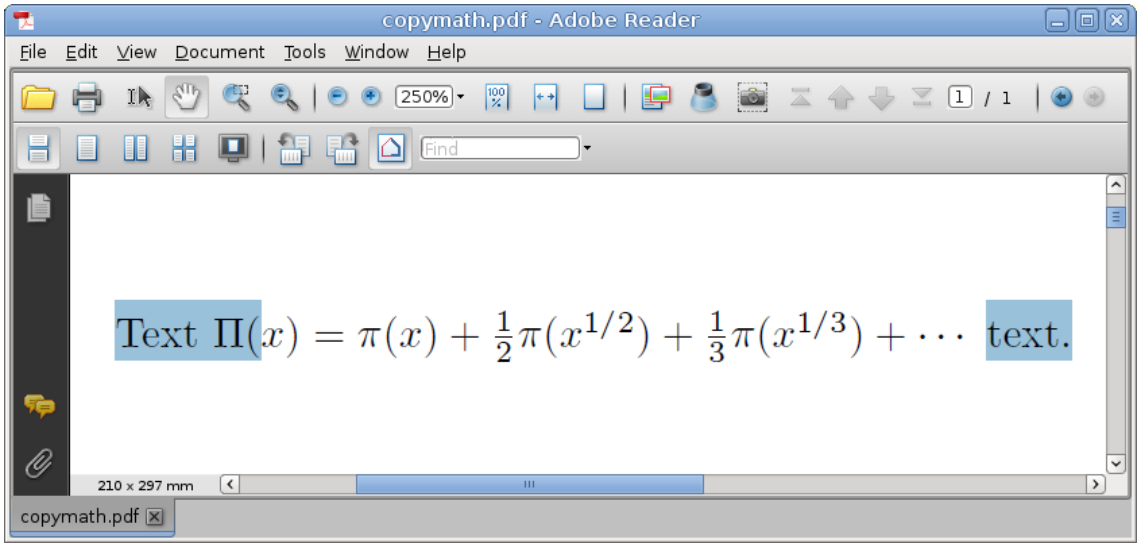
copypaste Enabled PDF Document



PDF code:

```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ
ET
1 0 0 1 171.959 707.125 cm
/Span <<
/ActualText<245C506920287829203D205C706920287829202B205C66726163207B317D7B32
7D5C70692028785E7B312F327D29202B205C66726163207B317D7B337D5C70692028785E7B31
2F337D29202B205C63646F74732024> >> BDC
1 0 0 1 -171.959 -707.125 cm
BT
/F17 9.9626 Tf 171.959 707.125 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 3.875 0 Td
[(x)]TJ/F17 9.9626 Tf 5.694 0 Td [(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923
Td [(1)]TJ
ET
```

copypaste Enabled PDF Document



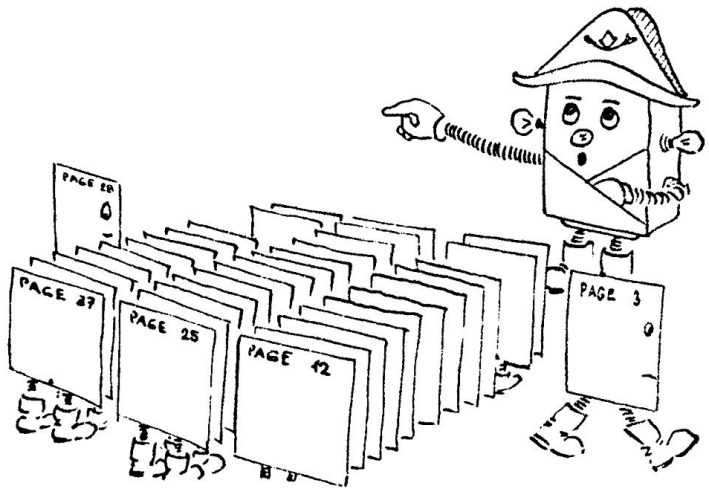
Text copied to clipboard:

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

Conclusion

Q: ‘What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?’

A: **‘Math formulae search.’**



EuDML : the European Digital Mathematics Library [online].
2011 [cit. 2011-11-30].
Available at WWW: <http://www.eudml.eu/>.

EuDML : the European Digital Mathematics Library Beta [online].
2011 [cit. 2011-11-30].
Available at WWW: <http://demo.eudml.eu/>.

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, P. (ed.): Towards a Digital Mathematics Library.
Masaryk University, Paris, France (Jul 2010), pp. 105–117, <http://dml.cz/dmlcz/702579>

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

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 (July 2011),
<http://dx.doi.org/10.1007/978-3-642-22673-1_16>

Sojka, P., Líška, M.: The Art of Mathematics Retrieval. In: Tompa, F., Hardy, M. (eds.) Proceedings of DocEng 2011 Conference. pp. 57–60. ACM. Mountain View, September 2011.


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 (July 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, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), 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 (eds.) *Proceedings of DML 2011*, pp. 77–84, Bertinoro, Italy, July 2011. Masaryk University. <http://dml.cz/dmlcz/702604>.