Semantics, Search and Digital Libraries (of Math)[1]

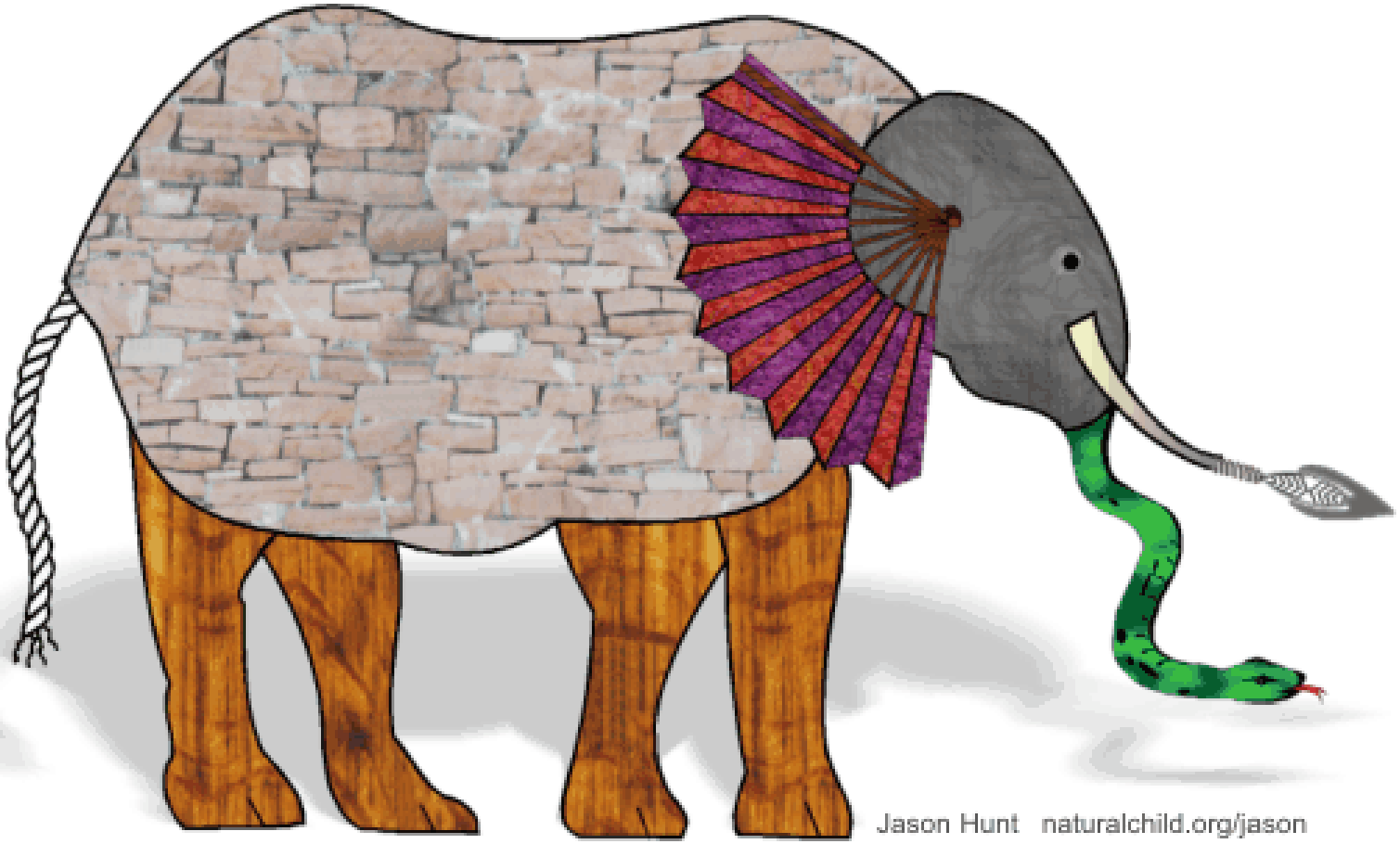
Petr Sojka

Faculty of Informatics, Masaryk University, Brno, CZ, EU

Sep 9th, 2009

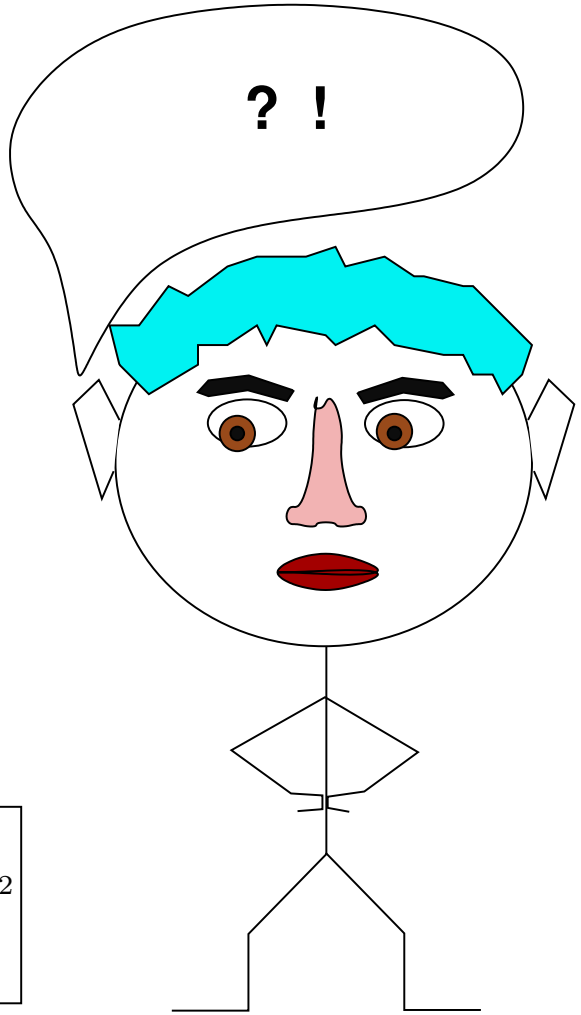
[1]Supported by JISC and AS CR grant #1ET200190513

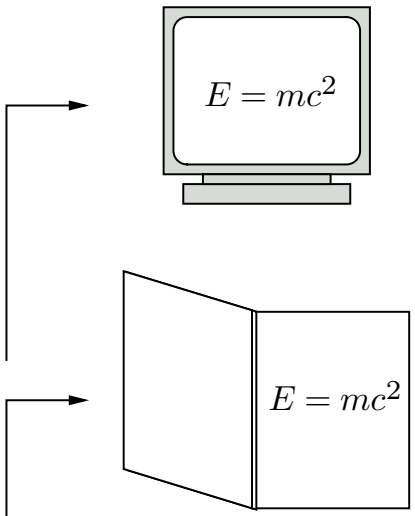

衆聲
探象之圖



Q: Is elephant a wall (belly), hand fan (ear), solid pipe (tusk), pillar (leg), rope (tail) or tree branch (trunk)?


$E = mc^2$


? !


$E = mc^2$
$E = mc^2$


Značkování
Markup
Návrh
Design
Sazba
Typesetting
Korektury
Proofreading
Předloha
Preprint
Tisk
Print
Distribuce
Distribution

Levels of text/math understanding/processing

1.0 lexical – words, **strings** of characters/TeX's $ $.

2.0 syntactical – phrases, **parsed** formulas (trees/MathML).

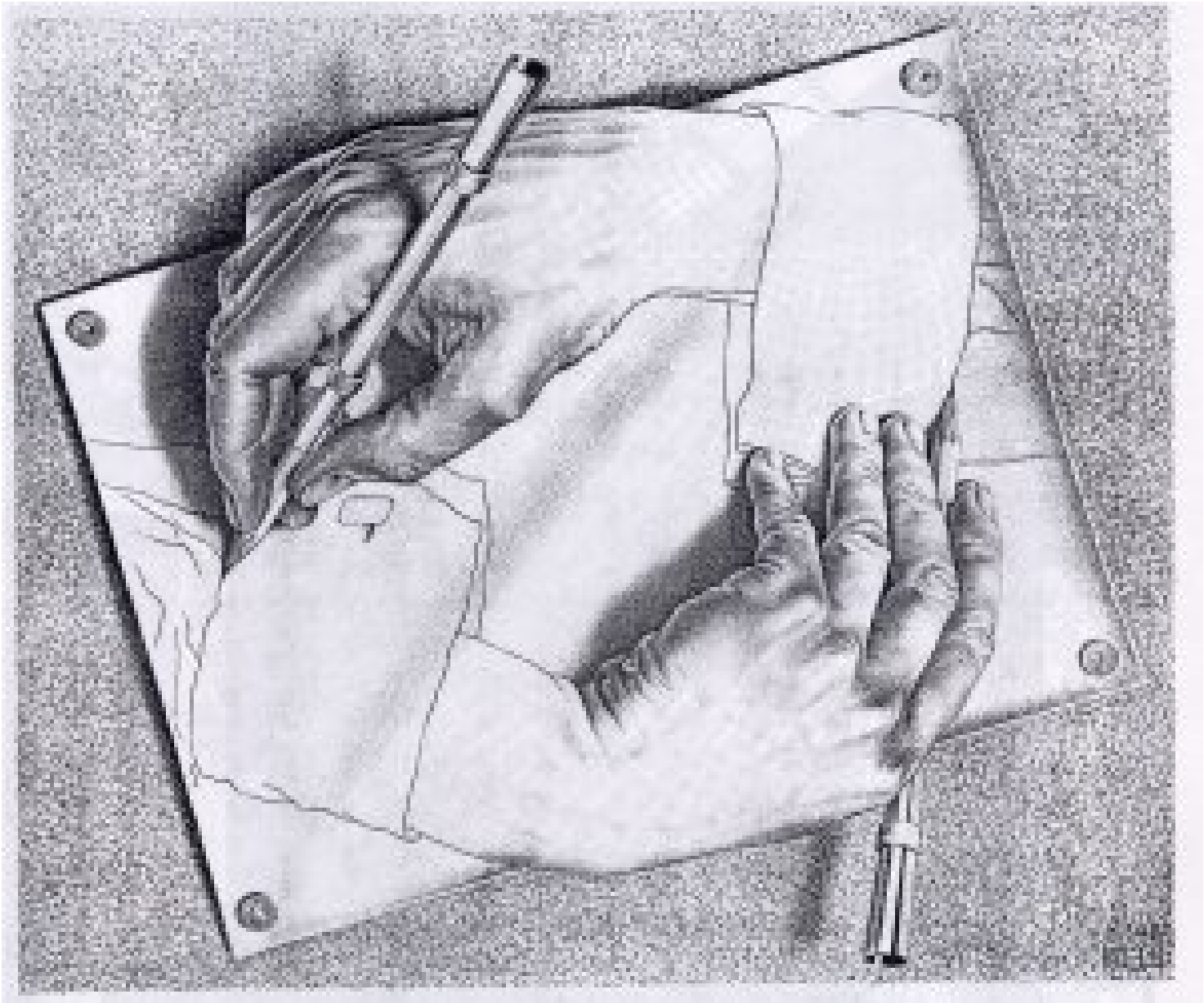
3.0 semantical – **meaning** of parsed phrases (cloud tags/ontologies/OpenMath).

Problem of message (content+form) representation (of math when transporting the message over the web).

Google around 1.5 now (no semantics, but for the purpose are people happy).

Many valid but different purposes for processing math

- Format choice **depends** on application's **purpose**.
- Most applications have its own internal format anyway.
- For **exchange** it seems that **XML/MathML** (but which one?) currently wins (cut&paste in Windows 7, CAS).
- For authoring it seems that (La)TEX is preferred.
- Quite different requirements have theorem proving systems and computer algebra systems.



Math authoring tools: LaTeX, AMSLaTeX

- Good for authors: authors may express as close as possible to their mental model in their brain (new macros, namespaces).
- This author's advantage make headaches to the editors, robots and those wishing to convert to some **standard** formalism (to index, evaluate, ...).
- Many different macropackages, and active development as possibilites grow (XeTeX, LuaTeX, pdfTeX), ...

Mark up (author)

```
&\elevenit I\kern.7ptllustrations by\cr
&DU\kern-1ptANE BIBBY\cr
\noalign{\vfill}
&\setbox0=\hbox{\manual77}%
\setbox2=\hbox to\wd0{\hss\manual6\hss}%
\raise2.3mm\box2\kern-\wd0\box0\cr % A-W logo
&ADDISON\kern.1em--WESLEY\cr
&PUBLISHING COMP\kern-.13emANY\kern-1.5mm\cr
```

?

NO! (for some purposes, e.g. web communication)

MathML: content vs. presentation

- MathML 2.0/3.0: XML namespace, W3C standard, supported and widely used.
- supported: in browsers (Firefox, IE, including fonts needed), symbolic computation sw (Mathematica, Maple), OCR sw (Infty :-)).
- de facto standard interapplication XML exchange format.
- extend to cover new things or not? (which DTD, symbol or notion e**X**tend/add?)

OpenMath and OMDoc

- OpenMath: markup language for specifying meaning of mathematical formula—complements MathML (used usually in it's presentation form only).
- Developed since 1993 in Europe (Helsinki).
- For more richly structured content dictionaries (and generally for arbitrary mathematical documents) the **OMDoc format** extends OpenMath by a **statement level** (including structures like definitions, theorems, proofs and examples, as well as means for interrelating them) and a **theory level**, where a theory is a collection of several contextually related statements.
- James Davenport's lightning talk.

```
<OMOBJ xmlns=''http://www.openmath.org/OpenMath''>
  <OMA cdbase=''http://www.openmath.org/cd''>
    <OMS cd=''relation1'' name=''eq''/>
    <OMV name=''x''/>
    <OMA>
     <OMS cd=''arith1'' name=''divide''/>
      <OMA>
        <OMS cdgroup=''http://www.example.com/mathops'' cd=''multiops'' name=''
        <OMA>
          <OMS cd=''arith1'' name=''unary_minus''/>
          <OMV name=''b''/>
        </OMA>
        <OMA>
          <OMS cd=''arith1'' name=''root''/>
          <OMA>
            <OMS cd=''arith1'' name=''minus''/>
            <OMA>
              <OMS cd=''arith1'' name=''power''/>
              <OMV name=''b''/>
              <OMI>2</OMI>
            </OMA>
            <OMA>
              <OMS cd=''arith1'' name=''times''/>
              <OMI>4</OMI>
```

Semantically enhanced TEX—sTEX (by Michael Kohlhase)

- LATEX macropackage which will enable the author to add semantic information to the document in a way that does not change the visual appearance. This process is also referred to as semantic pre-loading and the collection of macro packages is called Semantic TeX (sTeX). Thus, sTeX can serve as a conceptual interface between the document author and MKM systems: Technically, the semantically pre-loaded LaTeX documents are transformed into the (usually XML-based) MKM representation formats, but conceptually, the ability to semantically annotate the source document is sufficient.
- To convey semantics to be convertible to OMDoc.
- Grabing most abstract semantic level, but not in widespread use by authors (additional effort does not pay back). Lack of motivation to be used.

Other formats

- DITA Darwin Information Typing Architecture: XML-based, end-to-end architecture for authoring, producing, and delivering technical information. This architecture consists of a set of design principles for creating "information-typed" modules at a topic level and for using that content in delivery modes such as online help and product support portals on the Web.
- OOXML OpenOffice XML (XML+ZIP).
- ODF OpenDocument Format (XML+ZIP).
- LaTeXML (Bruce Miller's mathematical encyclopaedia).


衆聲
探象之圖

Levels of search

1–2 Google Demo.

1.5–2.5 SearchPoint Demo.

1.5–3 Math WebSearch.

Mathematical search specifics [Kohlhase, Sucan 2006]

- Mathematical notation is context-dependent, e.g. binomial coefficients: $\binom{n}{k}$, ${}_nC^k$, C_k^n, C_n^k.
- Identical presentations can stand for multiple distinct mathematical objects, e.g. $\int f(x)\,dx$ for several anti-derivative ioperators (Riemann, Lebesgue,...).
- Certain variations of notations are widely considered irrelevant, e.g. $\int f(x)\,dx$ and $\int f(y)\,dy$.

Math search systems and platforms

- MathWebSearch, I. Şucan, M. Kohlhase (Bremen, GE)
- MathDex, R. Miner (Design Science, US); DLMF search, A. Youssef (Washington, US)
- EgoMath, J. Mišutka, L. Galamboš (Prague, CZ)

Other notable related work:

- Mathematical formulae recognition from PDF, J. Baker, A. Sexton, V. Sorge, Birmingham, UK.
- Infty system, M. Suzuki, Kyushu, JP.
- ActiveMath web-based math-learning environment, P. Libbrecht, DKFI, Saarbrücken, GE.
- SWiM: A Semantic Wiki for Mathematical Knowledge Management, KWARC, Bremen, GE.

Some technical aspects of search (EgoMath)

- normalization.
- linearization (search engine may work on strings/words).
- partial evaluation (e.g. distributivity).
- generalization (introduction of variables in the index).
- ordering (for commutative operators).

Other approaches (MathWebSearch) possible, cf. Stefan Anca's lightning talk.


衆聲
探象之圖

Levels of Digital Libraries

1.0 classical library + scanned bitmaps.

2.0 interconnected, crosslinked and validated repository of peer reviewed docs, possibly fully (not only metadata) indexed on syntactic level.

3.0 dynamically personalized, formalized knowledge in semantic representation with inference.

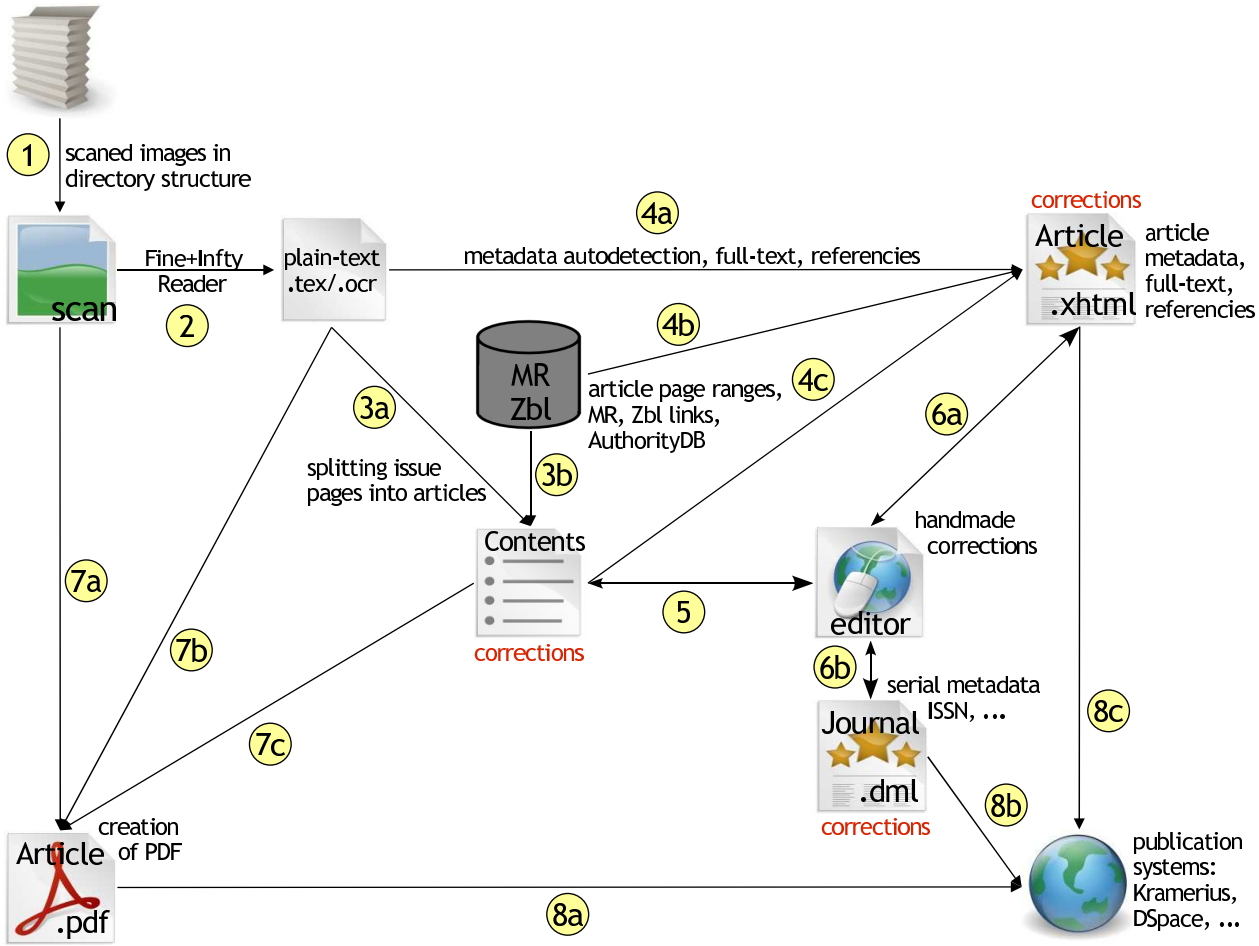
Google Scholar/Books now around 1.5.

Google Scholar demo.

From pixels (and minds) to minds – vision of WDML

- Google Scholar $\|_{\text{peer\_reviewed\_math}}$ but better!
- Vision of World Digital Math Library (WDML) that will bring the enduring mathematical legacy to researchers (and students) worldwide.
- High quality, checked content, crosslinking via reviewing databases Zentralblatt MATH or Mathematical Reviews (more than 2,800,000 reviewed articles).
- Estimation of 50–100,000,000 pages of math in total 'only' (to be stored in a pocket size disc. (demo :-).
- 250,000 distinct authors (**minds**) sent papers for a review in the last decade in mathematical sciences.

Top-level DML-CZ workflow (different primary data)



Digitization phases – math handling

acquisition preparation, document acquisition, copyright issues handling;

scanning document scanning, main metadata entering, scanning checks;

image processing main OCR, image enhancements;

semantic processing document markup enhancement, semantic processing, document classification, citation linking, document clustering, indexing;

presentation visualization techniques of document repository, digital library web portal, interfaces to other services and search engines for the semantic database document.

DML-EU project

- Recently accepted EU 1.6 mio EUR project ;-).
- Pilot project (36 months, 11 partners) to set up European portal for mathematical content.
- Virtual digital library to access and search data from existing (national, publisher) repositories like NUMDAM, DML-CZ, with metadata from Zentrallblatt MATH.
- Workpackage on ‘metadata augmentation and enhancements’.

Formalized mathematical knowledge/libraries

- CAS: Mathematica, Maple, Mathcad, or OS: Axiom, Maxima, PARI/GP, Reduce...
  comparison on Wikipedia
- MKM (proof assistants): Mizar (Trybulec, 1973), HOL, Isabelle, Coq,...: cf. Implementations on wikipedia.
- Interactive math docs: MathDox, ?Google Wave

MKM systems usually represent mathematical knowledge in the internal representation that allow inference. External interface format is usually MathML.


衆聲
探象之圖

Takeoff messages

- plethora of math content formats, tools and approaches for different purposes: LaTeX often wins for authoring, MathML for bot/program's exchanges.
- new possibilities (Google Wave), speed of changes/ development/ communication is increasing: people not tinkering with new tools and possibilities may loose (cf. Terry Tao).
- good math OCR is badly needed.

Blind monks and elephant metaphor

All of you are right. The reason every one of you is expressing, searching and storing the math differently is because each one of you are touching the different part of the elephant (the true of math web content). So, actually the elephant has all the features you mentioned.

Or try to talk to another nearby monk to share his feelings :-).