KWARC: Knowledge Adaptation and Reasoning for Content

Research in Foundations, Interaction, and Semantization

Michael Kohlhase

http://kwarc.info/kohlhase
Center for Advanced Systems Engineering
Jacobs University Bremen, Germany

April 12, 2013

The KWARC Group

Overview: KWARC Research and Projects

Applications: eMath 3.0, Active Documents, Semantic Spreadsheets, Semantic Help Systems, Semantic CAD/CAM, Change Mangagement, ...

Foundations of Math:

- MathML, `OpenMath`
- advanced Type Theories
- MMT: Modular Math Theories
- Logic Morphisms/Atlas
- Theorem Prover Interoperability

Knowledge Mgt. & Interaction:

- Semantic Interpretation
- JOBAD: Document-Embedded Interaction
- TNTBase: Versioned XML Storage
- Math Archives

Semantization:

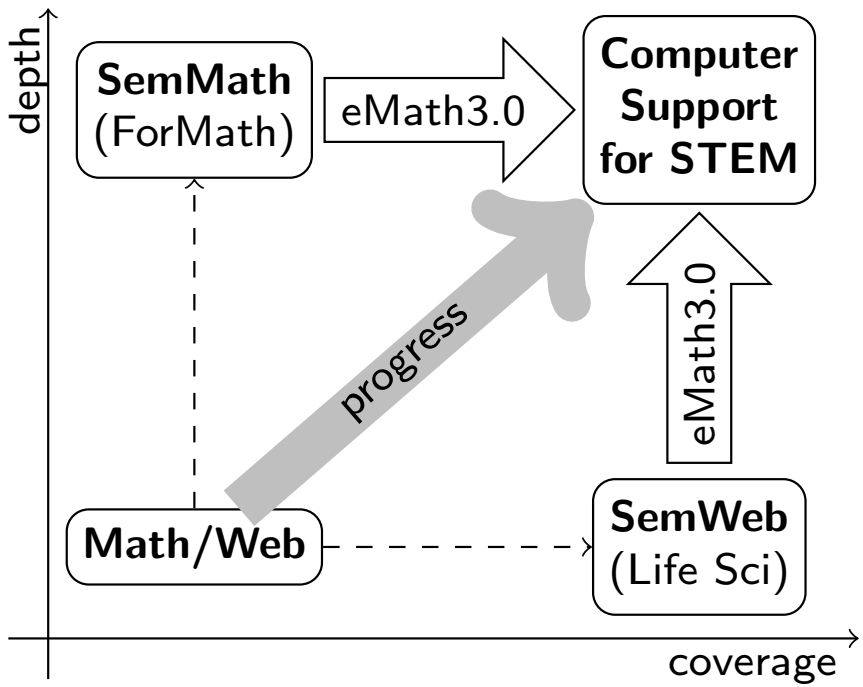
- LaTeXML: LaTeX $\rightarrow$ XML
- sTeX: Semantic LaTeX
- invasive editors
- Context-Aware IDEs
- Mathematical Corpora
- Linguistics of Math

Foundations: Computational Logic, Web Technologies, OMDoc

Contributions from KWARC@Jacobs@Bremen

- ⚠ STEM Knowledge: more like a Digital Library than the Open WWW ⚠
  (reviewed publication $\rightsquigarrow$ less junk, little duplication, partly inaccessible)

- Combination of SemMath and SemWeb
- Expertise in Semantics of STEM Docs
- Expressive Analysis Target Format (OMDoc)
- Software Stack for Semantic Processing
- eSTEM3.0 System Planetary (Active Docs)
- Invasive authoring (Office/LaTeX)
- Semantic Analysis for LaTeX-based Corpora
  (arXiv, ZBL, PlanetMath...)



- ⚠ We use Math as a test tube for STEM (Science, Tech, Eng, & Math) ⚠

KWARC People I

- Michael Kohlhase (Professor; Project Lead)
  - Projects: OMDoc, sTEX, arXMLiv, NL Semantics, MathSearch...
  - Specializes: "world domination" (ubiquitous computer-supp. math)

- Heinrich Stamerjohanns (Director of CS Labs; Senior Collaborator)
  - Projects: arXMLiv, PhysML, PhysNet, JGrader, ...
  - Specialties: Web Information Systems, Document Management...

- Florian Rabe (Jacobs PostDoc)
  - Projects: OMDoc2, LATIN, MMT (loves category theory)
  - Specializes: metalogics, language design, math foundations, ...

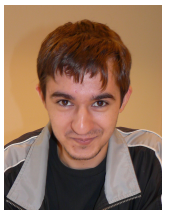
- Andrea Kohlhase (Jacobs PostDoc)
  - Projects: SiSsi, Planetary, MathSearch
  - Specializes: Semantic Interaction, Semantic Design, HCI

- Fulya Horozal (Doctoral Student (Metalogics))
  - Projects: LATIN
  - Thesis: Representing Logics and Logic Morphisms

KWARC People II

- Constantin Jucovschi (Doctoral Student (Editing Support))
  - Projects: SiSsi, PLANETARY
  - Thesis: Integrated Development Environemtns for STEM Documents

- Deyan Ginev (Doctoral Student (Math Linguistics))
  - Projects: LaMaPuN, `arXMLiv`, PLANETARY, sTEX, ...
  - Thesis: Semantizising Math Formulae



- Michnea Iancu (Doctoral Student (OMDoc2))
  - Projects: MathSearch
  - Thesis: informal MMT, OAFF
- M.Sc. Students: with thesis titles and ETA
  - Aivaras Jacubauskas: *Declarative Representation of Computation* (2013)
  - Corneliu Prodescu: *Formula Search Engines* (2014)
- B.Sc. Students Jan Dörrie, Felix Mance, Stefan Mirea, Daniel Rachev, Dharam Kapila (do thesis research and help with the KWARC projects)

Introduction/Background

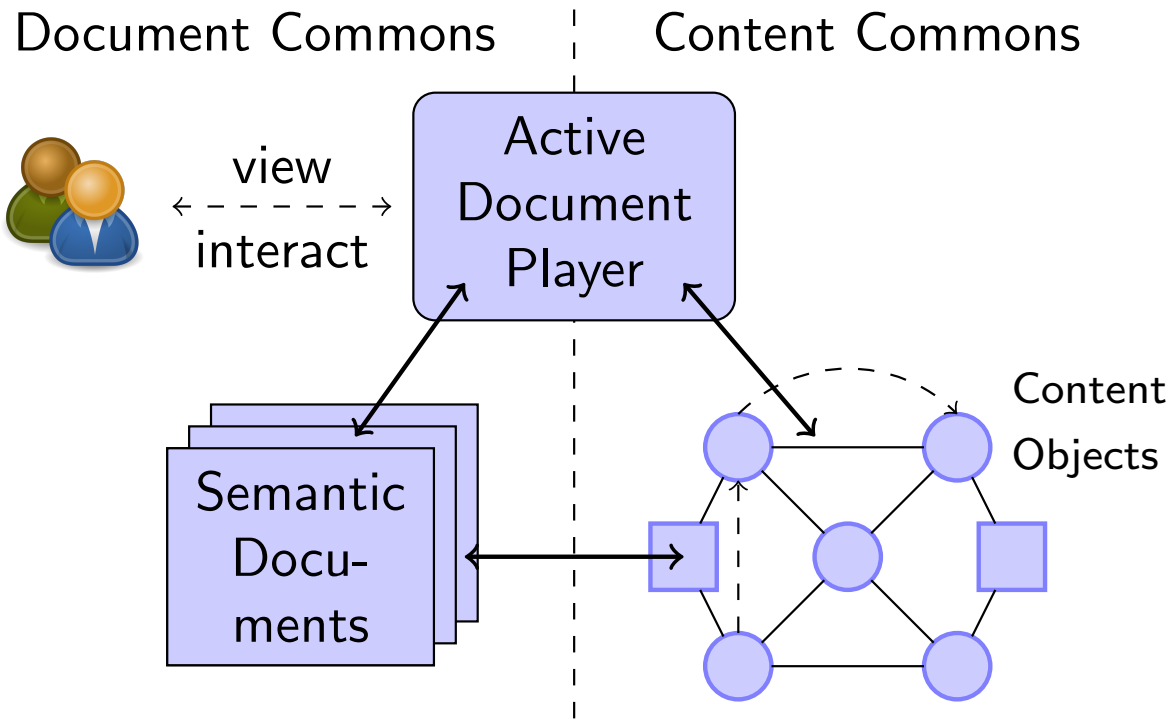
- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation,
- its conservation, dissemination, and utilization constitutes a challenge for the community and an attractive line of inquiry.
- Challenge: How can/should we do mathematics in the 21st century?
- Mathematical knowledge and objects are transported by documents
- Three levels of electronic documents:
  1. digitized (usually from print):
  2. presentational: encoded text interspersed with presentation markup
  3. semantic: encoded text with functional markup for the meaning

  transforming down is simple, transforming up needs humans or AI.
- Observation: Computer support for access, aggregation, and application is (largely) restricted to the semantic level.
- This talk: How do we do maths and math documents at the semantic level?

Semantic Interaction with Technical Documents

The Active Documents Paradigm

- **Definition 1** The active documents paradigm (ADP) consists of
  - *semantically annotated documents* together with
  - background ontologies (which we call the content commons),
  - *semantic services* that use this information
  - a document player application that embeds services to make documents executable.



- **Example 2** Services can be program (fragment) execution, computation, visualization, navigation, information aggregation and information retrieval

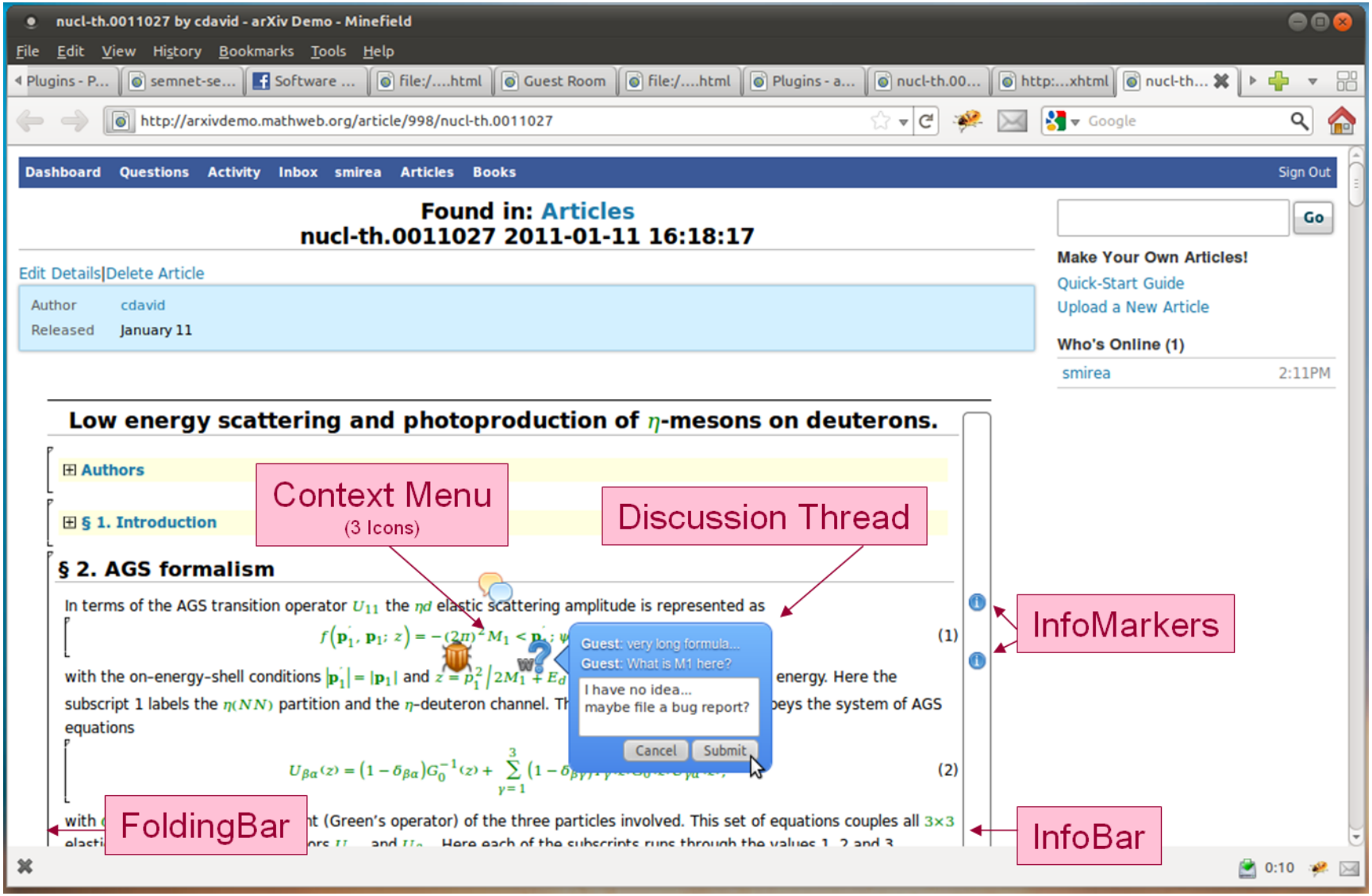
What you get out of it depends on what you put into it

- The level of support depends on the level of semantic enhancement!
- Presentation Structure Level: e.g. legacy-rich sites like arXiv.org and PlanetMath.org (or ScienceDirect.com)
- Semantic Level: e.g. flexiformal digital libraries: Course Notes and Manual
- Formal Level: e.g. specification and verification in the LATIN Logic Atlas

What you get out of it depends on what you put into it

- The level of support depends on the level of semantic enhancement!
- Presentation Structure Level: e.g. legacy-rich sites like arXiv.org and PlanetMath.org (or ScienceDirect.com)
  - discourse-level/formula folding
  - icon menu for embedded services,
  - localized discussions + notification bar (with Math)
  - localized refereeing
  - formula search engine
  - Ontology-based access (here: MSC)
- Semantic Level: e.g. flexiformal digital libraries: Course Notes and Manual
- Formal Level: e.g. specification and verification in the LATIN Logic Atlas

What you get out of it depends on what you put into it



What you get out of it depends on what you put into it

- The level of support depends on the level of semantic enhancement!
- Presentation Structure Level: e.g. legacy-rich sites like arXiv.org and PlanetMath.org (or ScienceDirect.com)
- Semantic Level: e.g. flexiformal digital libraries: Course Notes and Manual
  - prerequisites graph and guided tours
  - definition lookup
  - unit conversion
  - "monographs", that aggregate "modules" into narrative structures
  - semantic editing facilities
  - Management of Change (versioning and change impact analysis)
- Formal Level: e.g. specification and verification in the LATIN Logic Atlas

What you get out of it depends on what you put into it

Definition Lookup

$f \subseteq X \times Y$, is called a **partial function** , iff for all $x \in X$ there is at

Definition Lookup Results

DEFINITION:

Cartesian product :
$A \times B := \{\langle a, b\rangle \mid a \in A \wedge b \in B\}$, call $\langle a, b\rangle$ **pair** .

om

Semantic Folding

$s = s_i + v_i \Delta t + \frac{1}{2} a (\Delta t)^2$

Fold

Semantic Fold

$s = s_i + s_v + s$

contribution from acceleration

Unit Conversion

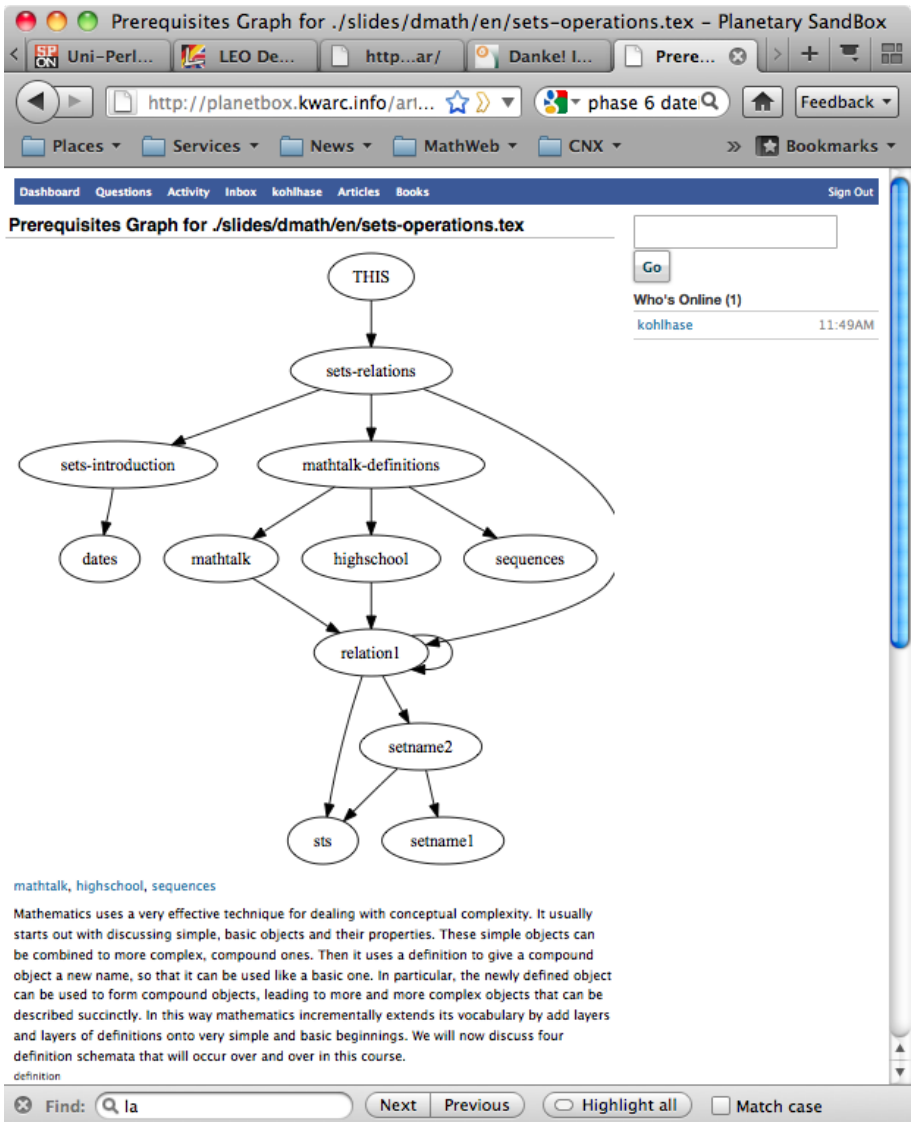
City A is 9144ft from city B and 5164ft from city C .

Look-up Definition
convert to miles
convert to meters
convert to feet
convert to inches
convert to yards

convert the units

City A is 3048m from city B and 5164ft from city C .

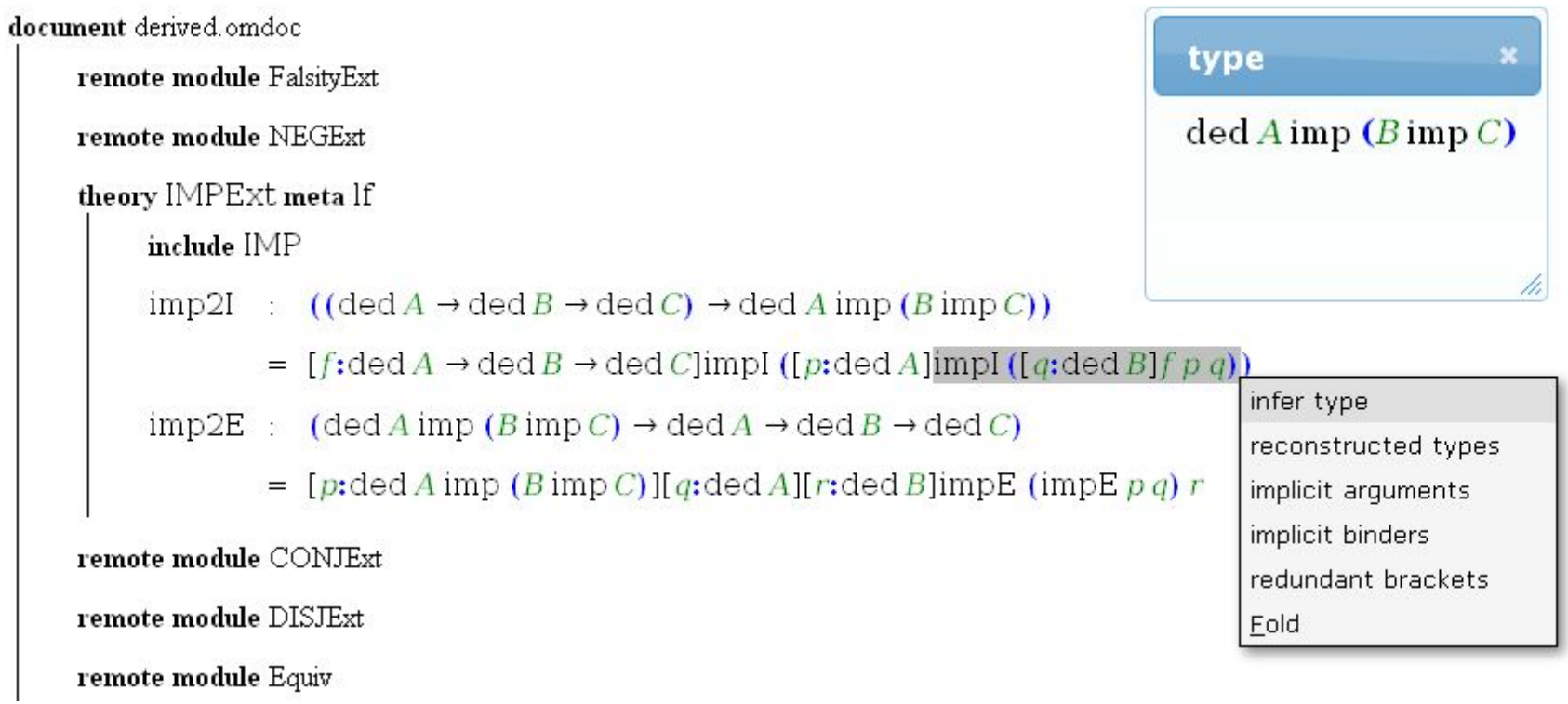
Prerequisites Navigation



What you get out of it depends on what you put into it

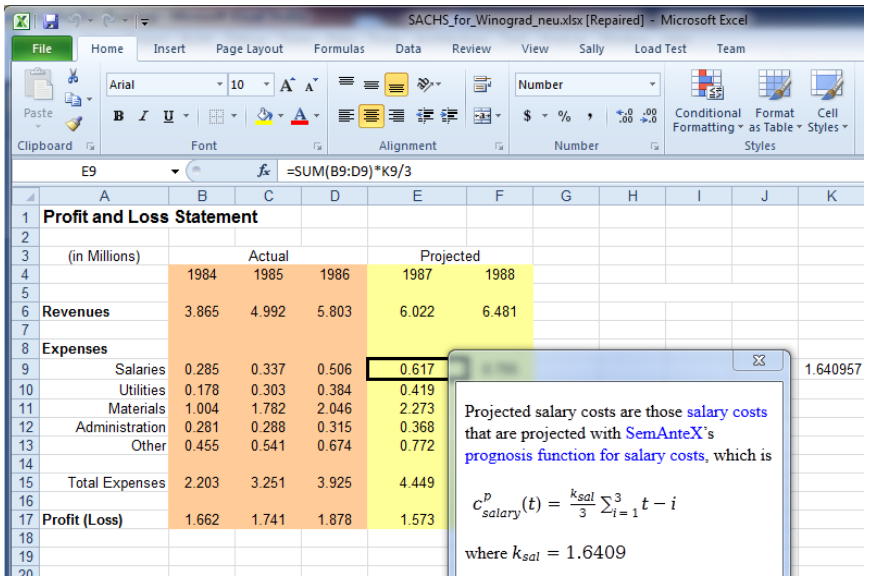
- The level of support depends on the level of semantic enhancement!
- Presentation Structure Level: e.g. legacy-rich sites like arXiv.org and PlanetMath.org (or ScienceDirect.com)
- Semantic Level: e.g. flexiformal digital libraries: Course Notes and Manual
- Formal Level: e.g. specification and verification in the LATIN Logic Atlas
  - flexible elision of brackets
  - argument reconstruction via external TWELF system
  - verification by the HETS system

What you get out of it depends on what you put into it

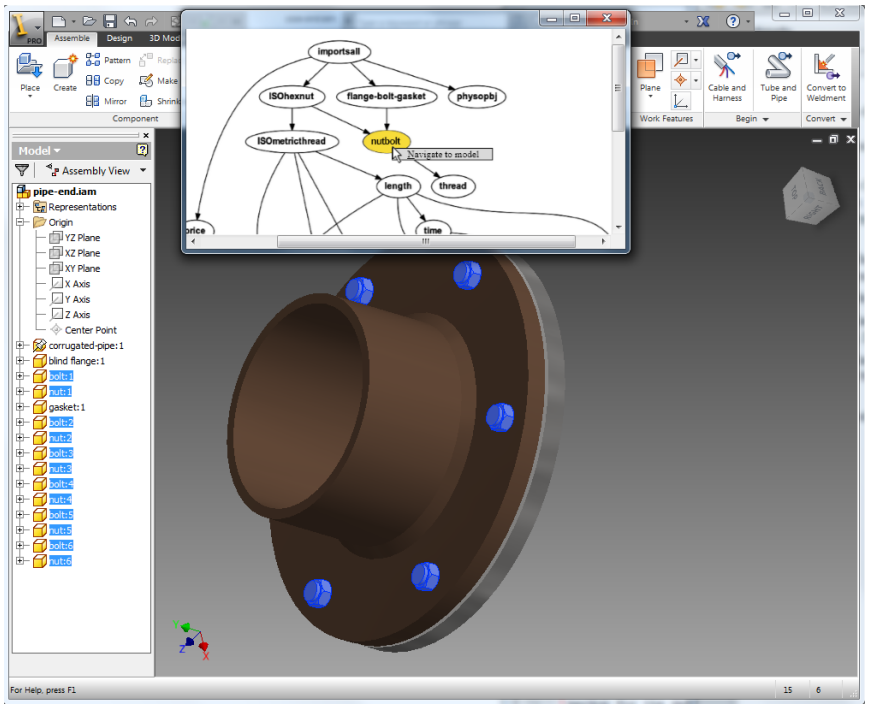


Embedding Semantic Services into Proprietary Applications

Definition Lookup in Excel



Semantic Navigation in a CAD System



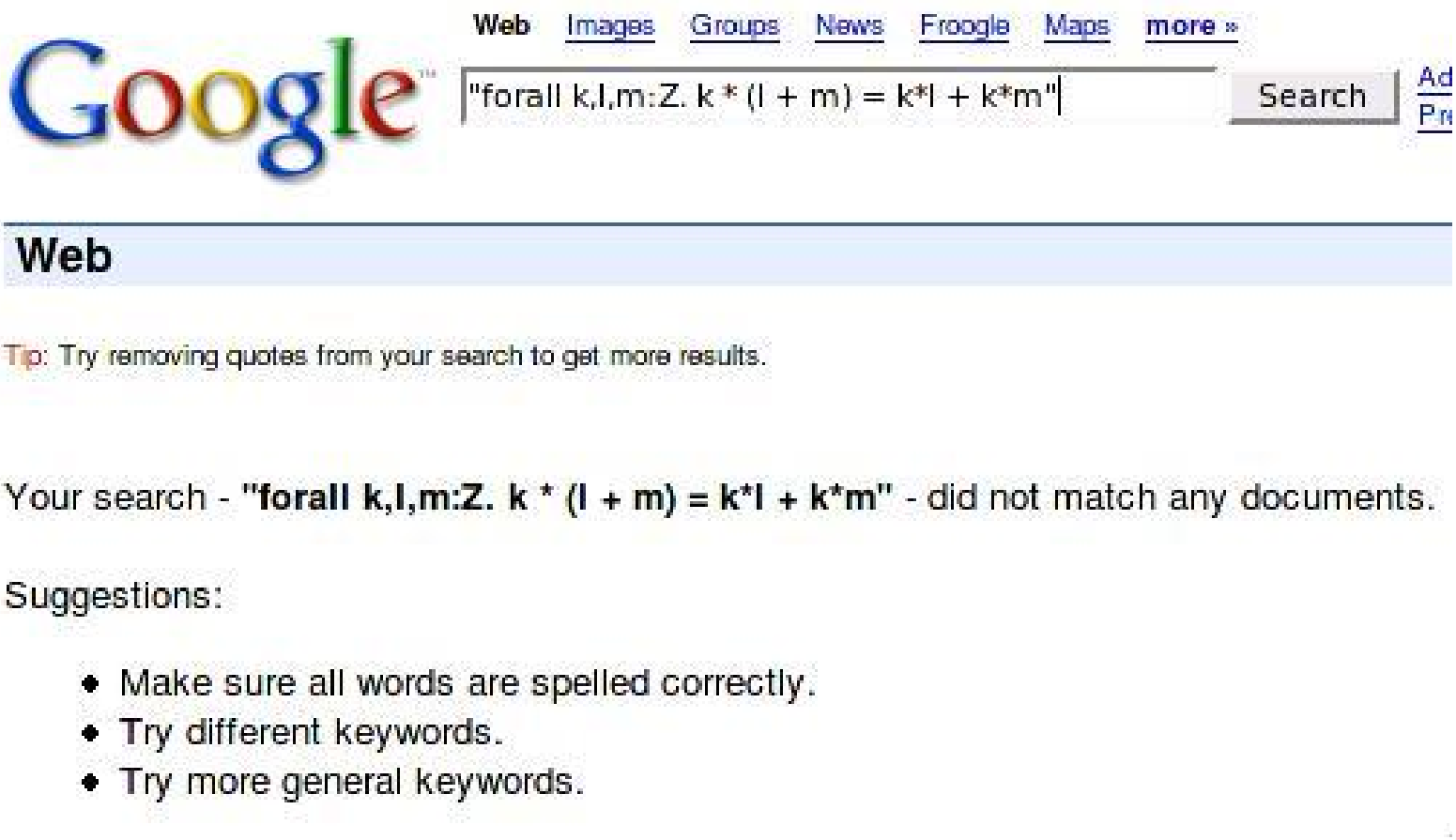
And of course. . . Math Search!

More Mathematics on the Web

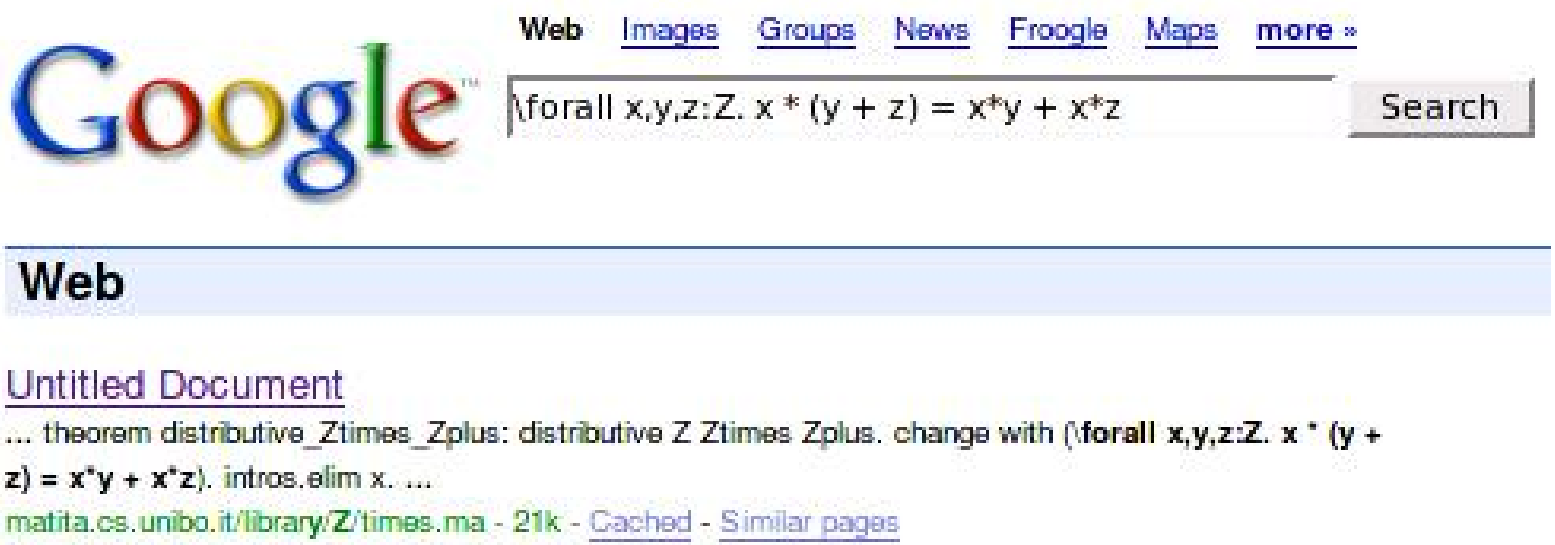
- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- ...
- Question: How will we find content that is relevant to our needs
- Idea: try Google (like we always do)
- Scenario: Try finding the distributivity property for $\mathbb{Z}$

$$(\forall k, l, m \in \mathbb{Z}.k \cdot (l + m) = (k \cdot l) + (k + m))$$

Searching for Distributivity



Searching for Distributivity

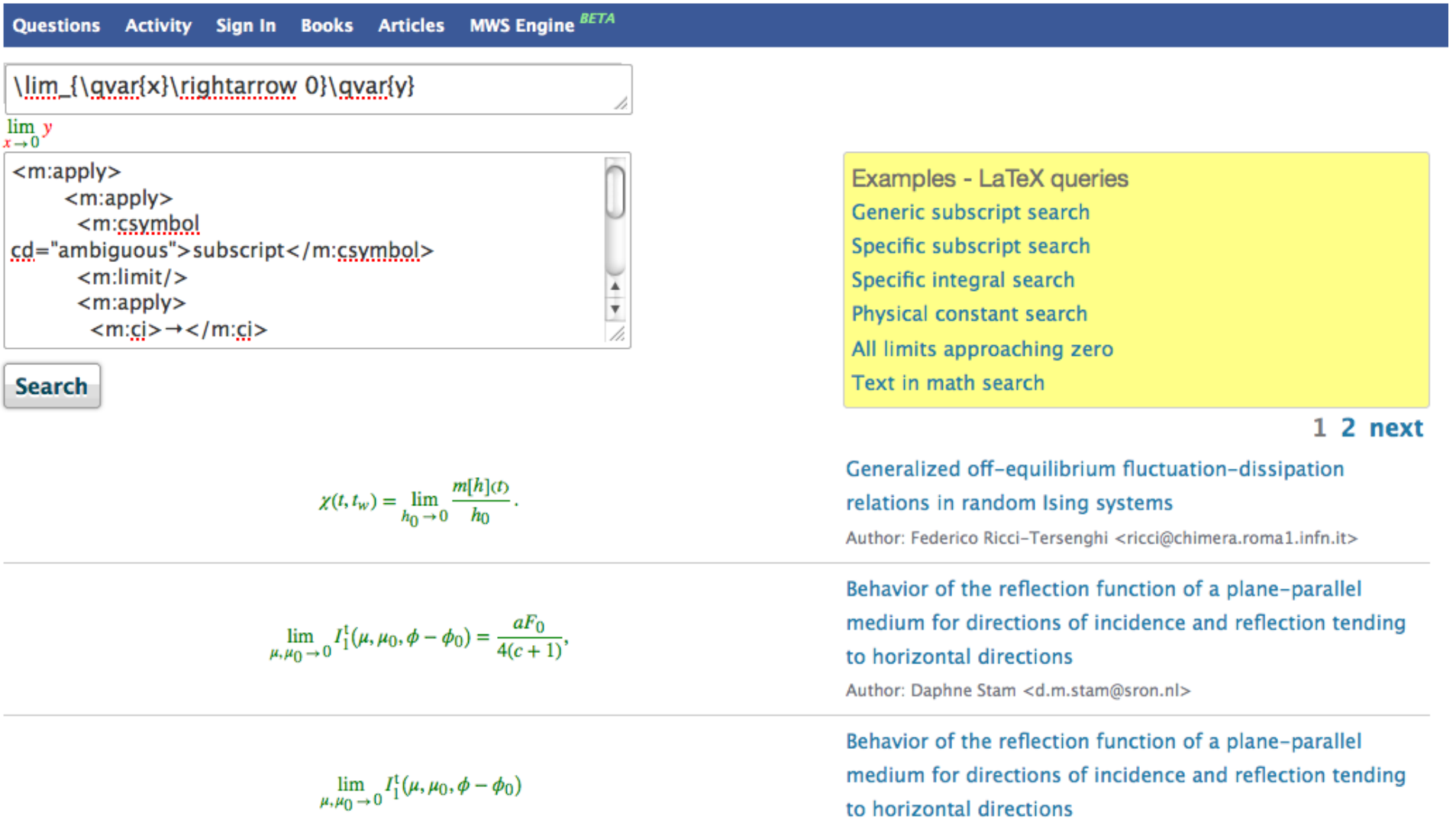


Searching for Distributivity

Of course Google cannot work out of the box

- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables.
    (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, MG11])
  ...and works surprisingly well (using Lucene as an indexing engine)
- Idea: Use database techniques (extract metadata and index it)
  Indeed this is done for the `Coq`/`HELM` corpus ([AGC+06])
- Our Idea: Use Automated Reasoning Techniques
  (free term indexing from theorem prover jails)

Instead of a Demo: LaTeX-based Search on the arXiv



Generalization Queries

- Application: Find (possibly) applicable theorems
- **Example 3** A researcher wants to estimate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ from above
  - Idea: Find inequation such that $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ matches left hand side.
  - Query: $\int_{\mathbb{R}^2} |\sin(x)\cos(x)|\,dx \leq \boxed{rhs}$
  - matches e.g. Hölder's Inequality in the index: ($\boxed{i}$ are universal variables)

$$\int_{\boxed{D}} \left|\boxed{f}(x)\boxed{g}(x)\right| dx \leq \left(\int_{\boxed{D}} \left|\boxed{f}(x)\right|^p dx\right)^{\frac{1}{p}} \left(\int_{\boxed{D}} \left|\boxed{g}(x)\right|^q dx\right)^{\frac{1}{q}}$$

  - Solution: Instantiate query accordingly and get

$$\int_{\mathbb{R}^2} |\sin(x)\cos(x)|\,dx \leq \left(\int_{\mathbb{R}^2} |\sin(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_{\mathbb{R}^2} |\cos(x)|^q\,dx\right)^{\frac{1}{q}}$$

Problem: Where do the index formulae come from in particular the universal variables (we'll come back to that later)

Instead of a Demo: Applicable Theorem Search in Mizar

```
definition
 let k, n be Ordinal;
 pred k divides n means :Def3: :: MTEST1:def 3
 ex a being Ordinal st n = k *^ a;
 reflexivity
 proof
  let n be Ordinal; :: thesis:
  thus ex a being Ordinal st n = n *^ a ;
```

ATP Proof not found

status: Timeout
Suggest hints, Unification query,

Suggested hints

t73_card_2, t39_ordinal2,

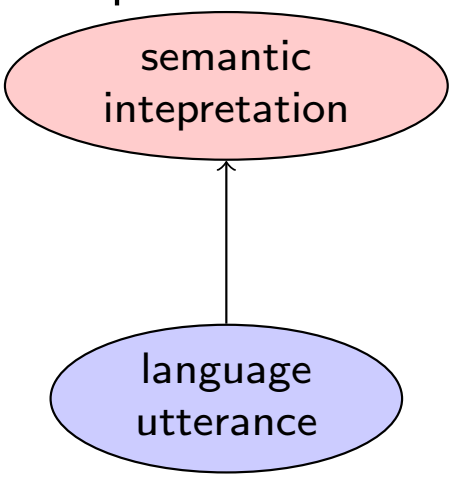
Try SPASS, Export problem to SystemOnTPTP

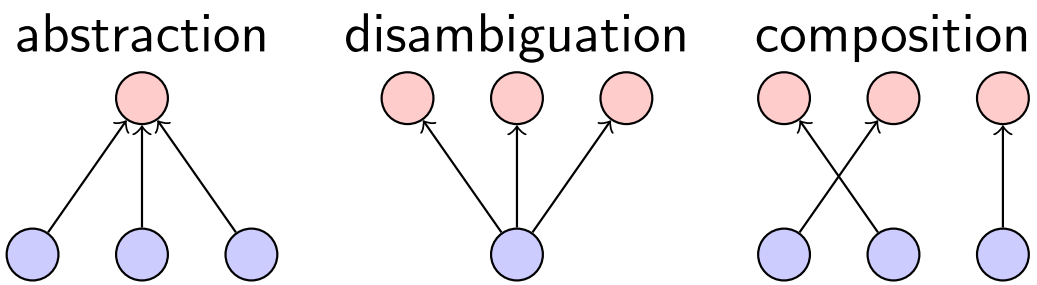
```
  :: thesis:
 end;
end;
```

Towards Semantic Math Libraries: Semantization

Language and Information

- humans use words (sentences, texts) in natural languages to represent information
- but:
- what really counts is not the words themselves, but the meaning information they carry.
- for questions/answers, it would be very useful to find out what words (sentences/texts) mean.
- Interpretation of natural language utterances: three problems





The Reconstruction Problem

- Mathematical communication relies on the inferential capability of the reader.
- semantically relevant arguments are left out (or ambiguous) to save notational overload (reader must disambiguate or fill in details.)

$$\log_2(x) \text{ vs. } \log(x) \qquad [\![\mathbf{A}]\!]^{\mathcal{M}} \text{ \_phi vs. } [\![\mathbf{A}]\!]$$

- condensed notation: $f(x+1)\pm 2\pi = g(x-1)\mp 2i$ (stands for 2 equations)
- ad hoc extensions: $\#(A \cup B) \leq \#A + \#B$ (exceptions for ∞)
- overt ambiguity: $\sin x/y$ vs. $\frac{\sin x}{y}$ vs. $\sin \frac{x}{y}$ vs. $-1 \leq \sin x/\pi \leq 1$
- size of the gaps varies with the intended readership and the space constraints.
- can be so substantial, that only a few specialists in the field can understand

The Notation/Context Problem

- idiosyncratic notations that are introduced, extended, discarded on the fly

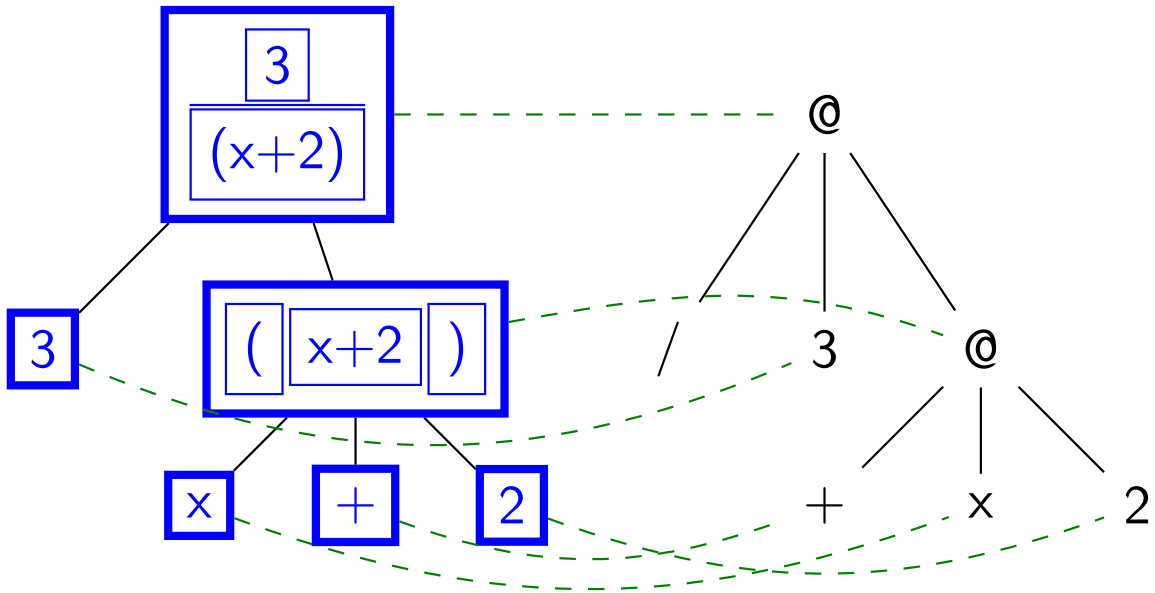
$$\lambda X_\alpha.X =_\alpha \lambda Y_\alpha.Y \hat{=} \mathbf{I}^\alpha$$

  meaning of α depends on context: object type vs. mnemonic vs. type label.
- even "standard notations" depend on the context, e.g. binomial coefficients: $\binom{n}{k}$, ${}_nC^k$, C^n_k, and C^k_n all mean the same thing: $\frac{n!}{k!(n-k)!}$ (cultural context)
- Notation scoping follows complex rules (notations must be introduced)
  - *We will write $\wp(S)$ for the set of subsets of S* (for the rest of the doc)
  - *We use the notation of [BrHa86], with the exception...* (by reference)
  - *Let S be a set and $f : S \to S$...* (scope local in definition)
  - *where w is the...* (scope local in preceding formula)
  - Book on group theory in Bourbaki series uses notation [Bou: Algebra]

Observation: Notation scoping is different from the one offered by TEX/LATEX

Parallel Markup e.g. in MathML I

- Idea: Combine the presentation and content markup and cross-reference

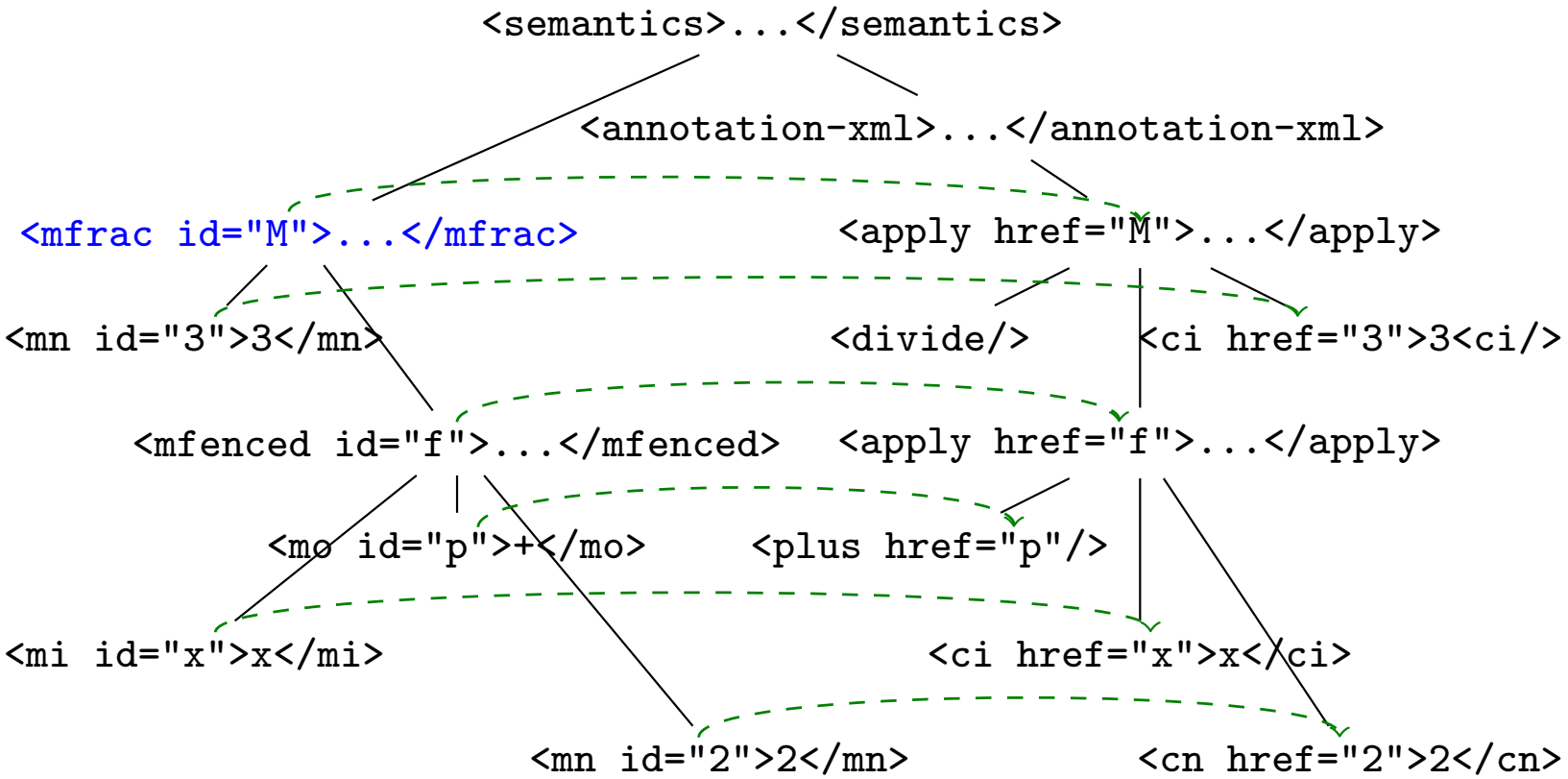


- use e.g. for semantic copy and paste.

(click on presentation, follow link and copy content)

Parallel Markup e.g. in MathML II

- Concrete Realization in MathML: `semantics` element with presentation as first child and content in `annotation-xml` child



STEM Corpora and Semantics Recovery

- Corpora at the presentation/structure level
  - The arXMLiv Corpus: arxmliv ca. 800k LaTeX documents from STEM discplines.
  - The ZBMath Corpus: ca. 3.3M abstracts/reviews for math articles in LaTeX.
  - PlanetMath.org: ca. 9k math encyclopedia articles in Planetary.
- Corpora at the content level
  - OpenMath/MathML Content Dictionaries: ca. 100 OpenMath documents designed as targets for semantic annotations.
  - The STC Corpus: ca. 4k semantically annotated course modules/problems in Planetary.
- Formal Corpora
  - The Mizar Corpus: ca. 100k formalized statements/proofs in Planetary.
  - The LATIN Logic Atlas: Meta-logical development of mathematical foundations.

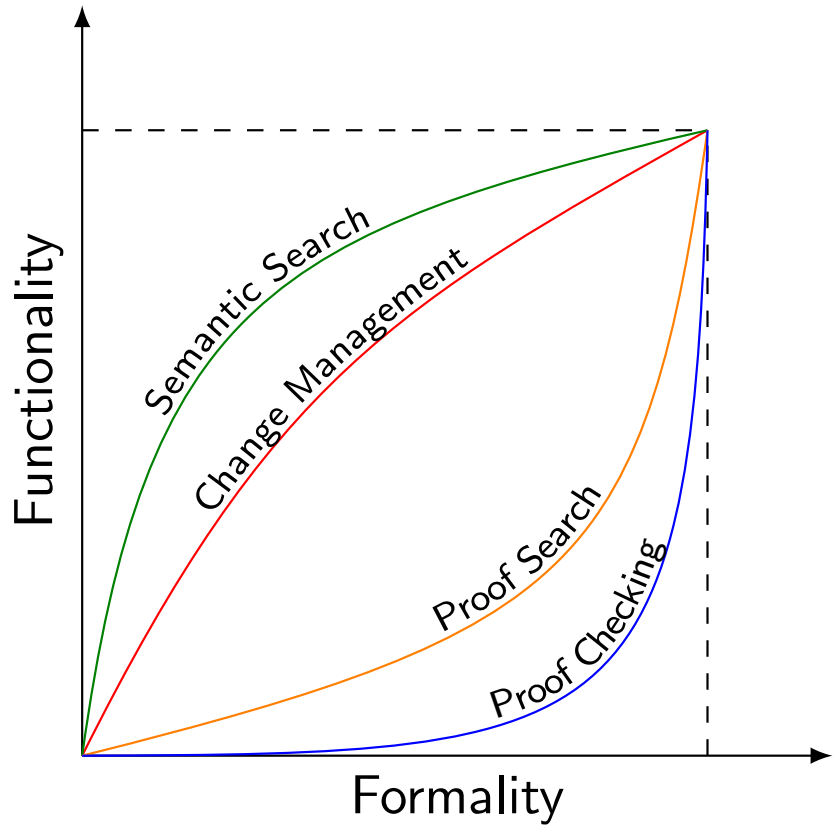
The Flexiformalist Program

The Flexiformalist Program (Details in [Koh13])

- The development of a regime of partially formalizing
  - mathematical knowledge into a modular ontology of mathematical theories (content commons), and
  - mathematical documents by semantic annotations and links into the content commons (semantic documents),
- The establishment of a software infrastructure with
  - a distributed network of archives that manage the content commons and collections of semantic documents,
  - semantic web services that perform tasks to support current and future mathematic practices
  - active document players that present semantic documents to readers and give access to respective
- the re-development of comprehensive part of mathematical knowledge and the mathematical documents that carries it into a flexiformal digital library of mathematics.

Flexible Levels of Formality/Semantics

- We are developing regimes of flexibly adding semantics (partial formalization)
- semantic services have distinct formalization thresholds
- start with the low-hanging fruit (80-20)



- Support Math Research by "global" services that cross the one-brain barrier
  - e.g. automatically finding representation theorems that transport theorems

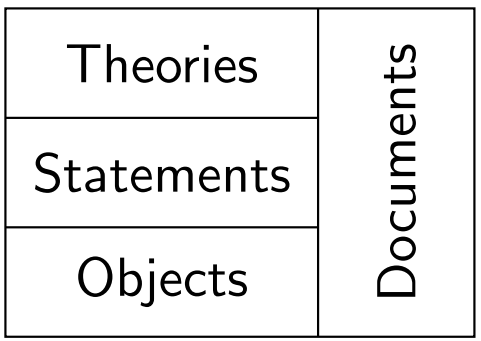
OMDoc: Open Mathematical Documents

OMDoc Philosophy

- OMDoc: Open Mathematical Documents: Representing Knowledge/Documents in Math/STEM.
- General features of knowledge representation languages (primitive/orthogonal?)

Feature	range		
aspect	functional	–	visual/narrative
paradigm	inferential	–	computational
grammar	natural	controlled	extensible
vocabulary	fixed	*vs.*	extensible
foundation	unspecified	fixed	expressible

- Domain Constraint for STEM Documents: Three levels:
  - object/phrase level, e.g. formulae
  - statement/paragraph level, e.g. definition, theorem, proof
  - theory/discourse level, theories, views, models
- OMDoc Philosophy: Interleave all features at all levels via parallel markup (general flexiformalization framework)



OMDoc in a Nutshell (three levels of modeling) [Koh06]

Formula level: OpenMath/C-MathML

- Objects as logical formulae
- symbol meaning by reference to theory level

```
<apply>
 <csymbol cd="ring">plus</c.>
 <csymbol cd="ring">zero</c.>
 <ci>N</ci>
</apply>
```

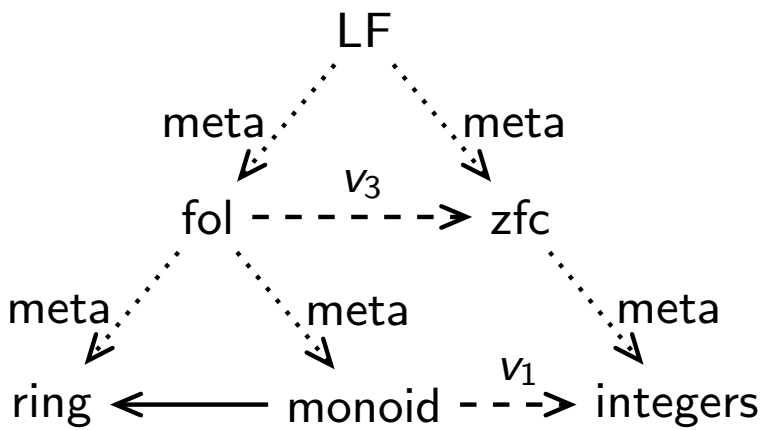
Statement level:

- Definition, Theorem, Proof, Example
- semantics via explicit forms and refs.
- parallel formal & natural language

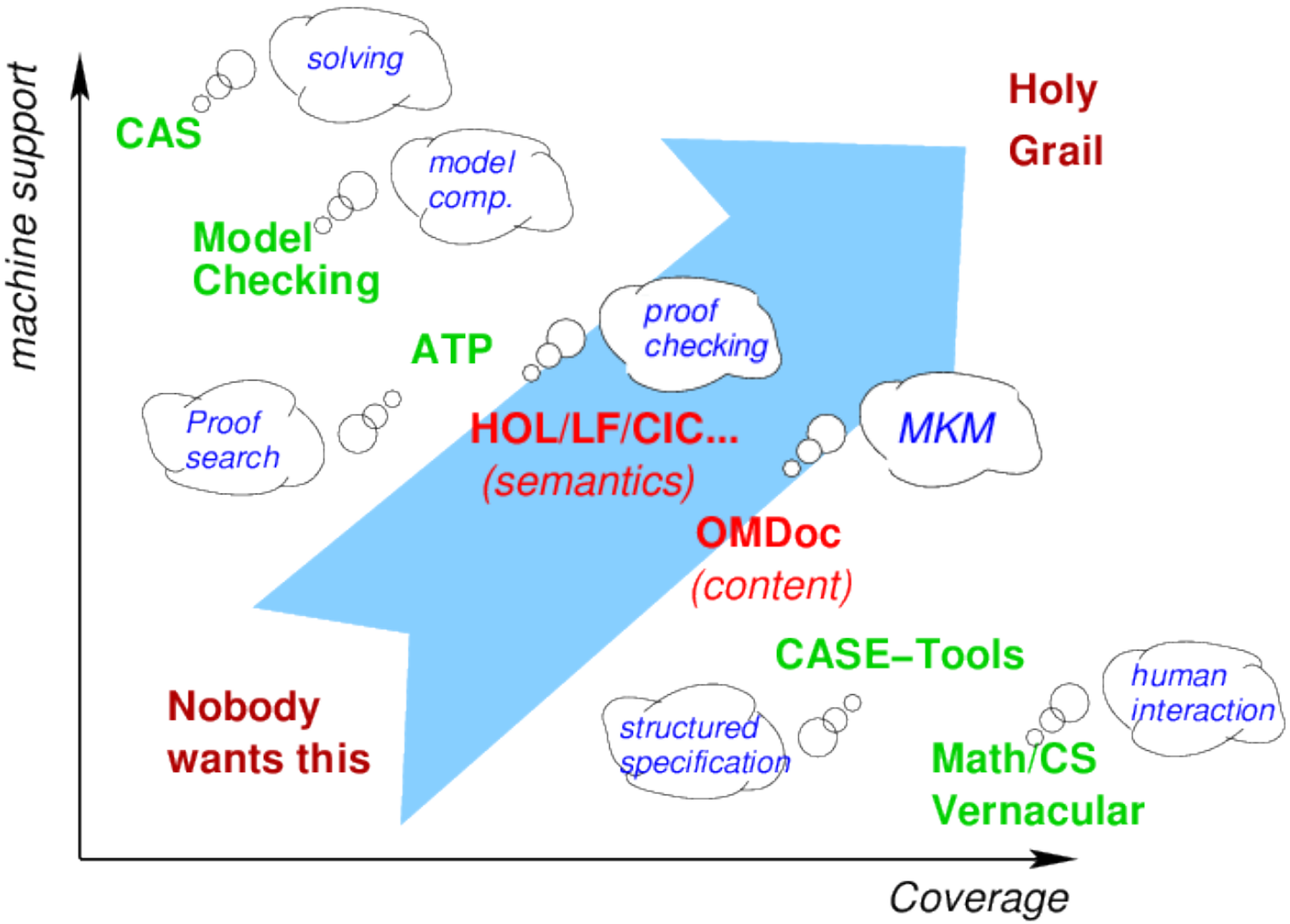
```
<defn for="plus" type="rec">
 <CMP>rec. eq. for plus</CMP>
 <FMP>X + 0 = X</FMP>
 <FMP>X + s(Y) = s(X + Y)</FMP>
</defn>
```

Module level: Theory Graph [RK13]

- inheritance via symbol-mapping
- views by proof-obligations
- logics as meta-theories (logic atlas)
- meta-logics as oracles for type/eq



Situating OMDoc: Math Knowledge Management




Andrea Asperti, Ferruccio Guidi, Claudio Sacerdoti Coen, Enrico Tassi, and Stefano Zacchiroli.
A content based mathematical search engine: Whelp.
In Jean-Christophe Filliâtre, Christine Paulin-Mohring, and Benjamin Werner, editors, *Types for Proofs and Programs, International Workshop, TYPES 2004, revised selected papers*, number 3839 in Lecture Notes in Computer Science, pages 17–32. Springer Verlag, 2006.


Michael Kohlhase.
OMDoc – *An open markup format for mathematical documents [Version 1.2]*.
Number 4180 in LNAI. Springer Verlag, August 2006.


Michael Kohlhase.
The flexiformalist manifesto.
In Andrei Voronkov, Viorel Negru, Tetsuo Ida, Tudor Jebelean, Dana Petcu, and Stephen M. Watt ane Daniela Zaharie, editors, *14th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012)*, pages 30–36, Timisoara, Romania, 2013.
in press.


Paul Libbrecht and Erica Melis.
Methods for Access and Retrieval of Mathematical Content in ActiveMath.

In N. Takayama and A. Iglesias, editors, *Proceedings of ICMS-2006*, number 4151 in LNAI, pages 331–342. Springer Verlag, 2006. http://www.activemath.org/publications/Libbrecht-Melis-Access-and-Retrieval-ActiveMath-ICMS-2006.pdf.

Jozef Misutka and Leo Galambos.
System description: Egomath2 as a tool for mathematical searching on wikipedia.org.
In James Davenport, William Farmer, Florian Rabe, and Josef Urban, editors, *Calculemus/MKM*, number 6824 in LNAI, pages 307–309. Springer Verlag, 2011.

Rajesh Munavalli and Robert Miner.
Mathfind: a math-aware search engine.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–735, New York, NY, USA, 2006. ACM Press.

Bruce R. Miller and Abdou Youssef.
Technical aspects of the digital library of mathematical functions.
Annals of Mathematics and Artificial Intelligence, 38(1-3):121–136, 2003.

Florian Rabe and Michael Kohlhase.
A scalable module system.

Manuscript, submitted to Information & Computation, 2013.