The Planetary System: Active Documents and a Web3.0 for Math.

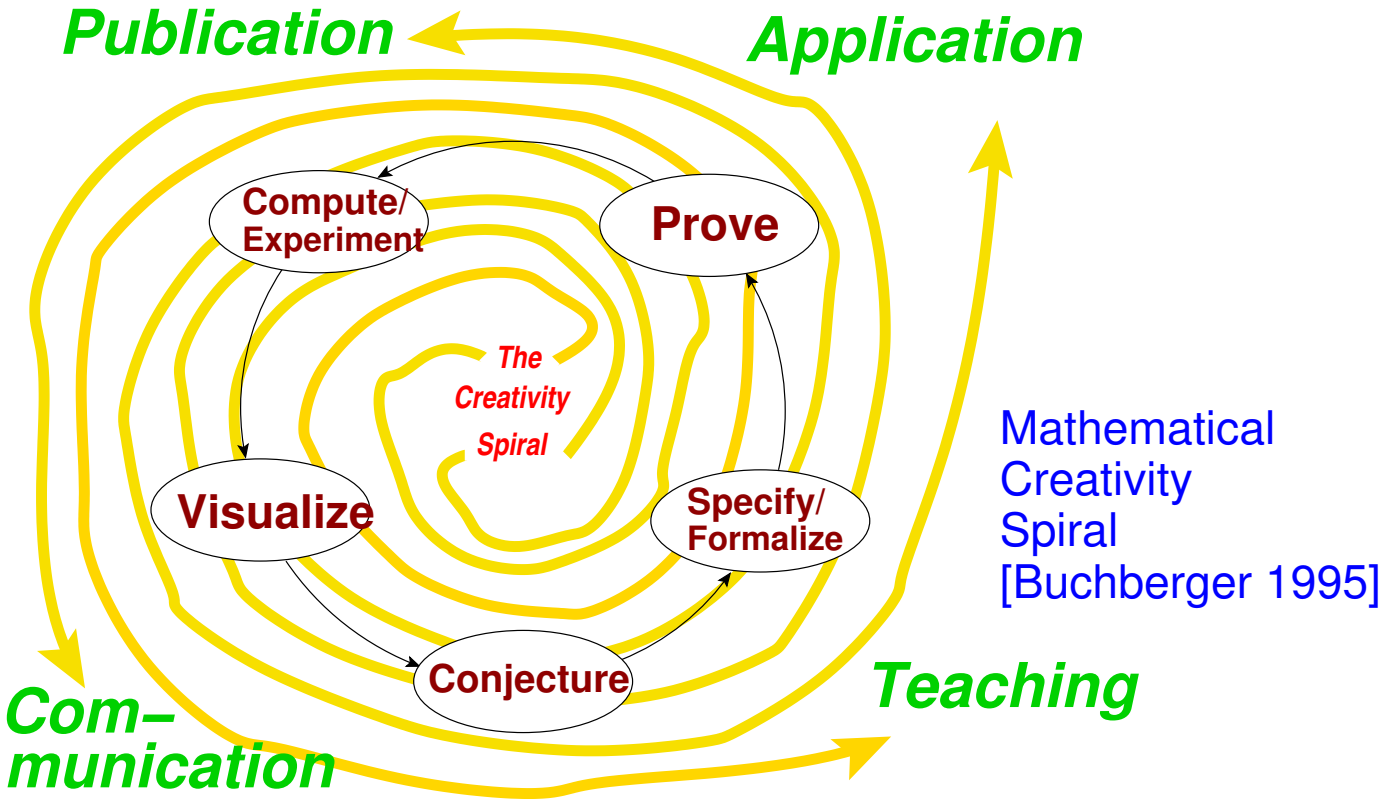
Michael Kohlhase

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

May 30. 2012, NIST

Introduction

The way we do math will change dramatically



- Every step will be supported by mathematical software systems
- Towards an infrastructure for web-based mathematics!

eMath 3.0: The Time is Ripe

- Background:
  - Web 2.0 is the term used for the "social Web" (tagging, blogs, wikis, facebook, ...)
  - The "Semantic Web" is a version of the Web, where humans & machines cooperate
  - Web 3.0 is the term used for the "social Semantic Web".
- We will apply these to eMath ( regular Math may or may not change)

eMath 3.0: The Time is Ripe

- Background:
  - Web 2.0 is the term used for the "social Web" (tagging, blogs, wikis, facebook, ...)
  - The "Semantic Web" is a version of the Web, where humans & machines cooperate
  - Web 3.0 is the term used for the "social Semantic Web".
- We will apply these to eMath ( regular Math may or may not change)
- Recent Improvements of the Math-on-the Web Environment:
  - MathML3 is out
  - MathML enabled in WebKit (in `Safari 5.1`, on the road for `Chrome`, `Konquerer`)
  - MathML is in HTML5 (without Namespaces though)
  - MathJax has reached 1.0 (Display MathML by JavaScript in many Browsers)

eMath 3.0: The Time is Ripe

- Background:
  - Web 2.0 is the term used for the "social Web" (tagging, blogs, wikis, facebook, ...)
  - The "Semantic Web" is a version of the Web, where humans & machines cooperate
  - Web 3.0 is the term used for the "social Semantic Web".
- We will apply these to eMath ( regular Math may or may not change)
- Recent Improvements of the Math-on-the Web Environment:
  - MathML3 is out
  - MathML enabled in WebKit (in `Safari 5.1`, on the road for `Chrome`, `Konquerer`)
  - MathML is in HTML5 (without Namespaces though)
  - MathJax has reached 1.0 (Display MathML by JavaScript in many Browsers)
- Recent Improvements of the Semantic Web:
  - RDF can be embedded into XML via RDFa (linked data export)
  - RDF querying via SPARQL (modulo OWL Ontologies) (semantic search)
  - OMDoc as a mathematical Ontology format (modularity, documentation, full Math)

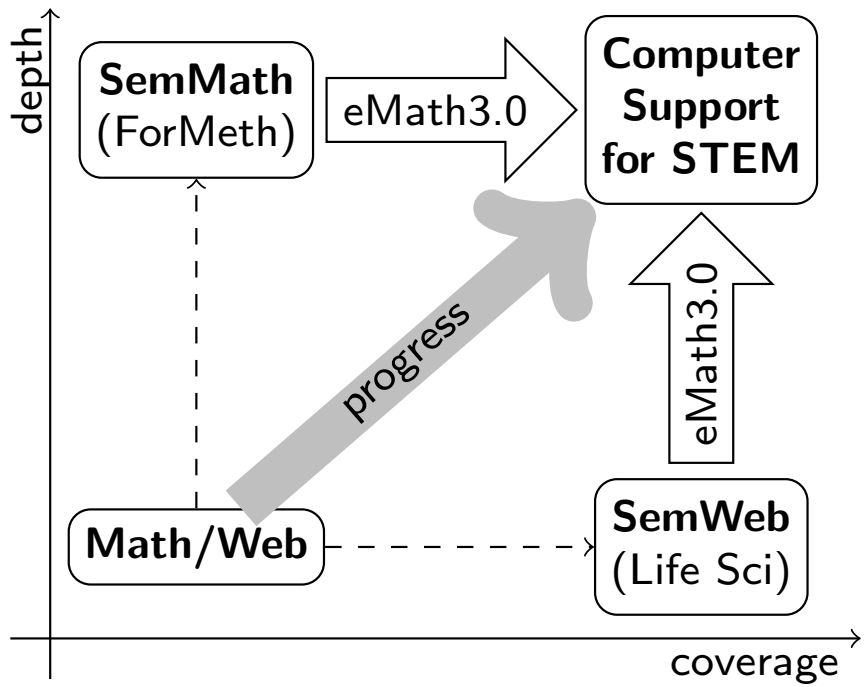
eMath 3.0: The Time is Ripe

- Background:
  - Web 2.0 is the term used for the "social Web" (tagging, blogs, wikis, facebook, ...)
  - The "Semantic Web" is a version of the Web, where humans & machines cooperate
  - Web 3.0 is the term used for the "social Semantic Web".
- We will apply these to eMath ( regular Math may or may not change)
- Recent Improvements of the Math-on-the Web Environment:
  - MathML3 is out
  - MathML enabled in WebKit (in `Safari 5.1`, on the road for `Chrome`, `Konquerer`)
  - MathML is in HTML5 (without Namespaces though)
  - MathJax has reached 1.0 (Display MathML by JavaScript in many Browsers)
- Recent Improvements of the Semantic Web:
  - RDF can be embedded into XML via RDFa (linked data export)
  - RDF querying via SPARQL (modulo OWL Ontologies) (semantic search)
  - OMDoc as a mathematical Ontology format (modularity, documentation, full Math)
- Overview over the talk:
  - MathML3 brings more semantics (strict content Math, elementary Math)
  - integrating MathML/LaTeX into the Web 2.0
  - A LaTeX-based Semantic Web for Mathematics

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...)



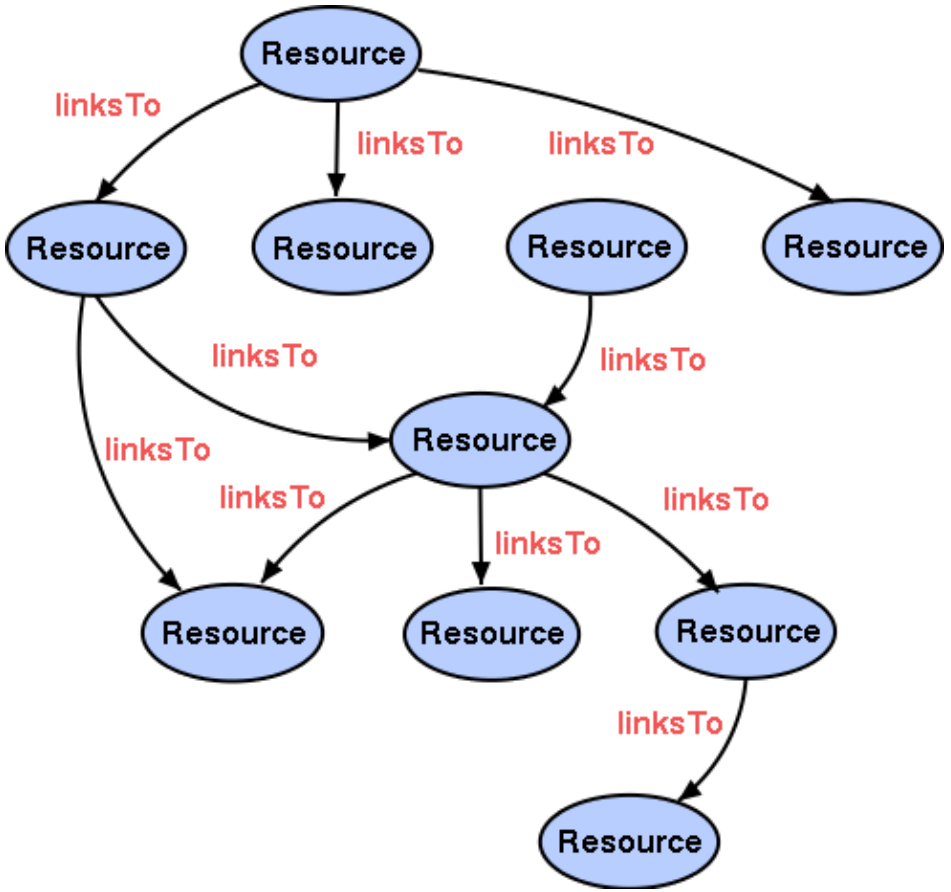
- ⚠ We use Math as a test tube for STEM (Science, Tech, Eng, & Math) ⚠

Foundations

The Semantic Web

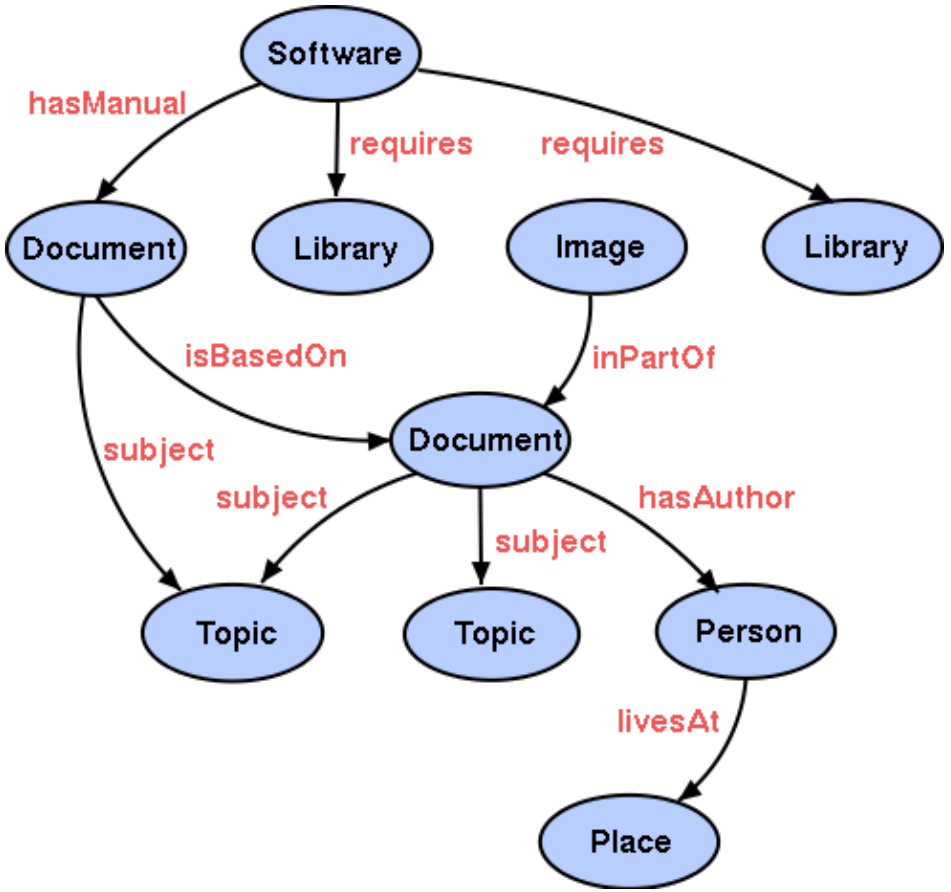
The Current Web

- Resources: identified by URI's, untyped
- Links: `href`, `src`, ... limited, non-descriptive
- User: Exciting world - semantics of the resource, however, gleaned from content
- Machine: Very little information available - significance of the links only evident from the context around the anchor.



The Semantic Web

- Resources: Globally Identified by URI's or Locally scoped (Blank), Extensible, Relational
- Links: Identified by URI's, Extensible, Relational
- User: Even more exciting world, richer user experience
- Machine: More processable information is available (Data Web)
- Computers and people: Work, learn and exchange knowledge effectively



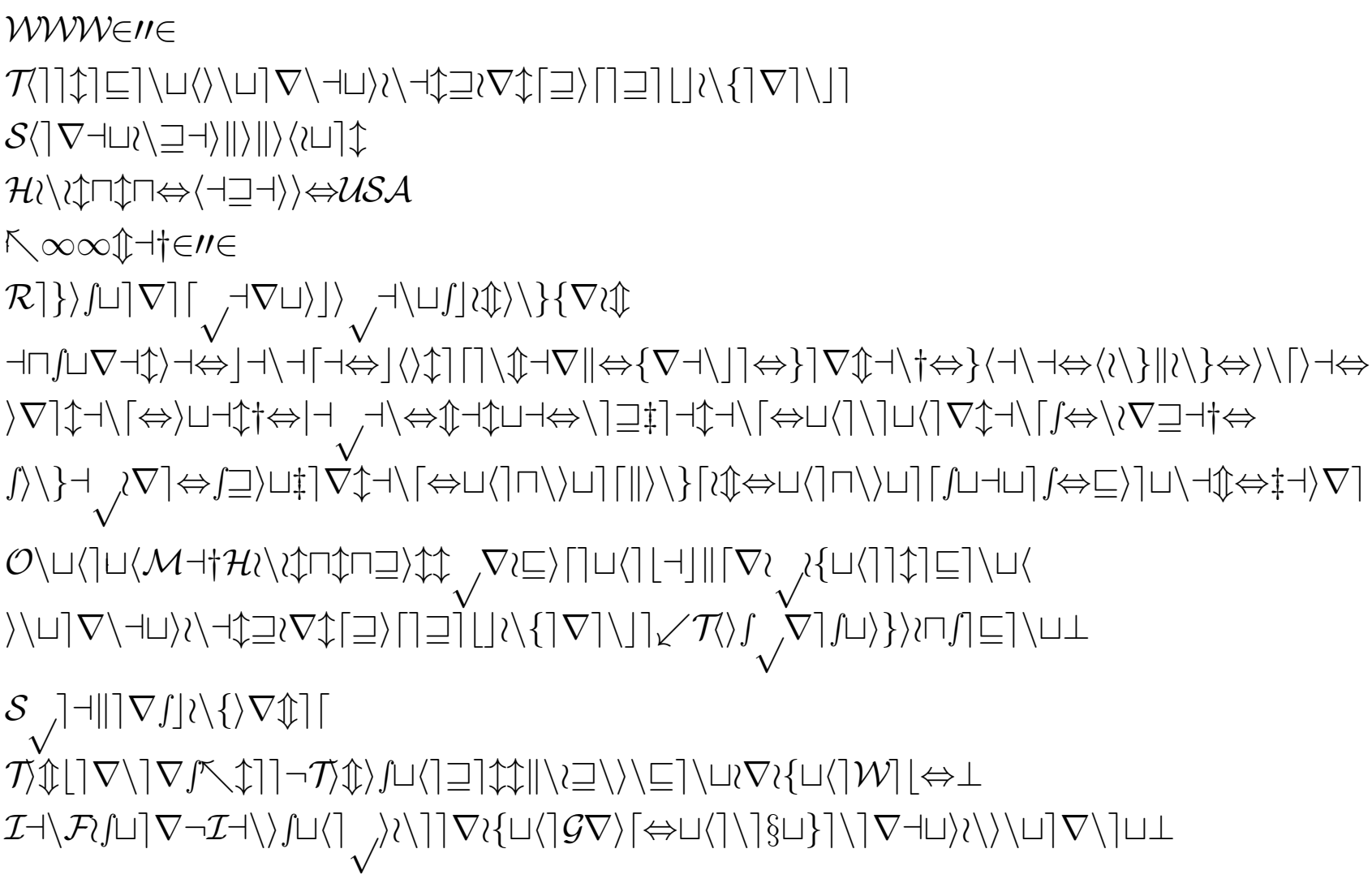
What is the Information a User sees?

WWW2002
The eleventh international world wide web conference
Sheraton waikiki hotel
Honolulu, hawaii, USA
7-11 may 2002
1 location 5 days learn interact

Registered participants coming from
australia, canada, chile denmark, france, germany, ghana, hong kong, india, ireland, italy, japan, malta, new zealand, the netherlands, norway, singapore, switzerland, the united kingdom, the united states, vietnam, zaire

On the 7th May Honolulu will provide the backdrop of the eleventh international world wide web conference. This prestigious event ?
Speakers confirmed
Tim Berners-Lee: Tim is the well known inventor of the Web, ?
Ian Foster: Ian is the pioneer of the Grid, the next generation internet ?

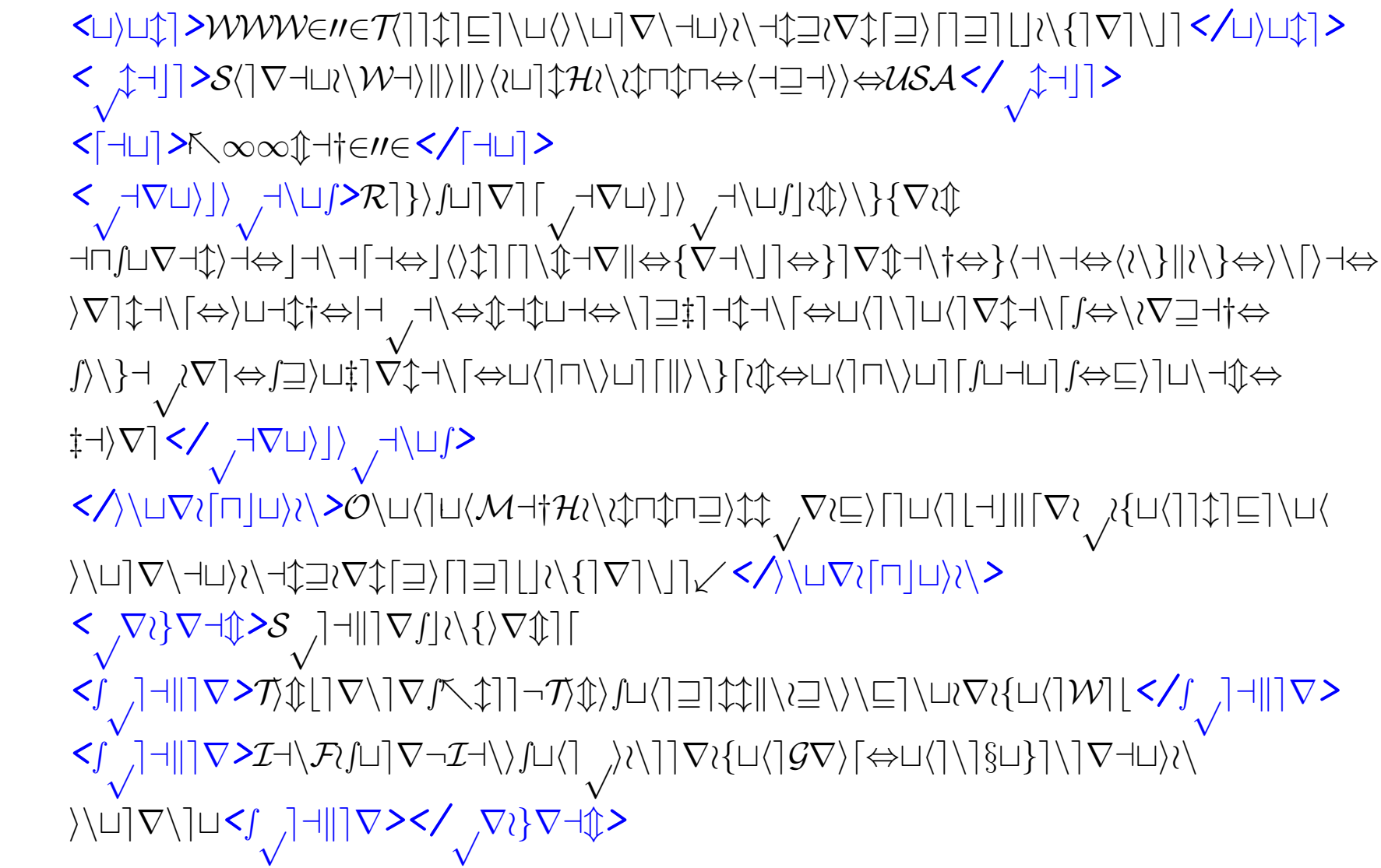
What the machine sees



Solution: XML markup with "meaningful" Tags

<title>WWW∈//∈T⟨⌉⌉↕⌉⊑⌉\⊔⟨⟩\⊔⌉∇\⊣⊔⟩ɾ\⊣↕⊒ɾ∇↕⌈⊒⟩⌈⌉⊒⌉⌊⌋ɾ\{⌉∇⌉\⌋⌉</title>
<place>S⟨⌉∇⊣⊔ɾ\W⊣⟩‖⟩‖⟩⟨ɾ⊔⌉↕Hɾ\ɾ↕⊓↕⊓⇔⟨⊣⊒⊣⟩⟩⇔USA</place>
<date>↖∞∞⇳⊣†∈//∈</date>
<participants>R⌉}⟩∫⊔⌉∇⌉⌈ √⊣∇⊔⟩⌋⟩ √⊣\⊔∫⌋ɾ⇳⟩\}{∇ɾ⇳
⊣⊓∫⊔∇⊣↕⟩⊣⇔⌋⊣\⊣⌈⊣⇔⌋⟨⟩↕⌉⌈⌉\⇳⊣∇‖⇔{∇⊣\⌋⌉⇔}⌉∇⇳⊣\†⇔}⟨⊣\⊣⇔⟨ɾ\}‖ɾ\}⇔⟩\⌈⟩⊣⇔
⟩∇⌉↕⊣\⌈⇔⟩⊔⊣⇳†⇔|⊣ √⊣\⇔⇳⊣⇳⊔⊣⇔\⌉⊒‡⌉⊣⇳⊣\⌈⇔⊔⟨⌉\⌉⊔⟨⌉∇↕⊣\⌈∫⇔\ɾ∇⊒⊣†⇔
∫⟩\}⊣ √ɾ∇⌉⇔∫⊒⟩⊔‡⌉∇↕⊣\⌈⇔⊔⟨⌉⊓\⟩⊔⌉⌈‖⟩\}⌈ɾ⇳⇔⊔⟨⌉⊓\⟩⊔⌉⌈∫⊔⊣⊔⌉∫⇔⊑⟩⌉⊔\⊣⇳⇔
‡⊣⟩∇⌉</participants>
</introduction>O\⊔⟨⌉⊔⟨M⊣†Hɾ\ɾ↕⊓↕⊓⊒⟩↕↕ √∇ɾ⊑⟩⌈⌉⊔⟨⌉⌊⊣⌋‖⌈∇ɾ √ɾ{⊔⟨⌉⌉↕⌉⊑⌉\⊔⟨
⟩\⊔⌉∇\⊣⊔⟩ɾ\⊣↕⊒ɾ∇↕⌈⊒⟩⌈⌉⊒⌉⌊⌋ɾ\{⌉∇⌉\⌋⌉↙</introduction>
<program>S √⌉⊣‖⌉∇∫⌋ɾ\{⟩∇⇳⌉⌈
<speaker>T⇳⇳⌊⌉∇\⌉∇∫↖↕⌉⌉¬T⇳⇳⟩∫⊔⟨⌉⊒⌉↕↕‖\ɾ⊒\⟩\⊑⌉\⊔ɾ∇ɾ{⊔⟨⌉W⌉⌊</speaker>
<speaker>I⊣\Fɾ∫⊔⌉∇¬I⊣\⟩∫⊔⟨⌉ √⟩ɾ\⌉⌉∇ɾ{⊔⟨⌉G∇⟩⌈⇔⊔⟨⌉\⌉§⊔}⌉\⌉∇⊣⊔⟩ɾ\
⟩\⊔⌉∇\⌉⊔<speaker></program>

What the machine sees of the XML



Need to add "Semantics"

- External agreement on meaning of annotations E.g., Dublin Core
  - Agree on the meaning of a set of annotation tags
  - Problems with this approach: Inflexible, Limited number of things can be expressed

Need to add "Semantics"

- External agreement on meaning of annotations E.g., Dublin Core
  - Agree on the meaning of a set of annotation tags
  - Problems with this approach: Inflexible, Limited number of things can be expressed
- Use Ontologies to specify meaning of annotations
  - Ontologies provide a vocabulary of terms
  - New terms can be formed by combining existing ones
  - Meaning (semantics) of such terms is formally specified
  - Can also specify relationships between terms in multiple ontologies

Need to add "Semantics"

- External agreement on meaning of annotations E.g., Dublin Core
  - Agree on the meaning of a set of annotation tags
  - Problems with this approach: Inflexible, Limited number of things can be expressed
- Use Ontologies to specify meaning of annotations
  - Ontologies provide a vocabulary of terms
  - New terms can be formed by combining existing ones
  - Meaning (semantics) of such terms is formally specified
  - Can also specify relationships between terms in multiple ontologies
- Inference with annotations and ontologies (get out more than you put in!)
  - Standardize annotations in RDF [KC04] or RDFa [BAHS] and ontologies on OWL [w3c09]
  - Harvest RDF and RDFa in to a triplestore or OWL
  -

MathML: Presentation and Content of Mathematical Formulae

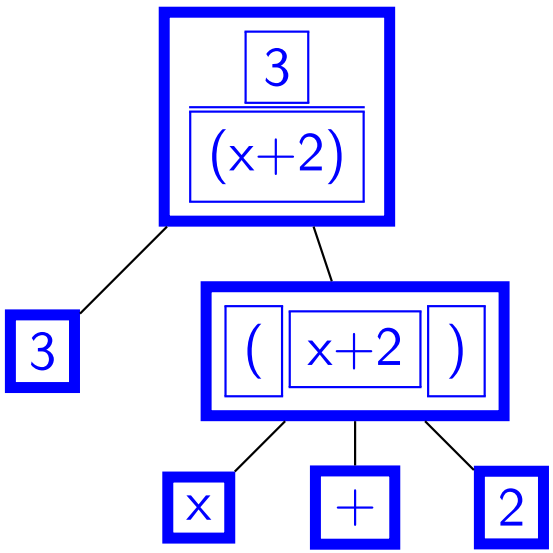
Representation of Formulae as Expression Trees

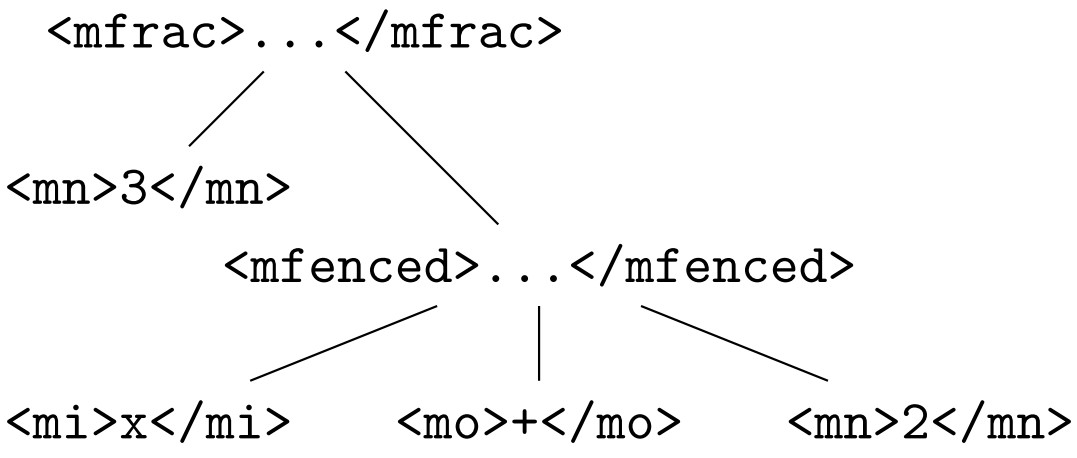
- Mathematical Expressions are build up as expression trees
  - of layout schemata in Presentation-MathML
  - of functional subexpressions in Content-MathML
- Example: $\frac{3}{x+2}$

```
<mfrac>
  <mn>3</mn>
  <mfenced>
     <mi>x</mi>
     <mo>+</mo>
     <mn>2</mn>
   </mfenced>
</mfrac>
```

```
<apply>
  <divide/>
  <cn>3</cn>
  <apply>
    <plus/>
    <ci>x</ci>
    <cn>2</cn>
  </apply>
</apply>
```

Layout Schemata and the MathML Box model





Content Mathml: Expression Trees in Prefix Notation

- Prefix Notation saves parentheses (so does postfix, BTW)

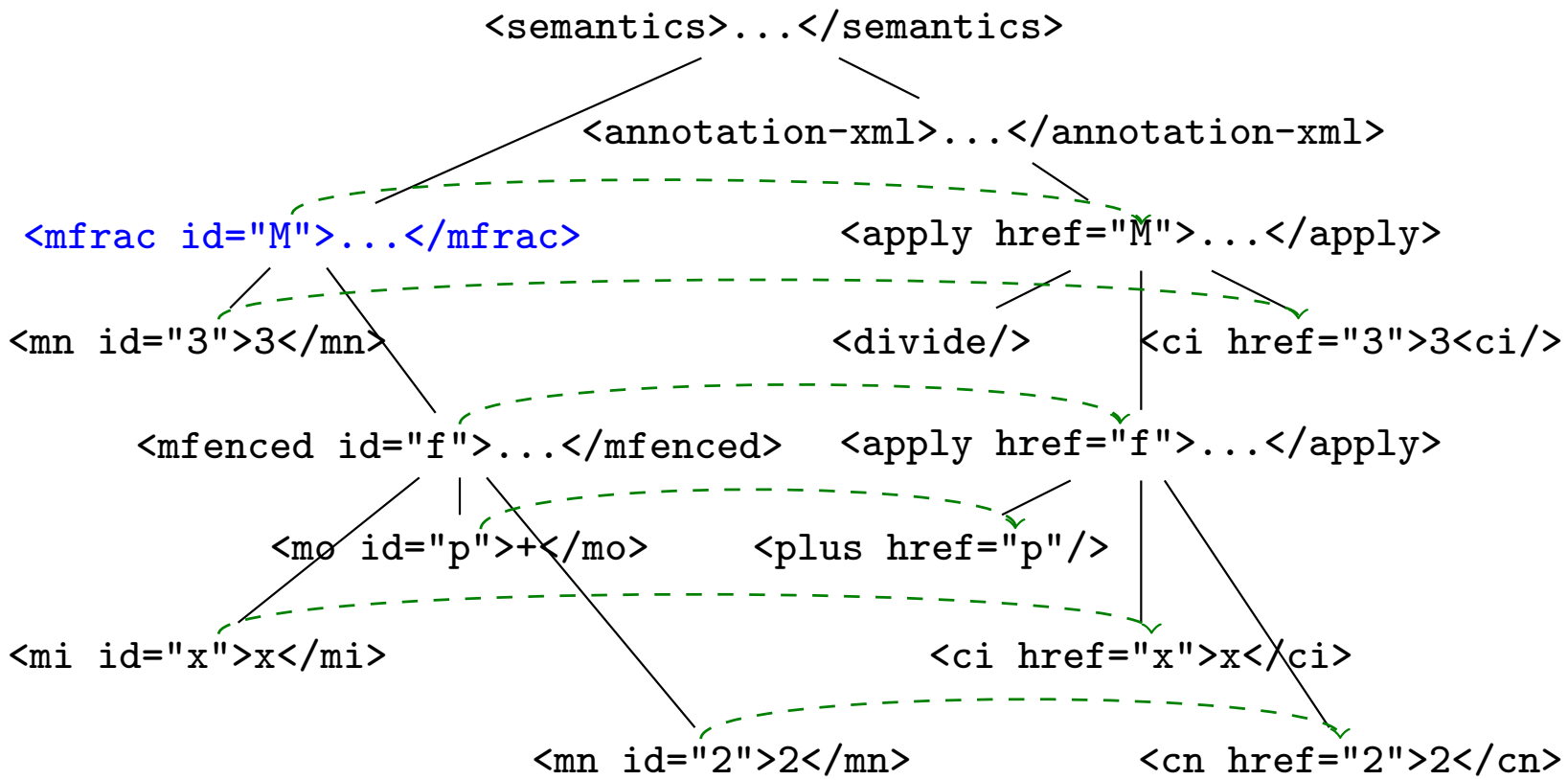
$(x-y)/2$	$x-(y/2)$

`<apply>`
`  <divide/>`
`  <apply>`
`  <minus/>`
`  <ci>x</ci>`
`  <ci>y</ci>`
`  </apply>`
`  <cn>2</cn>`
`</apply>`
	`<apply>`
	`  <minus/>`
	`  <ci>x</ci>`
	`  <apply>`
	`  <divide/>`
	`  <ci>y</ci>`
	`  <cn>2</cn>`
	`  </apply>`
	`</apply>`

- Function Application: `<apply>function arg1 ... argn </apply>`
- Operators and Functions: $\sim$ 100 empty elements `<sin/>`, `<plus/>`, `<eq/>`, `<compose/>`,...
- Token elements: `ci`, `cn` (identifiers and numbers)
- Extra Operators: `<csymbol definitionURL="...">...</csymbol>`

Parallel Markup e.g. in MathML

- Combine the presentation and content markup in one tree and crosss-reference



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

(click on presentation, follow link and copy content)

Mixing Presentation and Content MathML

```
<semantics>
 <mrow>
  <mrow><mo>(</mo><mi>a</mi> <mo>+</mo> <mi>b</mi><mo>)</mo></mrow>
  <mo>&InvisibleTimes;</mo>
  <mrow><mo>(</mo><mi>c</mi> <mo>+</mo> <mi>d</mi><mo>)</mo></mrow>
 </mrow>
 <annotation-xml encoding="MathML-Content">
  <apply><times/>
   <apply><plus/><ci>a</ci> <ci>b</ci></apply>
   <apply><plus/><ci>c</ci> <ci>d</ci></apply>
  </apply>
 </annotation-xml>
 <annotation-xml encoding="openmath">
  <OMA><OMS cd="arith1" name="times"/>
   <OMA><OMS cd="arith1" name="plus"/><OMV name="a"/><OMV name="b"/></OMA>
   <OMA><OMS cd="arith1" name="plus"/><OMV name="c"/><OMV name="d"/></OMA>
  </OMA>
 </annotation-xml>
</semantics>
```

Converting the arXiv

The arXMLiv Project: arXiv to semantic XML

- Idea: Develop a large corpus of knowledge in OMDoc/PhysML
  - to get around the chicken-and-egg problem of MKM
  - corpus-linguistic methods for semantics recovery (linguists interested)
- **Definition 1 (The Cornell Preprint arXiv)** (http://www.arxiv.org)
  Open access to ca. 700K e-prints in Physics, Mathematics, Computer Science and Quantitative Biology.
- **Definition 2 (The arXMLiv Project)** (http://arxmliv.kwarc.info)
  - use Bruce Miller's LaTeXML to transform to XHTML+MathML
  - extend to LaTeXML daemon (RESTful web service) (http://latexml.mathweb.org)
  - we have an automated, distributed build system (ca. 2 CPU-years)
  - create ca. 12K LaTeXML binding files (8 Jacobs students help)
  - use MathWebSearch to index XML version (realistic search corpus)
- More semantic information will enable more added-value services, e.g.
  - filter hits by model assumptions (expanding, stationary, or contracting universe)
  - use linguistic techniques to add the necessary semantics

Why reimplement the TEX parser? I

- Problem: The TEX parser can change the tokenizer while at runtime (`\catcode`)
- **Example 3 (Obfuscated TEX)** David Carlisle posted the following, when someone claimed that word counting is simple in TEX/LATEX

```
\let~\catcode~‘76~‘A13~‘F1~‘j00~‘P2jdefA71F~‘7113jdefPALLF
PA’’FwPA;;FPAZZFLaLPA//71F71iPAHHFLPAzzFenPASSFthP;A$$FevP
A@@FfPARR717273F737271P;ADDFRgniPAWW71FPATTFvePA**FstRsamP
AGGFRruoPAqq71.72.F717271PAYY7172F727171PA??Fi*LmPA&&71jfi
Fjfi71PAVVFjbigskipRPWGAUU71727374 75,76Fjpar71727375Djifx
:76jelse&U76jfiPLAKK7172F7117271PAXX71FVLnOSeL71SLRyadR@oL
RrhC?yLRurtKFeLPFovPgaTLtReRomL;PABB71 72,73:Fjif.73.jelse
B73:jfiXF71PU71 72,73:PWs;AMM71F71diPAJJFRdriPAQQFRsreLPAI
I71Fo71dPA!!FRgiePBt’el@ lTLqdrYmu.Q.,Ke;vz vzLqpip.Q.,tz;
;Lql.IrsZ.eap,qn.i. i.eLlMaesLdRcna,;!;h htLqm.MRasZ.ilk,%
s$;z zLqs’.ansZ.Ymi,/sx ;LYegseZRyal,@i;@ TLRlogdLrDsW,@;G
LcYlaDLbJsW,SWXJW ree @rzchLhzsW,;WERcesInW qt.’oL.Rtrul;e
doTsW,Wk;Rri@stW aHAHHFndZPpqar.tridgeLinZpe.LtYer.W,:jbye
```

When formatted by TeX, this leads to the full lyrics of “The twelve days of christmas”. When formattet by LATEXML, it gives

Why reimplement the TEX parser? II

```
<song>
  <verse>
    <line>On the first day of Christmas my true love gave to me</line>
    <line>a partridge in a pear tree.</line>
  </verse>
  <verse>
    <line>On the second day of Christmas my true love gave to me</line>
    <line>two turtle doves</line>
    <line>and a partridge in a pear tree.</line>
  </verse>
  <verse>
    <line>On the third day of Christmas my true love gave to me</line>
    <line>three french hens</line>
    <line>two turtle doves</line>
    <line>and a partridge in a pear tree.</line>
  </verse>
  <verse>
    <line>On the fourth day of Christmas my true love gave to me</line>
    <line>four calling birds</line>
    <line>three french hens</line>
    <line>two turtle doves</line>
    <line>and a partridge in a pear tree.</line>
  </verse>
  ...
```

Why reimplement the TEX parser? III

- But the real reason is: that we can take advantage of the semantics in the LATEX.
- LATEXML does not need to expand macros, we can tell it about XML equivalents.
- **Example 4 (Recovering the Semantics of Proofs)**
  Add the following magic incantation to `amsthm.sty.ltxml` (LATEXML binding)

```
DefEnvironment('{proof}',"<xhtml:div class='proof'>#body</xhtml:div>");
```

The `arXMLiv` approach: Try to cover most packages and classes in the arXiv
(Jacobs undergrads' intro to research)

Future Plans for arXMLiv

- State: LaTeX-to-XHTML+MathML Format Conversion works (65% success)
- Over the summer: Bump up success rate to 75%, daily downloads, web site, instrumentation,...
- Soon: Integrate user-level quality control (integrate JS feedback into html)
- starting Fall: Extend post-processing by linguistic methods for semantic analysis
  - build semantics blackboard/database for linguistic information (rdf triples)
  - extend build system for arbitrary XML2BB processes
  - invite the linguists over (they leave semantics results in BB)
  - harvest the semantics BB to get OMDoc representations

Current and Possible Applications

- the arxmliv build system `http://arxmliv.kwarc.info`
- the transformation web service `http://tex2xml.kwarc.info`
- LaTeXML daemon to avoid `perl` and LaTeX startup times (Deyan Ginev)
  - keep LaTeXML alive as a daemon that can process multiple files/fragments (patch memory leaks)
  - a LaTeXML client just passes files/fragments along ($\frac{10}{s}$ to $\frac{100}{s}$)
- embedding/editing LaTeX in web pages `http://tex2xml.kwarc.info/test`
- a MathML version of the arXiv allows vision-impared readers to understand the texts
- generalization search (need to know sentence structure for detecting universal variables)
- semantic search by academic discipline or theory assumption (need discourse structure)
- development of scientific vocabularies (over the past 18 years; drink from the source)

Planetary: An Integrated Platform for eMath3.0

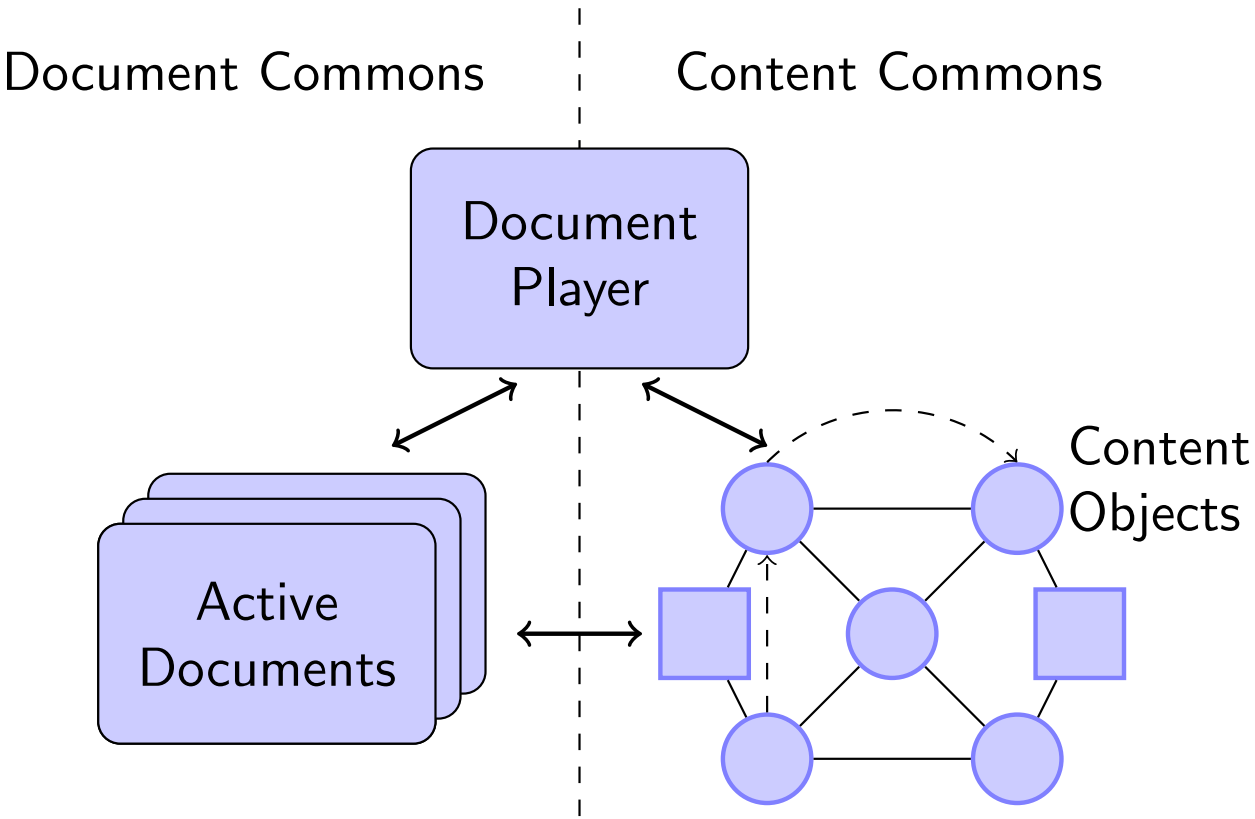
Planetary: A Social Semantic eScience System

The PLANETARY System

- The PLANETARY system is a Web 3.0 system for semantically annotated document collections in Science, Technology, Engineering and Mathematics (STEM).
- Web 3.0 stands for extension of the Social Web with Semantic Web/Linked Open Data technologies.
- documents published in the PLANETARY system become flexible, adaptive interfaces to a content commons of domain objects, context, and their relations.
- PLANETARY is based on the Active Documents Paradigm (see next)
- **Example 5 (Example installments)**
  - `arxivdemo.mathweb.org` (presentation/structural Level: arXiv)
  - `panta.kwarc.info` (semantic level: PantaRhei course system)
  - `logicatlas.omdoc.org` (fully formal level: Logic Representations)
  - `planetbox.kwarc.info` (Technology Sandbox)
- The PLANETARY system is finalist in the Elsevier Executable Papers Challenge.

The Active Documents Paradigm

- **Definition 6** 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 tha embeds services to make documents executable.



- **Example 7** Services can be program (fragment) execution, computation, visualization, navigation, information aggregation and information retrieval

OMDoc in a Nutshell (three levels of modeling)

Formula level: OpenMath/C-MathML

- Objects as logical formulae
- semantics by ref. to theory level

```
<OMA>
 <OMS cd="arith1" name="plus"/>
 <OMS cd="nat" name="zero"/>
 <OMV name="N"/>
</OMA>
```

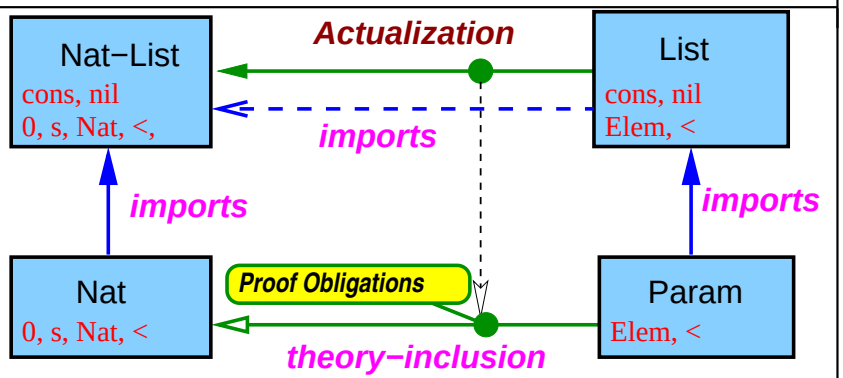
Statement level:

- Definition, Theorem, Proof, Ex.
- semantics explicit forms and refs.

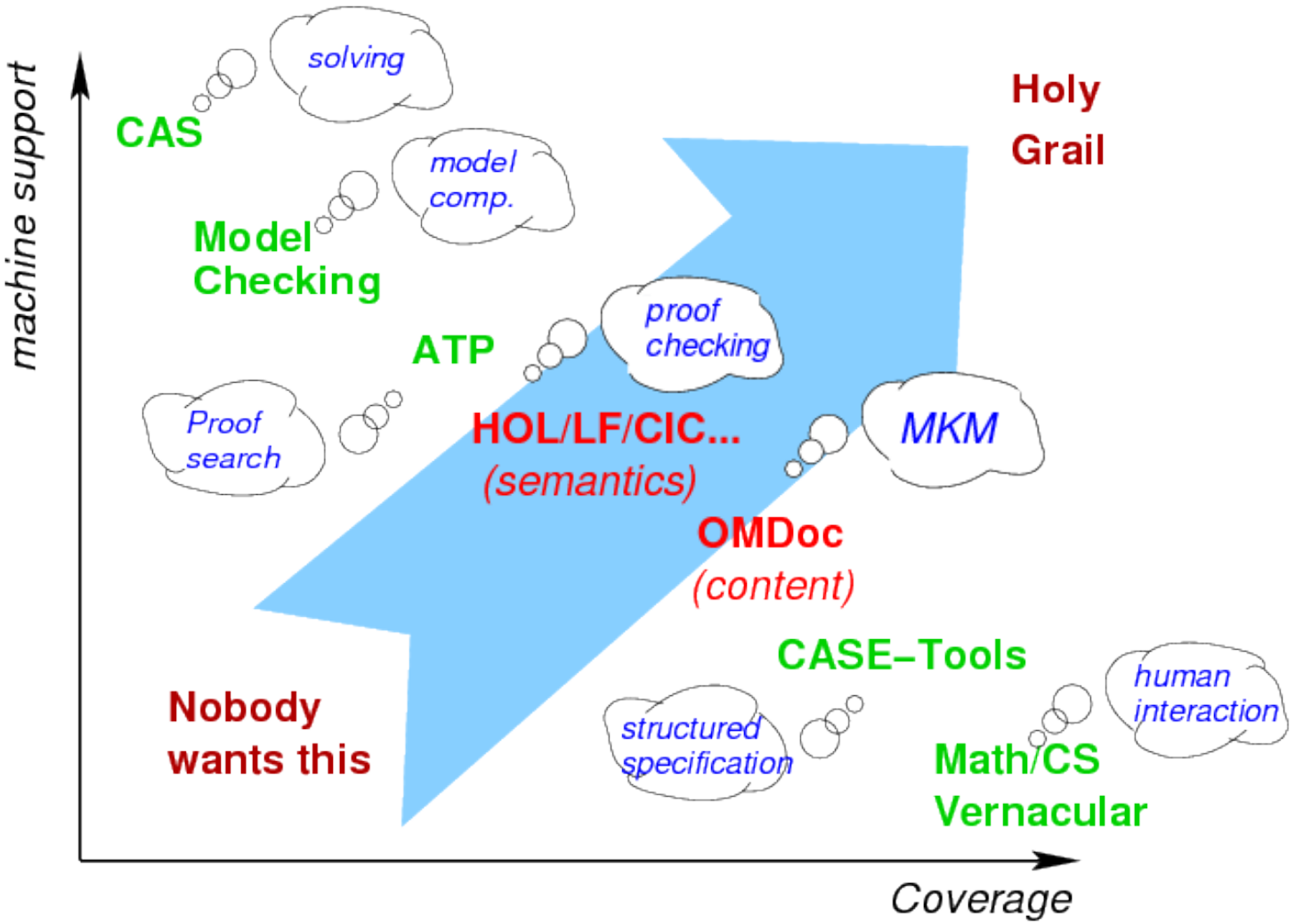
```
<defn for="plus" type="rec">
 <CMP>rec. eq. for plus</CMP>
 <FMP>X + 0 = X</FMP>
 <FMP>X + s(Y) = s(X + Y)</FMP>
</defn>
```

Theory level: Development Graph

- inheritance via symbol-mapping
- theory-inclusion by proof-obligations
- local (one-step) vs. global links



Situating OMDoc: Math Knowledge Management



sTEX: A Semantic Variant of LaTeX

TEX/LATEX as MKM Format: 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

TEX/LATEX as MKM Format: 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.)

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

- 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 sTEX approach

- The reconstruction and the notation/context problem have to be solved to turn or translate TEX/LATEX into a MKM format
- Problem: This is impossible in the general case (AI-hard)
- Idea: Enable the author to make structure explicit and disambiguate meanings
  - use the TEX macro mechanism for this (well established)
  - the author knows the semantics best (at least she understands)
  - the burden is is alleviated by manageability savings (MKM on TEX/LATEX)
- **Definition 8 (sTEX Approach)** Semantic pre-loading of TEX/LATEX documents.
  - Introduce semantic macros: e.g. `\union{a,b,c}` $\rightsquigarrow a \cup b \cup c$
  - Mark up discourse structure: (largely invisible)
    e.g. `\begin{proof}[id=Wiles,for=Fermat]...\end{sproof}`
  - Generate PDF and XML from that (via LATEXML [Miller])

sTeX Modules help with the Notation/Context Problem

- Note: *the context of notations coincides with the context of the concepts they denote*
- Idea: Use the theory structure for notational contexts
  - The scoping rules of TeX/LaTeX follow a hierarchical model:
  - a TeX macro is either globally defined or defined exactly inside the group induced by the TeX/LaTeX curly braces hierarchy.
- Solution: provide explicit grouping for scope with inheritance.
  - new sTeX environment `module`,
  - new macro definition `\symdef`, scoped in `module`
  - specify the inheritance of `\symdef`-macros in `module` explicitly
  - `\symdef`-macros are undefined unless in home module or inherited.

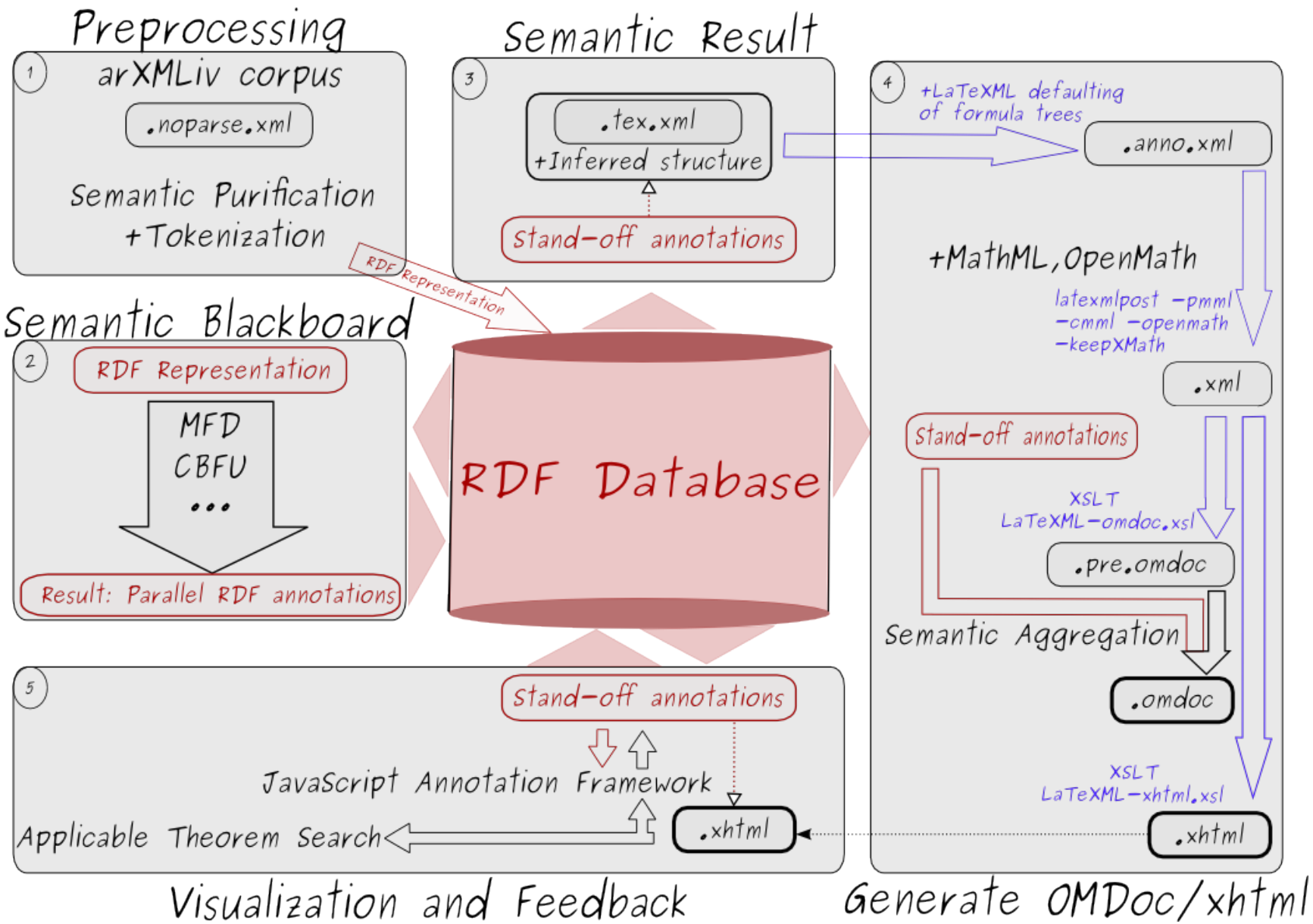
sTEX Modules: Example

```
\begin{module}[id=pairs]\symdef{pair}[2]{\langle#1,#2\rangle} ...\end{module}

\begin{module}[id=sets]
  \symdef{member}[2]{#1\in #2} % set membership
  \symdef{mmember}[2]{#1\in #2} ... % aggregated set membership
\end{module}

\begin{module}[id=setoid]
  \importmodule{pairs}
  \importmodule{sets}
  \symdef{sset}{\mathcal{S}} % the base set
  \symdef{sopa}{\circ} % the operation symbol
  \symdef{sop}[2]{(#1\sopa #2)} % the operation applied
  \begin{definition}[id=setoid.def]
    A structure $\pair\sset\sopa$ is called a \defi{setoid}, if $\sset$ is closed under
    $\sopa$, i.e. if $\member{\sop{a}{b}}\sset$ for all $\mmember{a,b}\sset$.
  \end{definition}
\end{module}

\begin{module}[id=semigroup]
  \importmodule{setoid}
  \begin{definition}[id=monoid.def]
    A \trefi[setoid]{setoid} $\pair\sset\sopa$ is called a \defi{monoid}, if $\sopa$ is
    associative on $\sset$, i.e. if $\sop{a}{\sop{b}{c}}=\sop{\sop{a}{b}}{c}$ for all
    $\mmember{a,b,c}\sset$.
  \end{definition}
\end{module}
```

The Result of the Example

- Empirically: Explicit module structure
  - is a little overhead (can be automated/supported by IDE [JK10])
  - more semantic/portable (but I might be brainwashed)
- In our case study: 320 slides, 160 modules, depth ~ 25

LaMaPUn:Semantic Analysis for Docs with Math (LaTeX)

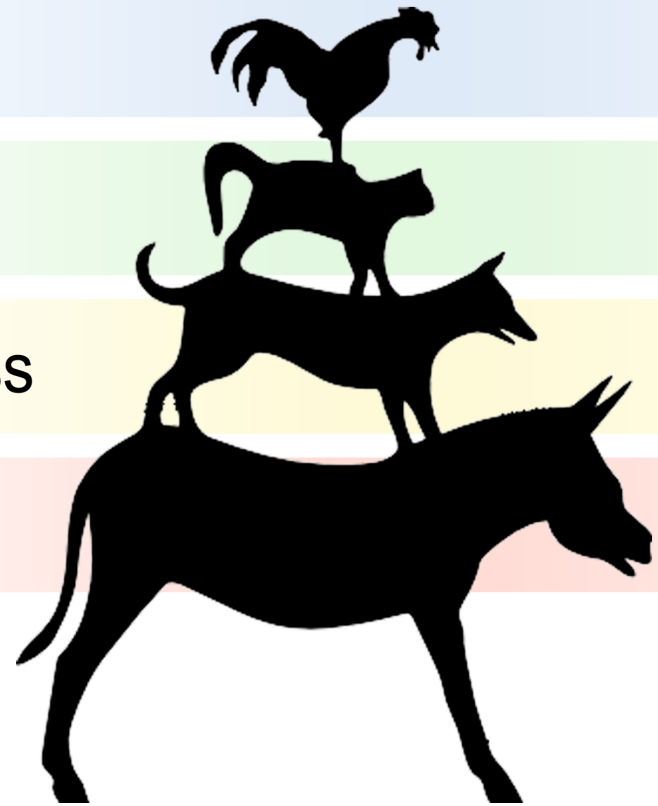


Realizing Planetary

Realizing PLANETARY: The KWARC stack

We have already developed the necessary tools/systems over the last decade

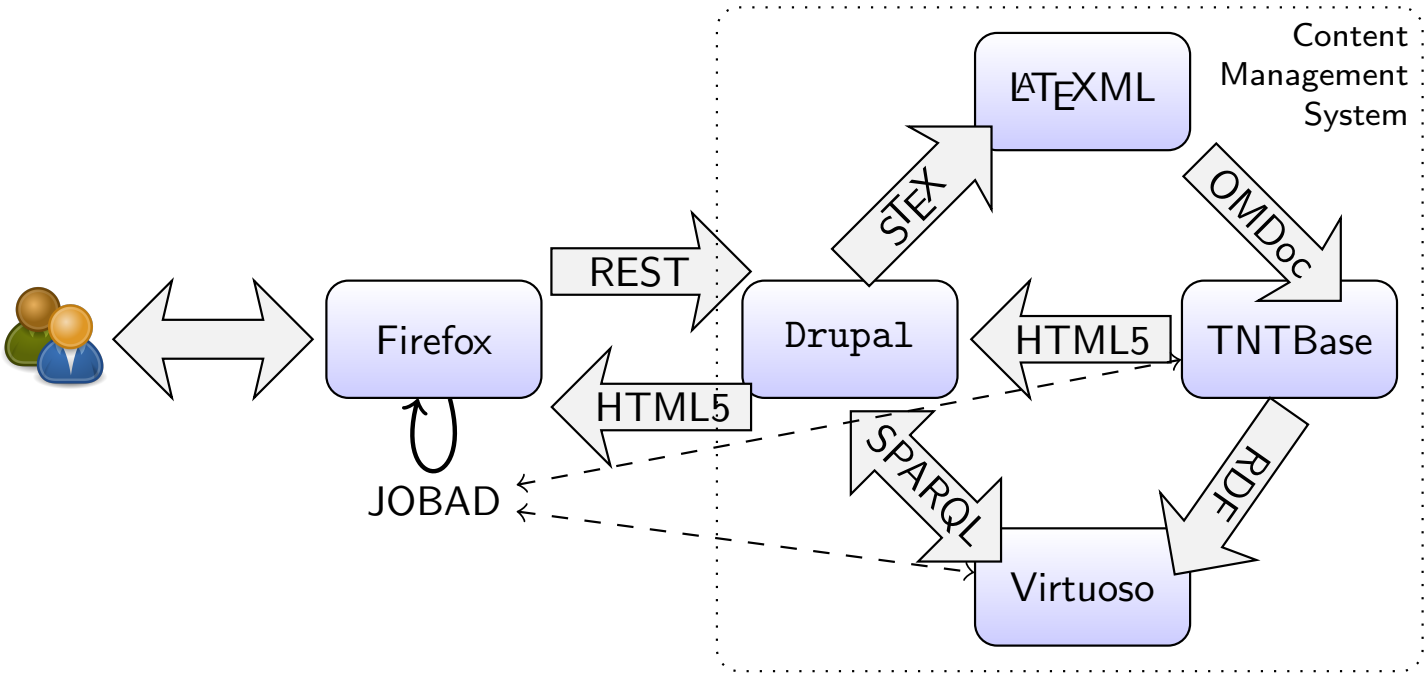
Interaction	JOBAD | SIDER
Processing	JOMDoc | Krextor
Storage | Access	TNTBase
Representation	OMDoc



PLANETARY is the ideal test bed to integrate them.

Assembling PLANETARY: System Architecture

- PLANETARY functionality can be achieved by integrating existing components.

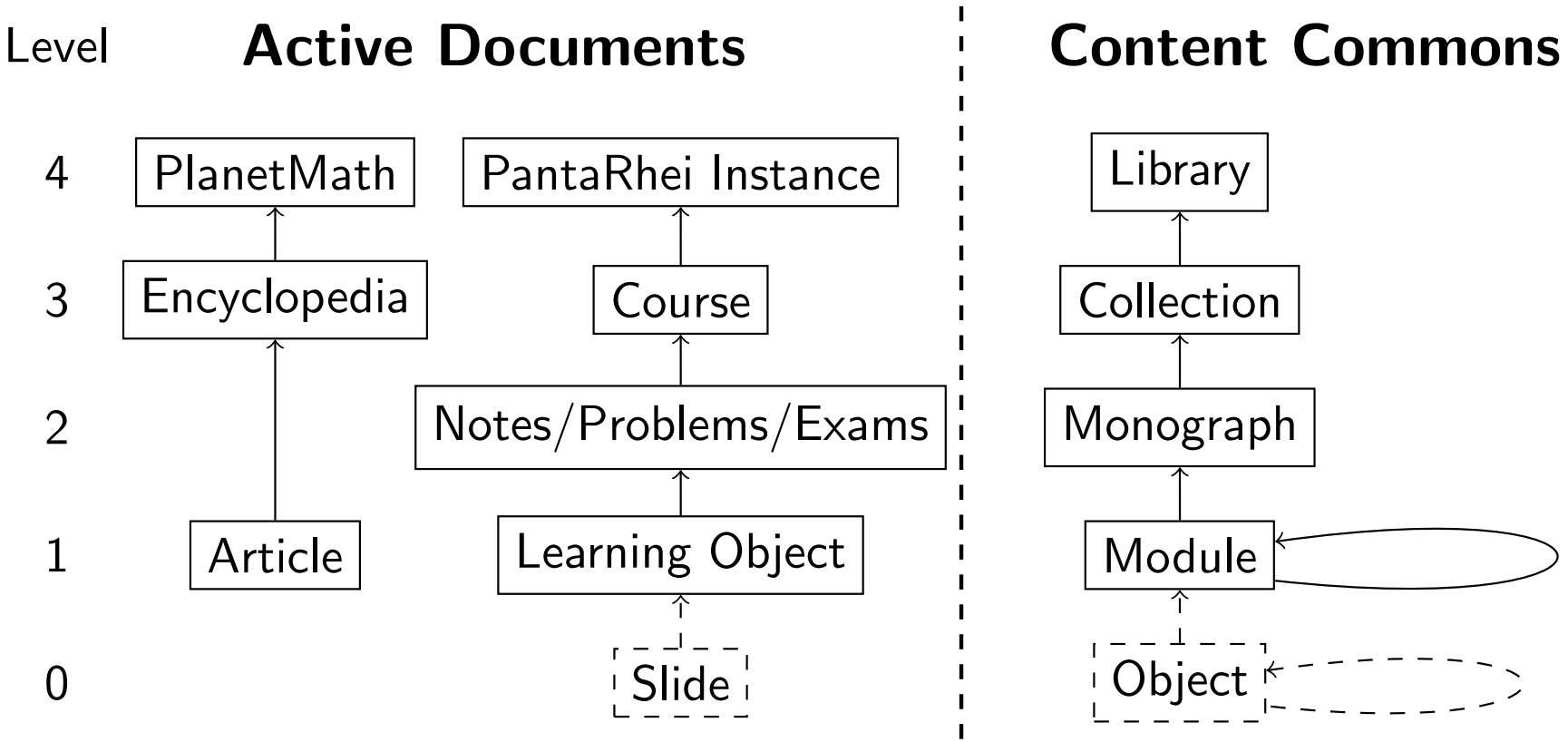


- `Drupal` for discussions, user management, caching,
- TNTBase for versioned XML storage, OMDoc presentation
- JOBAD integrates semantic services into documents
- Virtuoso is a triple store for semantic relations
- LaTeXML transforms LaTeX/sTeX to XHTML+MathML+RDFa

Organization of Content/Narrative Structure

Layers of Documents/Content

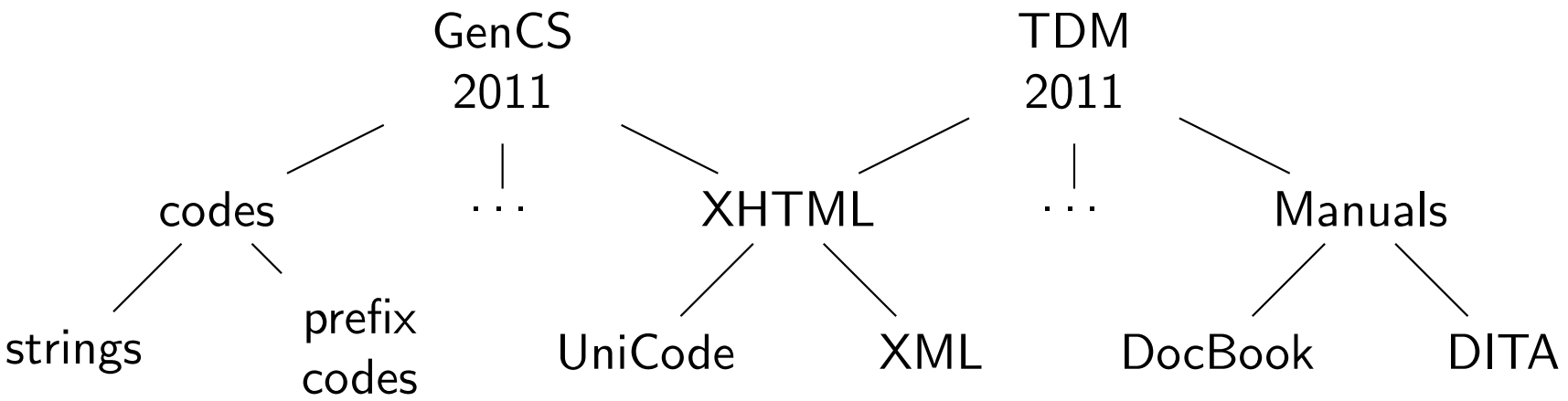
- Content and narrative structures come at different conceptual layers



- Different layers support different functionality

Monographs as Module Graphs foster Reuse

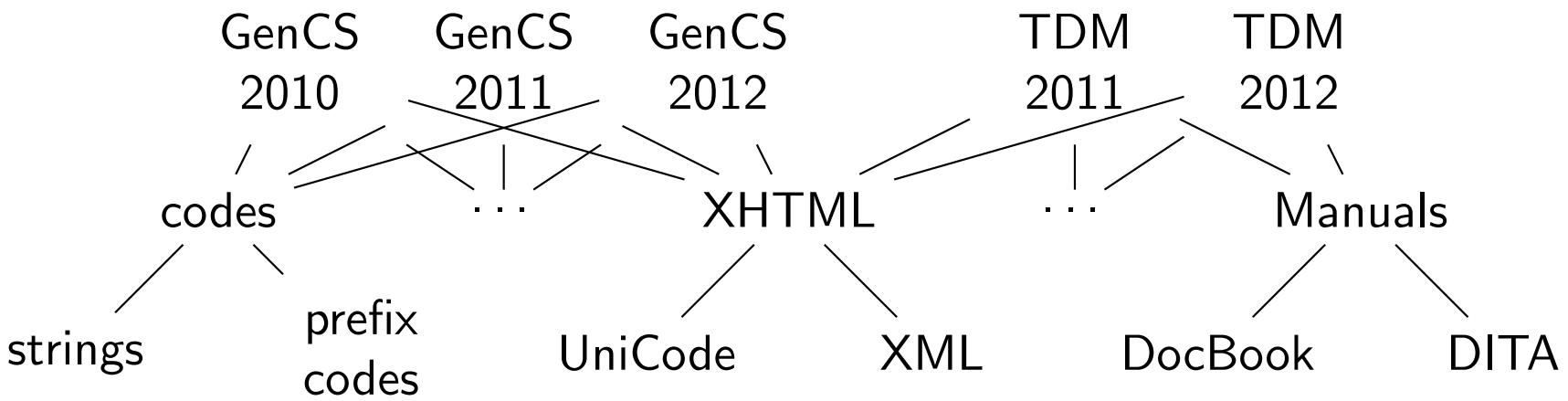
- Idea: Modules can be reused in more than one monograph
- Note: Similar to, but more general (nesting) than DITA concepts and DITA maps. (but no conditional processing (yet))
- **Example 9** For instance a module on HTML/XML in the courses "General Computer Science" and "Text and Digital Media".



Observation: These graphs can get quite large: Our corpus has 3300 nodes with 130 roots.

Monographs as Module Graphs foster Reuse

- Idea: Modules can be reused in more than one monograph
- Note: Similar to, but more general (nesting) than DITA concepts and DITA maps. (but no conditional processing (yet))
- **Example 10** For instance a module on HTML/XML in the courses "General Computer Science" and "Text and Digital Media".

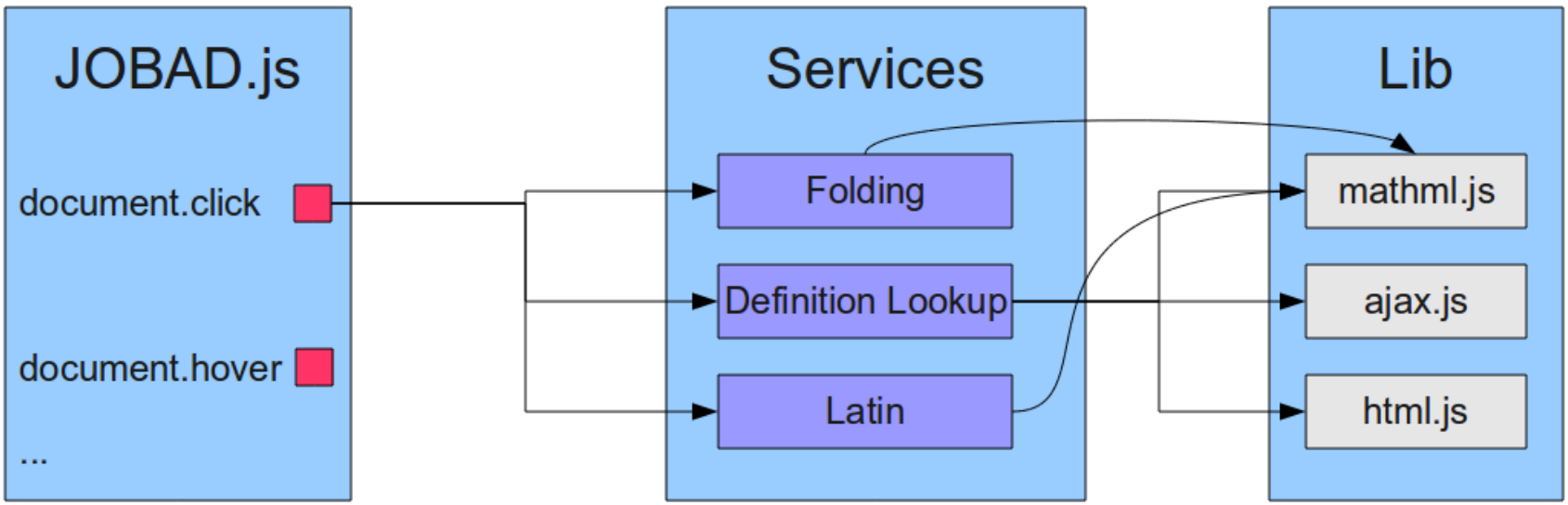


Courses given in different years share most of their content (but not all)

- Observation: These graphs can get quite large: Our corpus has 3300 nodes with 130 roots.

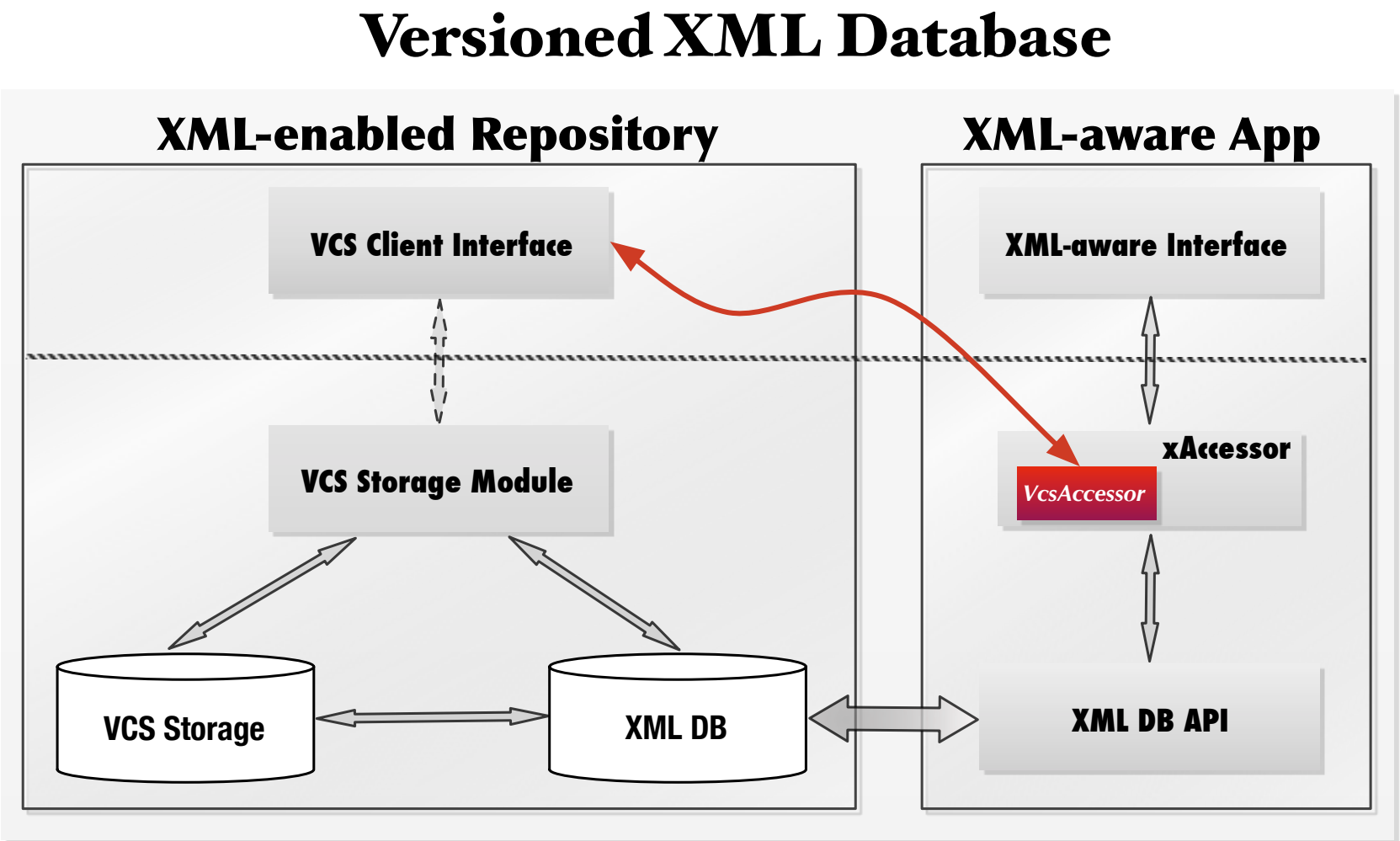
JOBAD: Embedding Semantic Services into Web Docs I

- JavaScript API for (J)OMDoc Based Active Documents
- runs inside client browser (`FireFox` currently)
- provides client-only or server-based features (extensible framework)
  based on semantic annotations in XHTML+MathML+RDFa documents
- Project home page: `https://jomdoc.omdoc.org/wiki/JOBAD`

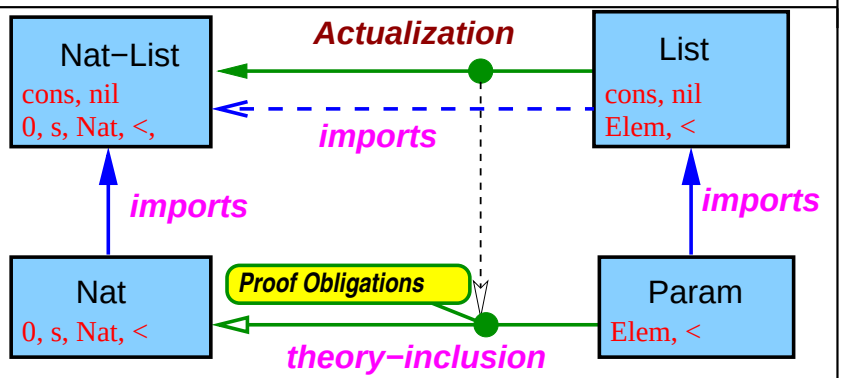


TNTBase: Versioned Storage for XML

- The TNTBase system is a versioned storage system for XML documents. It combines the functionality and interfaces of Subversion with those of an XML database.



OMDoc in a Nutshell (three levels of modeling)

Formula level: OpenMath/C-MathML • Objects as logical formulae • semantics by ref. to theory level	<OMA> <OMS cd="arith1" name="plus"/> <OMS cd="nat" name="zero"/> <OMV name="N"/> </OMA>
Statement level: • Definition, Theorem, Proof, Ex. • semantics explicit forms and refs.	<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>
Theory level: Development Graph • inheritance via symbol-mapping • theory-inclusion by proof-obligations • local (one-step) vs. global links	

LaTeXML: Converting TeX/LaTeX Documents to XML

- **Definition 11** LaTeXML converts LaTeX documents to XHTML+MathML
  - re-implement the TeX parser in `perl`. (do not expand semantic macros)
  - needs LaTeXML bindings for all LaTeX packages and classes (specify the XML for the emitter)

  Case Study: Converting the arXiv into XHTML+MathML (70% coverage of 550 k documents)

sTeX, a Semantic Variant of TeX/LaTeX

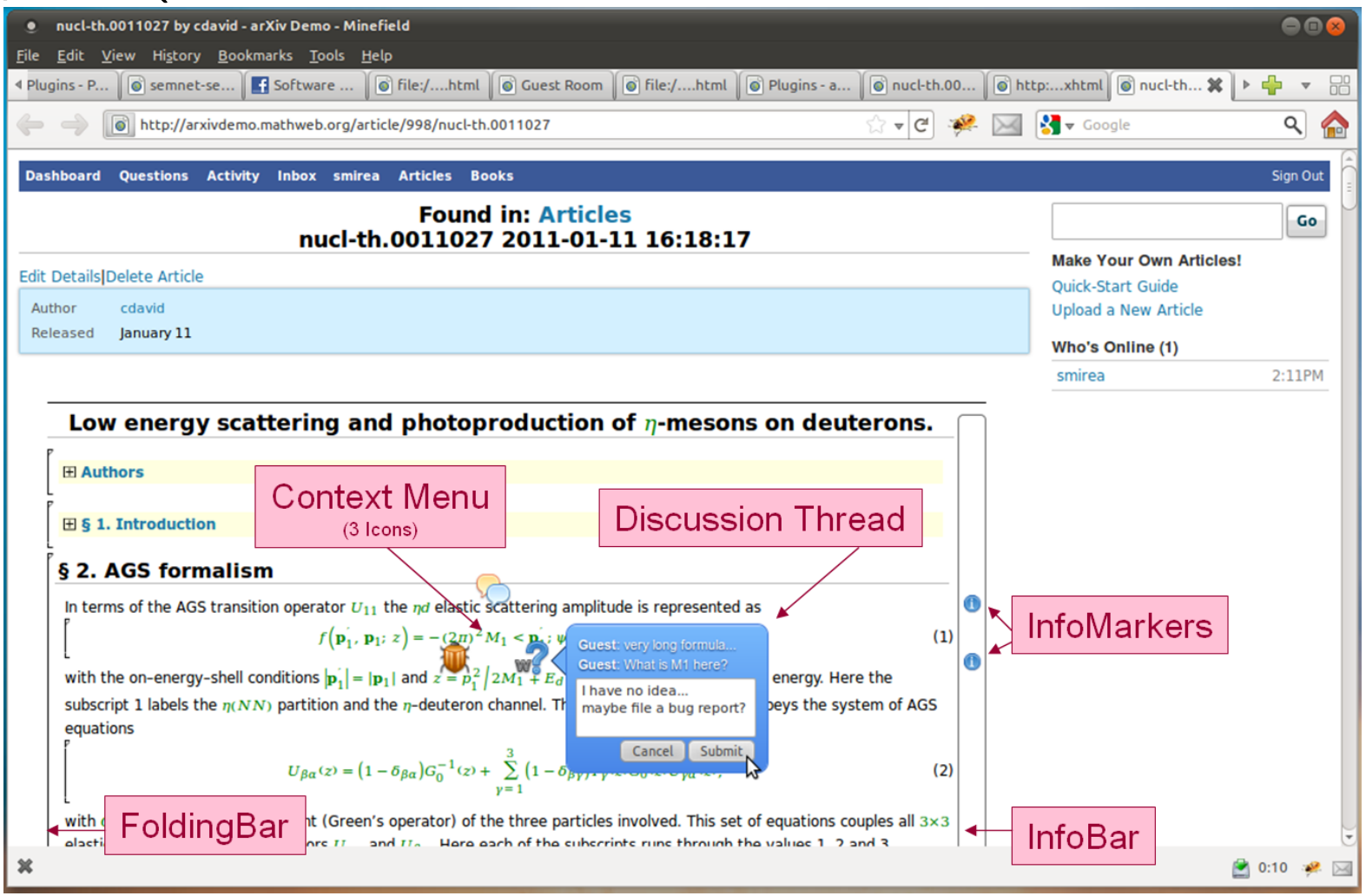
- Problem: Need content markup formats for semantic services, but Mathematicians write LaTeX
- Idea: Enable the author to make structure explicit and disambiguate meanings
  - use the TeX macro mechanism for this (well established)
  - the author knows the semantics best (at least she understands)
  - the burden is is alleviated by manageability savings (MKM on TeX/LaTeX)
- **Definition 12 (sTeX Approach)** Semantic pre-loading of TeX/LaTeX documents.
  - Introduce semantic macros: e.g. `\union{a,b,c}` $\rightsquigarrow a \cup b \cup c$
  - Mark up discourse structure: (largely invisible)
    e.g. `\begin{sproof}[id=Wiles,for=Fermat]...\end{sproof}`
  - Generate PDF and OMDoc from that (via LaTeXML [Mil])

`http://trac.kwarc.info/sTeX/`

Levels of Service in Planetary

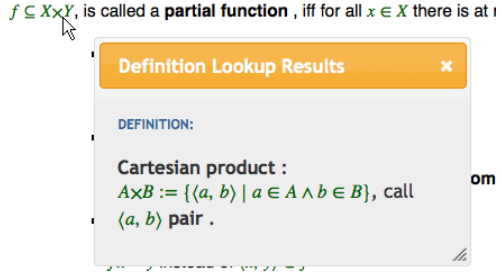
PLANETARY at the Presentation/Structural Level

- PLANETARY can make use objects and relations at various levels,
- **Example 13 (arXivdemo: Document Structure and Presentational Math)**

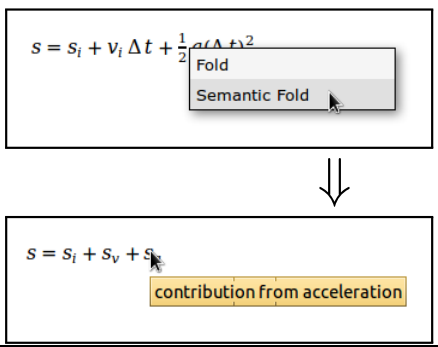


User Services at the Semantic Level in PLANETARY

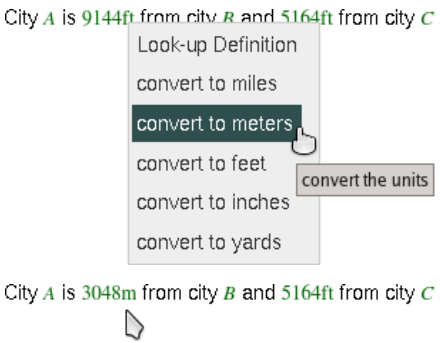
Definition Lookup



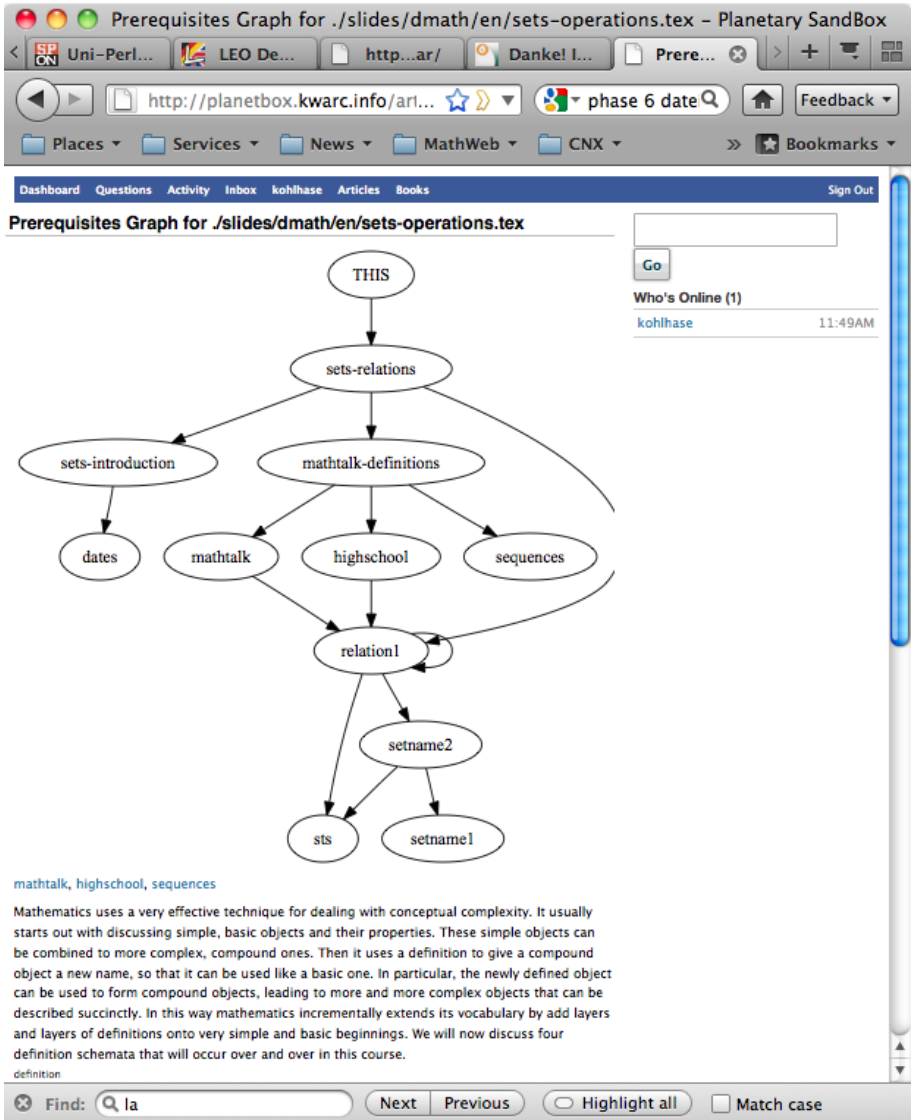
Semantic Folding



Unit Conversion

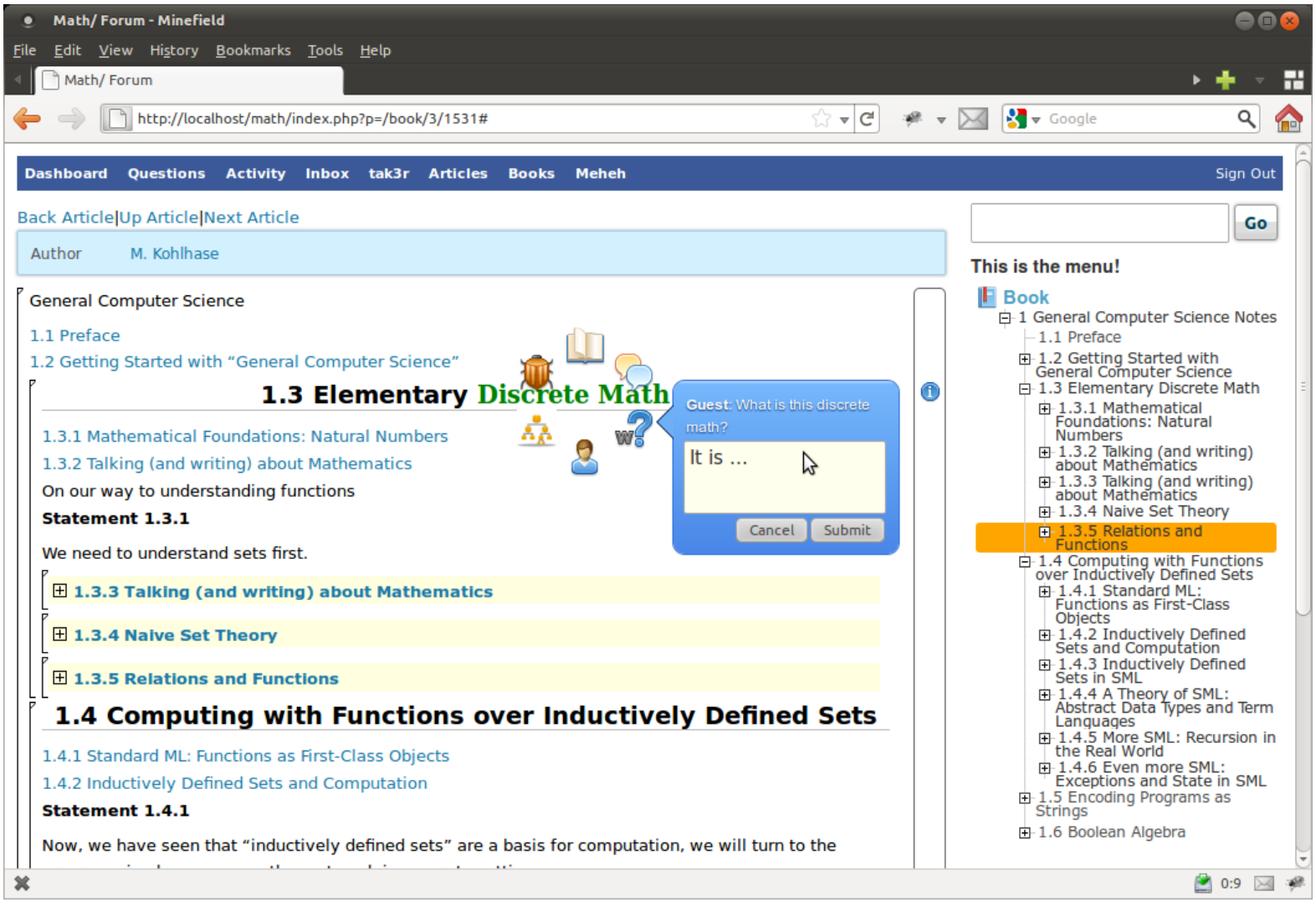


Prerequisites Navigation



PantaRhei: Semantic Course Knowledge Exploration

- PantaRhei is a semantic course knowledge exploration system based on the PLANETARY system.



User Services at the Formal Level in PLANETARY

- Formal Representations Adapted to Distinct User Settings
  (Customized via the Dashboard Widget on the Right)

AlgebraTest

- unfold ☑
  view OppositeMagmaCommut:MagmaCommut→MagmaCommut
  - mag⟼OppositeMagma ; MagmaCommut/mag
  - commut⟼forallI (λx:i.(forall (λx1:i.(x1 mag/* x == x mag/* x1))))
    (λx:i.(forallI (λx1:i.(x1 mag/* x == x mag/* x1)) (λy:i.(forall2E
    (λx1:i.(λx2:i.(x1 mag/* x2 == x2 mag/* x1))) commut y x))))

show definitions ☐
show reconstructed types ☑
show implicit arguments ☑
show redundant brackets (high value = more brackets) 50

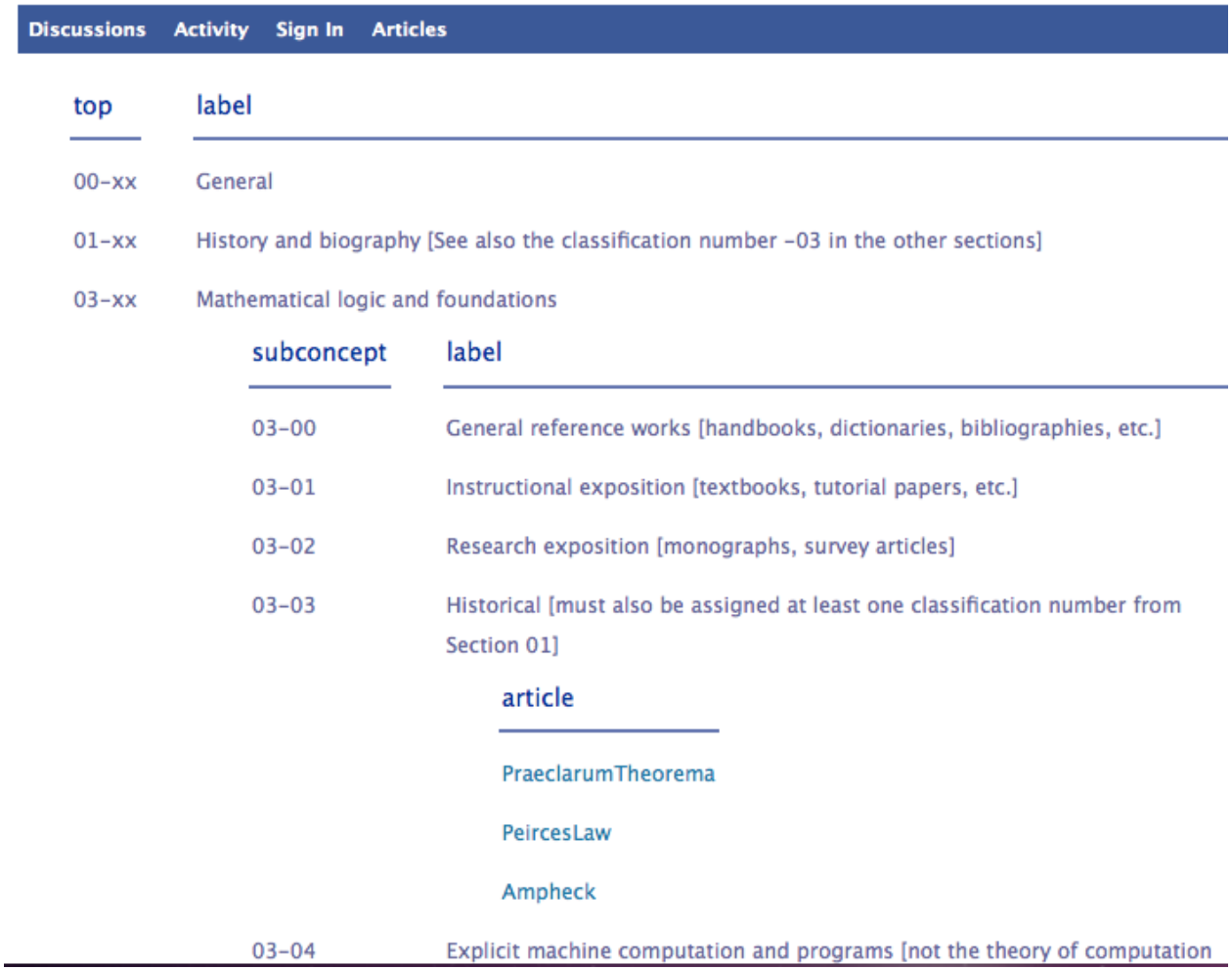
AlgebraTest

- unfold ☑
  view OppositeMagmaCommut:MagmaCommut→MagmaCommut
  - mag⟼OppositeMagma ; MagmaCommut/mag
  - commut⟼forallI (λx.forallI (λy.forall2E commut y x))

show definitions ☐
show reconstructed types ☐
show implicit arguments ☐
show redundant brackets (high value = more brackets) 0

Accessing Encyclopedias via Ontologies

- Idea: add classification metadata to articles, harvest as RDF into triplestore, compute access methods via SPARQL queries and SKOS ontology.
- **Example 14 (MSC View in PlanetMath)** use the Math Subject Classification



Discussions Activity Sign In Articles

top	label
00-xx	General
01-xx	History and biography [See also the classification number -03 in the other sections]
03-xx	Mathematical logic and foundations

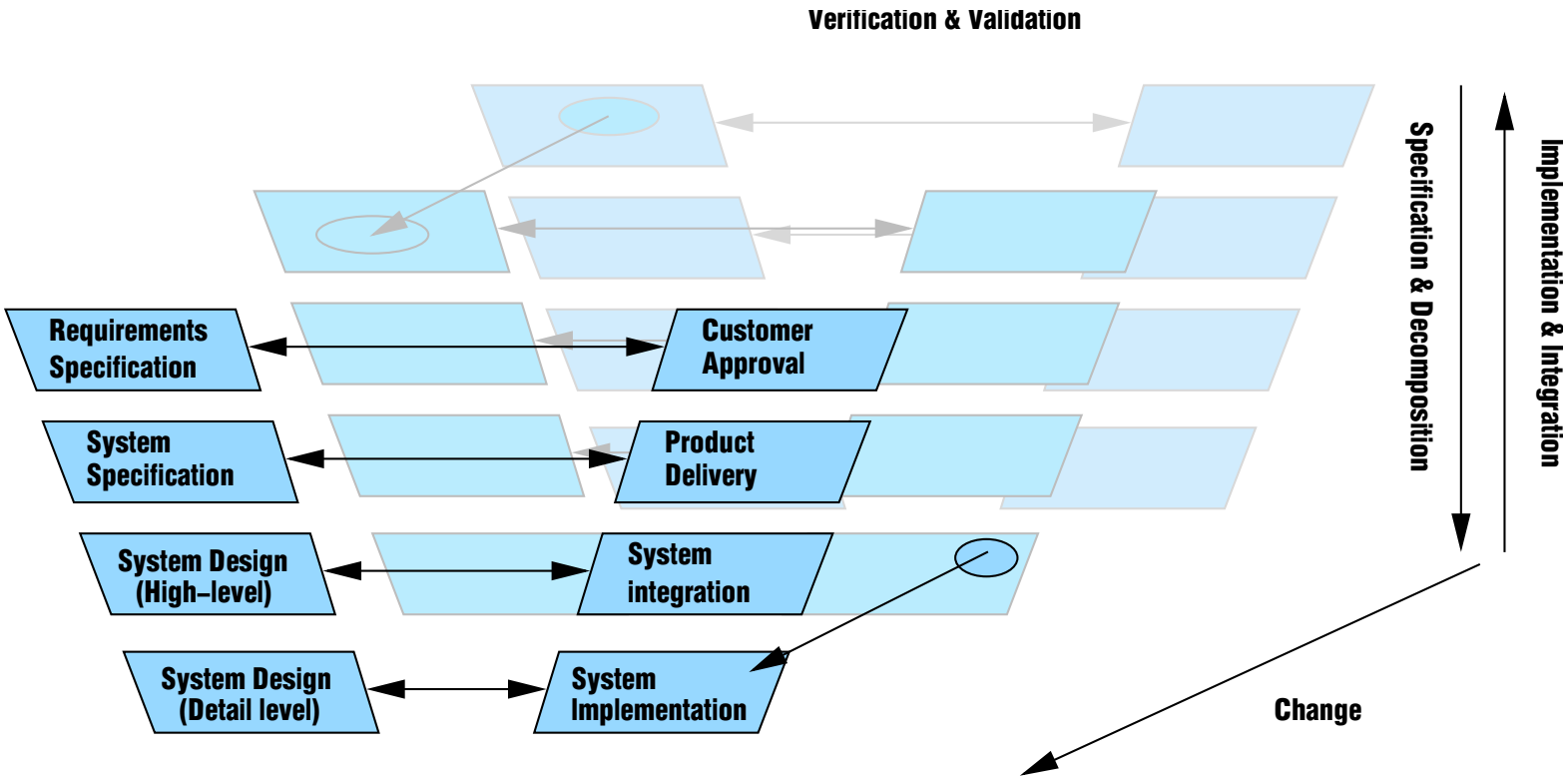
subconcept	label
03-00	General reference works [handbooks, dictionaries, bibliographies, etc.]
03-01	Instructional exposition [textbooks, tutorial papers, etc.]
03-02	Research exposition [monographs, survey articles]
03-03	Historical [must also be assigned at least one classification number from Section 01]
03-04	Explicit machine computation and programs [not the theory of computation

article
PraeclarumTheorema
PeircesLaw
Ampheck

Ontology-Based Management of Change; A Killer Application for Semantic Techniques

Application: Formal Software Development

- Idea: Understand, markup, & version development documents
- **Example 15** For instance in the V Model



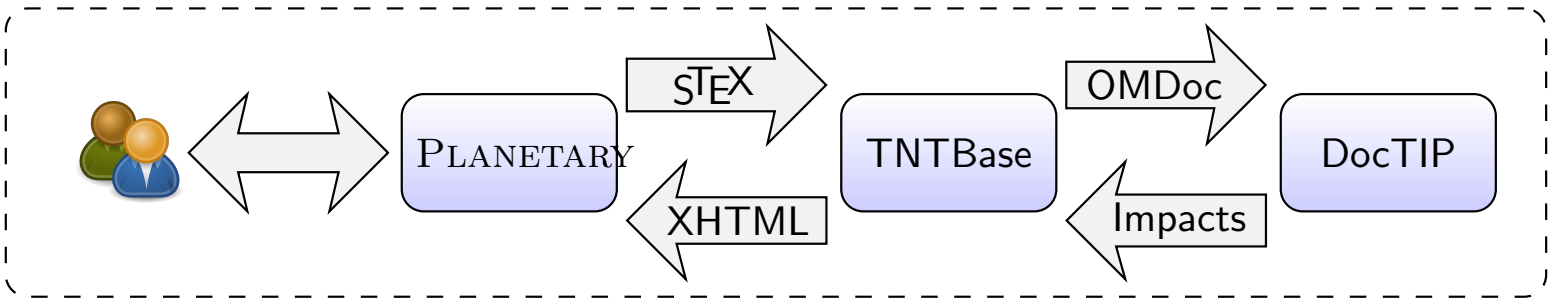
Problem: We need to understand hybrid documents (text, math, UML, code)

Management of Change in Planetary

Management of Change in PLANETARY

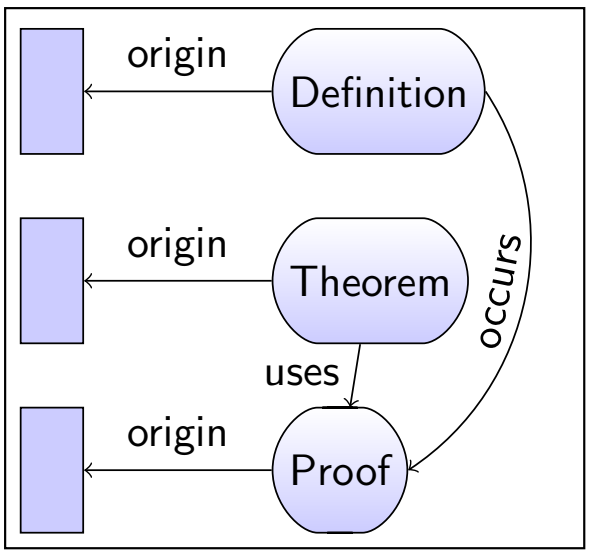
- Observation: In an eScience3.0 System, the content is constantly changing.
- Problem: How do we maintain consistency and coherence
- Idea: Integrate functionality for Management of Change.
  - Make use of the semantic relations already in place in PLANETARY.
  - If A depends on B, then a change in B impacts A.
  - Extend PLANETARY by the DocTIP system from OMoC.

    (Joint project with DFKI Bremen).
- Prototypical Integration in PLANETARY available [ADD+11]

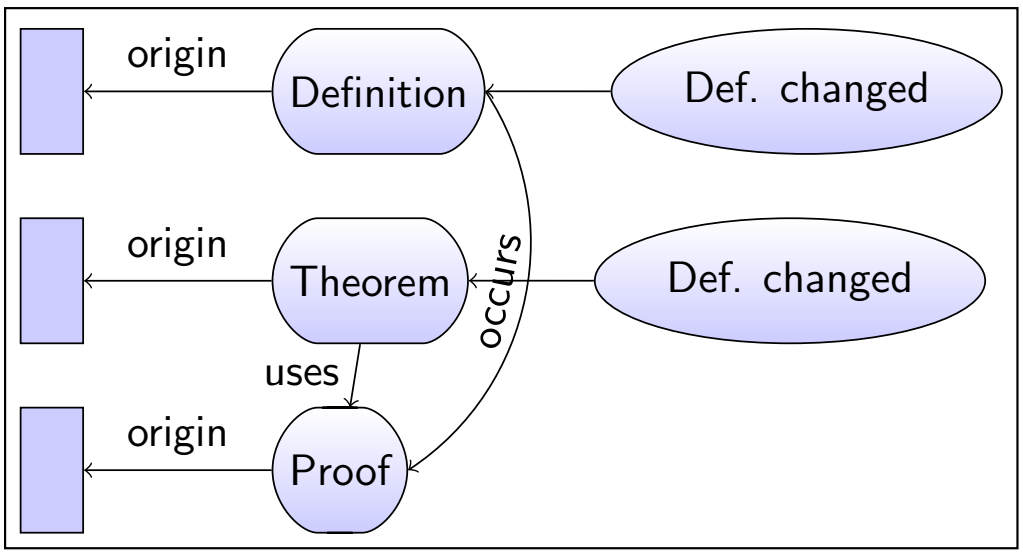


Change Impact Analysis in DocTIP

- Idea: If A depends on B, then a change in B impacts A.
- **Definition 16** Change Imact Analysis (CIA) is a process for computing potentially impacted fragments in a document collection $\mathcal{C}$ from a change description and semantic relations in $\mathcal{C}$.
- In DocTIP, CIA is computed by graph rewriting rules on the document ontology.
- **Example 17** CIA propagation rules for OMDoc



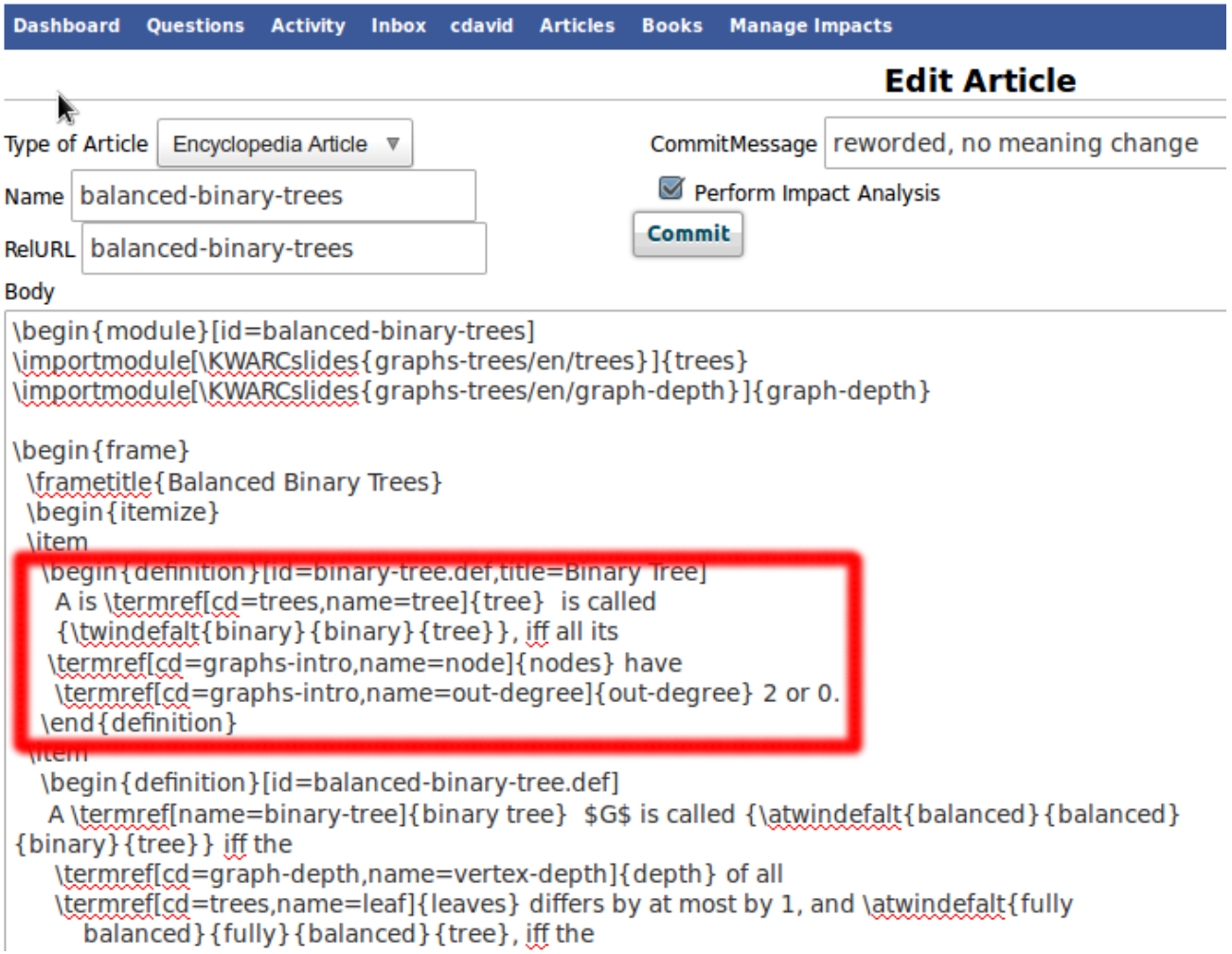
(a) Initial Syntax and Semantics



(b) Propagated Impacts after Definition Change

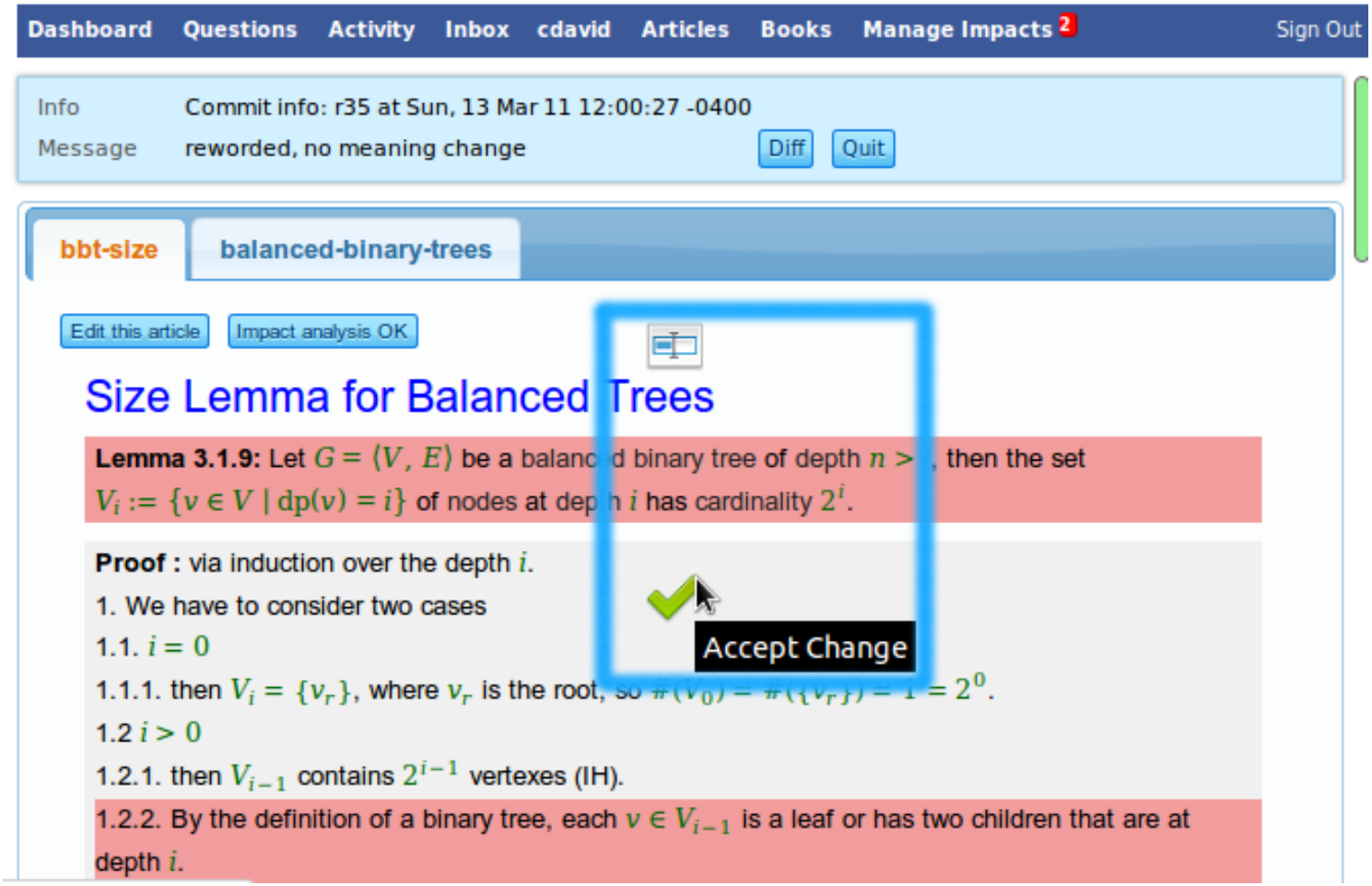
MoC in Planetary I

- Extend the commit dialog with CIA



MoC in Planetary II

- The Impact Resolution Dialog



Searching for Mathematical Formulae

Introduction & Motivation

Why we need a search engine for Mathematics

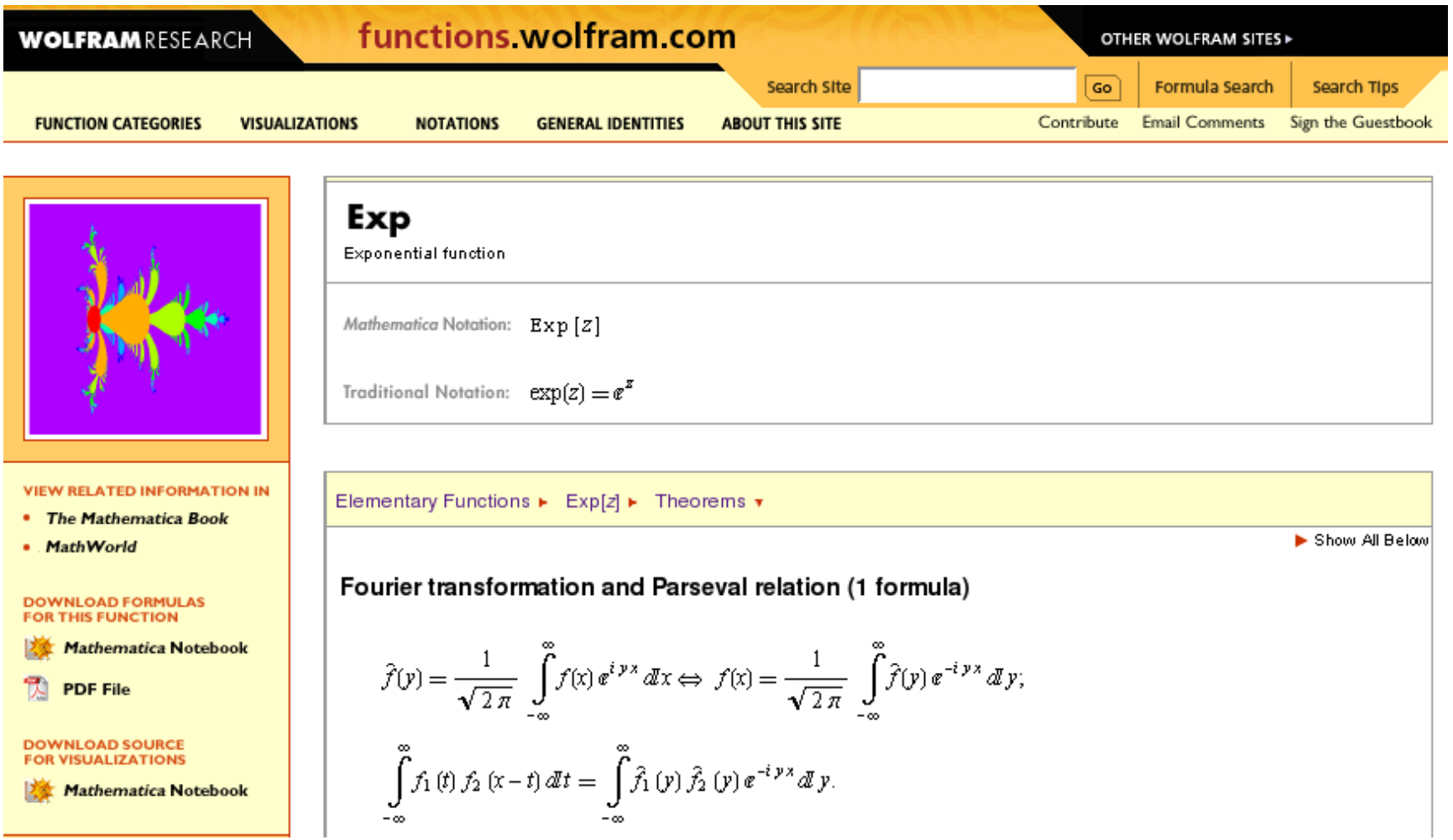
- We have come to rely on the World Wide Web for almost all of our information needs.
- The Internet is only the transport layer. We see a succession of techniques

  Early Web navigation via explicitly represented hyperlinks

  Mature Web finding relevant content via search engines
  (bag of words techniques for textual content)

  Semantic Web Representation of meaning and inferring content that is not explicitly represented
- For scientific content, we are still in the "Early Web" phase
  - Need a "Semantic Web for Science" (talk about OMDoc some other time)
  - Today: provide techniques for the "Mature Web"
  - Concretely: a search engine for math. formulae
    (a prominent non-textual part of science)

Mathematics Resources on the Web



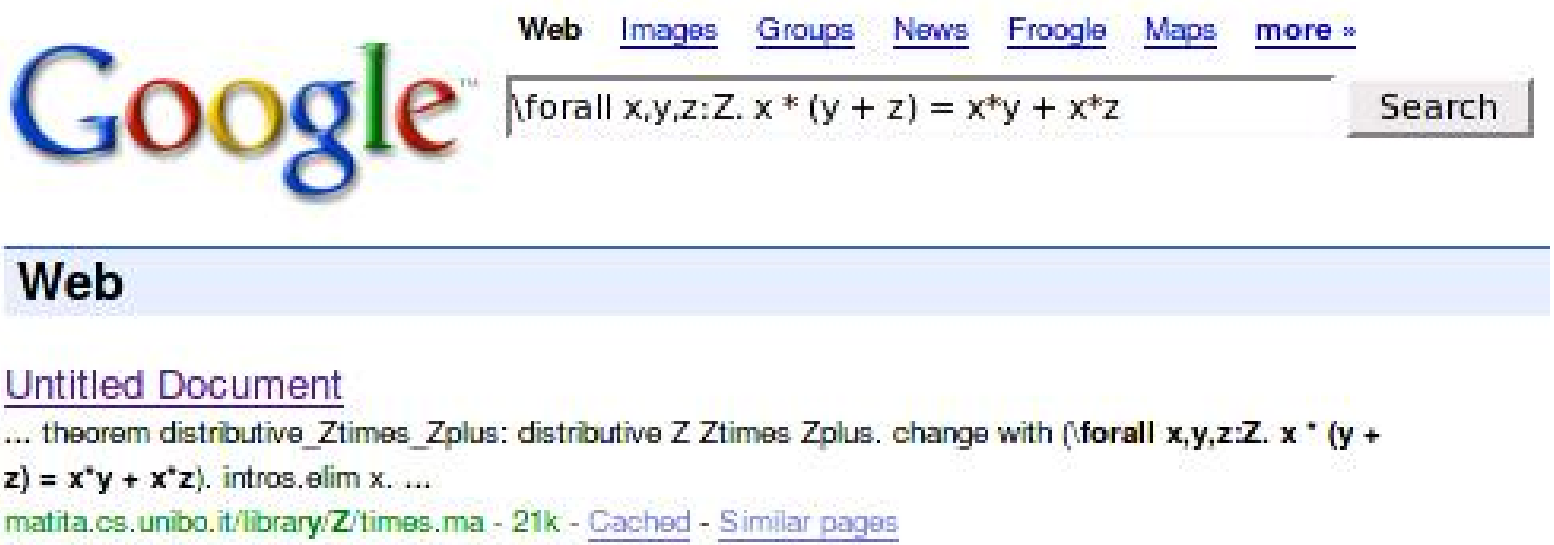
More Mathematics on the Web

- 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 ([Asperti&al'04])
- Our Idea: Use Automated Reasoning Techniques
  (free term indexing from theorem prover jails)

A running example: The Power of a Signal

- An engineer wants to compute the power of a given signal $s(t)$
- She remembers that it involves integrating the square of s.
- Problem: But how to compute the necessary integrals
- Idea: call up `MathWebSearch` with $\int_?^? s^2(t)dt$.
- `MathWebSearch` finds a document about Parseval's Theorem and $\frac{1}{T}\int_0^T s^2(t)dt = \Sigma_{k=-\infty}^{\infty}|c_k|^2$ where c_k are the Fourier coefficients of $s(t)$.

Some other Problems (Why do we need more?)

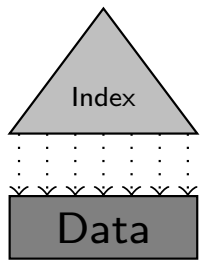
- Substitution Instances: search for $x^2 + y^2 = z^2$, find $3^3 + 4^2 = 5^2$
- Homonymy: $\binom{n}{k}$, ${}_nC^k$, C^n_k, and C^k_n all mean the same thing (binomial coeff.)
- Solution: use content-based representations (MathML, `OpenMath`)
- Mathematical Equivalence: e.g. $\int f(x)dx$ means the same as $\int f(y)dy$ (α-equivalence)
- Solution: build equivalence (e.g. α or ACI) into the search engine (or normalize first [Normann'06])
- Subterms: Retrieve formulae by specifying some sub-formulae
- Solution: record locations of all sub-formulae as well

Term Indexing

Term-Indexing

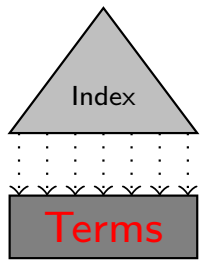
- Motivation: Automated theorem proving (efficient systems)
- Problem: Decreasing inference rate (basic operations linear in # of formulae)
- Idea: Make use of structural equality between terms (term indexing)

database systems (Algorithms: select, meet, join)



- Data: `PERSON(hans, manager, 32)`
- Query: "find all 40-year old persons"

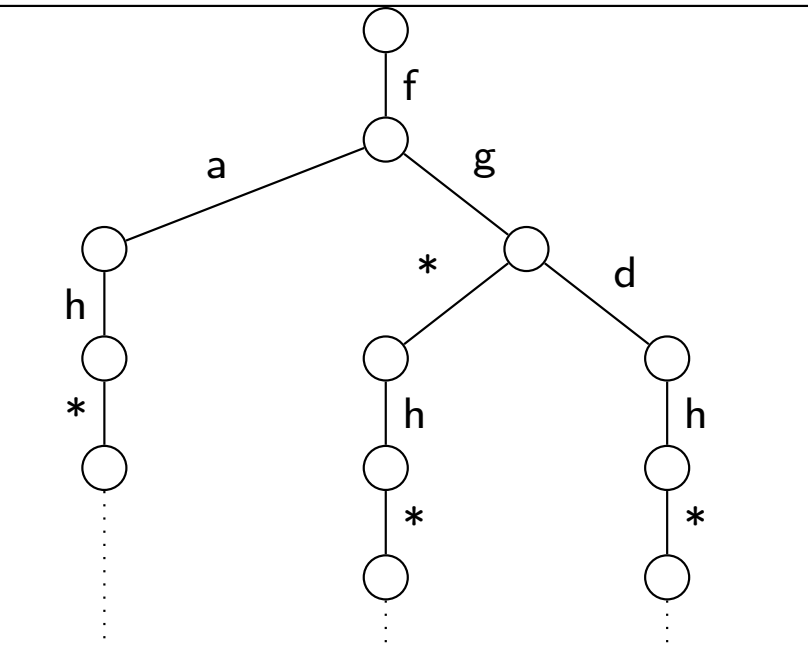
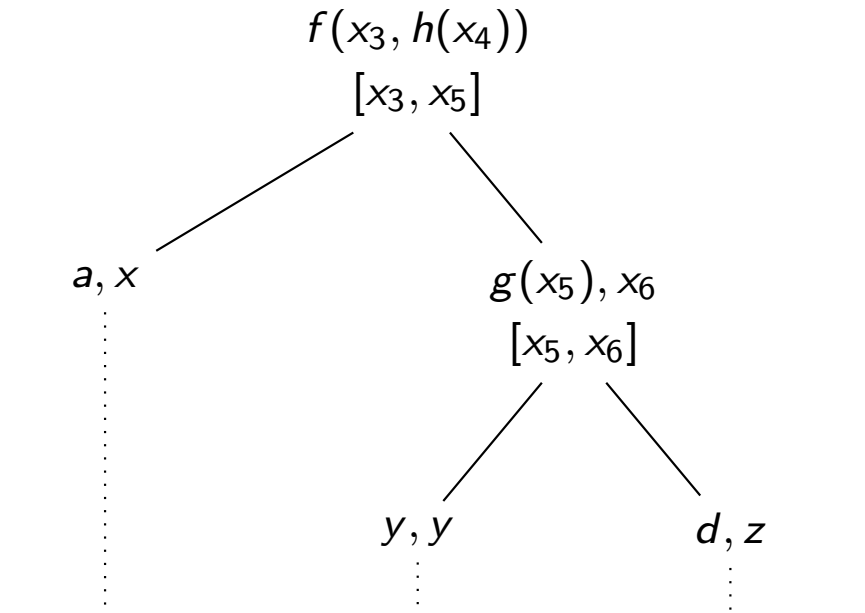
automated theorem proving (Algorithm: Unification)



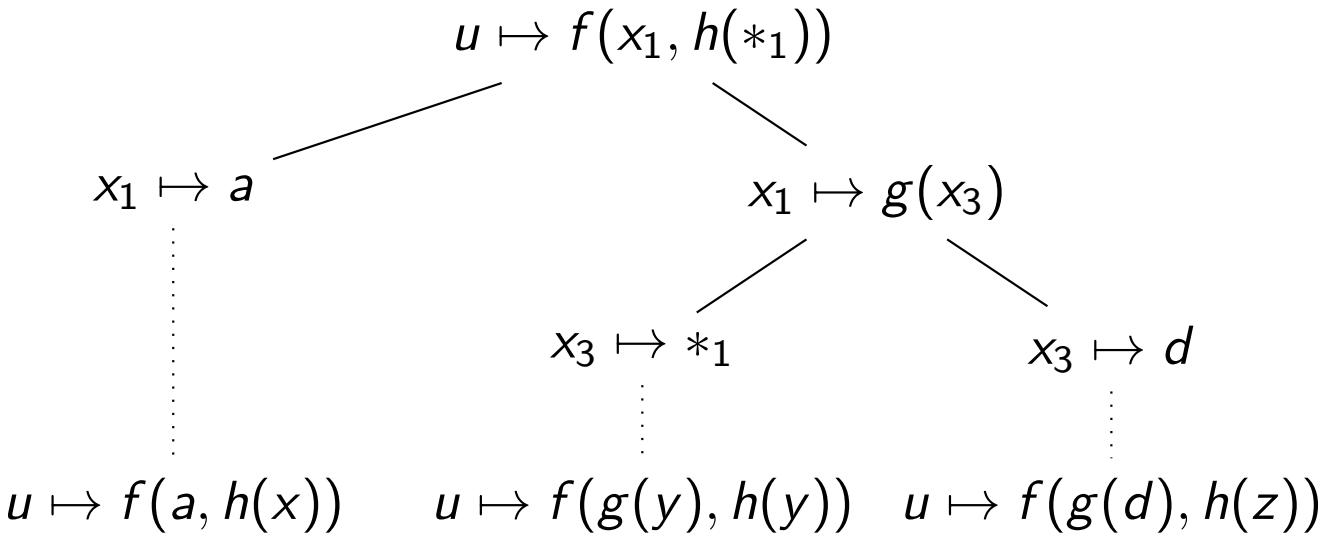
- Data: $P(f(x, g(a, b)))$
- Queries: "find all literals that are unifiable with $P(f(c, y))$"

An (additional) index data structure can make the retrieval logarithmic

Tree-based Indexing: use structural similarity of terms

Discrimiation Tree	Abstraction Tree
	
Terms as linear chains, shared in trees	Trees of substitution instances

Substitution Tree [Graf '94]



- Variant of abstraction trees that indexes Substitutions
  (Nodes labeled with Substitutions)
- includes Variable renaming
  ($*_i \hat{=}$ i^{th} variable)
- less redundant than abstraction trees
- allows $n : m$ indexing

Unification-based Search

Unification-Based Querying for Math. Formulae

- Theory: Substitution Tree Indexing is a perfect filter for
  - Variants: $\{\sigma \mid \sigma \in \mathbf{GEN}(\tau,\rho) \ \wedge \ \mathbf{supp}(\sigma) \cap \mathbf{V}^* = \emptyset\}$
  - Instances: $\{\sigma \mid \sigma \in \mathbf{UNIF}(\tau,\rho) \ \wedge \ \mathbf{supp}(\sigma) \cap \mathbf{V}^* = \emptyset\}$
  - Generalization: $\{\sigma \mid \forall x_i \in \mathbf{supp}(\tau).\tau\rho\sigma(x_i) = \rho(x_i)\}$
  - Unification: $\{\sigma \mid \forall x_i \in \mathbf{supp}(\tau).\tau\rho\sigma(x_i) = \rho\sigma(x_i)\sigma \text{ mgu}\}$
- Idea: Use all of them for querying Formulae mathematical Formulae
  - Variants: To find formulae of a given structure
  - Instances: To find formulae of a partially remembered structure
  - Generalization: To find applicable Theorems
  - Unification: A mixture of all three.

MathWebSearch: Search Math. Formulae on the Web

- Idea 1: Crawl the Web for math. formulae (in OpenMath or CMathML)
- Idea 2: Math. formulae can be represented as first order terms (see below)
- Idea 3: Index them in a substitution tree index (for efficient retrieval)
- Problem: Find a query language that is intuitive to learn
- Idea 4: Reuse the XML syntax of OpenMath and CMathML, add variables

Indexing Math Formulae as First-Order Terms?

- Mathematical Expression: $\int_0^\infty s^2(t)dt$.
- Content MathML: Formulae built up by function application, and binding from constants and variables.

```
<math>
  <apply><defint/>
    <apply><interval/><cn>0</cn><infinity/></apply>
    <bind>
      <lambda/>
      <bvar><ci>t</ci></bvar>
      <apply><power/>
        <apply><ci>s</ci><ci>t</ci></apply>
        <cn>2</cn>
      </apply>
    </bind>
  </apply>
</math>
```

Idea: Extend Substitution Tree Indexing with bound variables and α-renaming

- Technically: Use deBruijn Indexes
  (bvars as name-less pointers interact well with substitution)

Instantiation Queries

- Application: Find partially remembered formulae
- **Example 18** An engineer might face the problem remembering the energy of a given signal $f(x)$
  - Problem: hmmmm, have to square it and integrate
  - Query Term: $\int_{\boxed{min}}^{\boxed{max}} \boxed{f}(x)^2 dx$ ($\boxed{i}$ are search variables)
  - One Hit: Parseval's Theorem $\frac{1}{T}\int^{T_0} s^2(t)dt = \sum_{k=-\infty}^{\infty} \|c_k\|^2$ (nice, I can compute it)
- This works out of the box (has ween working in `MathWebSearch` for some time)
- Another Application: Underspecified Conjectures/Theorem Proving
  - during theory exploration we often have some freedom
  - express that using metavariables in conjectures
  - instantiate the conjecture metavariables as the proof as the proof dictates

  applied e.g. in Alan Bundy's "middle-out reasoning" in proof planing

Generalization Queries

- Application: Find (possibly) applicable theorems
- **Example 19** A researcher wants to estimate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ from above
  - Problem: Find inequation such that $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ matches left hand side.
  - e.g. Hölder's Inequality: ($\boxed{i}$ 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: Take the instance

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

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

Unification Queries

- Application: Find appliccable theorems for underspecified formulae
- **Example 20** estimate $g^2\cos(\boxed{x}) + b\sin(\sqrt{y})$
  - this unifies with $\boxed{a}\cos(\boxed{t}) + \boxed{b}\sin(\boxed{t}) \leq ?$
  - result: $g^2\cos(\sqrt{y}) + b\sin(\sqrt{y}) \leq \sigma(?)$, where σ is the mgu

  Problem: Users find it difficult to state the exact unification query

- Solution: (from Databases again)
  - express the query in SELECT FROM WHERE form.
  - e.g. SELECT instance $\left(\int_{\boxed{D}} \frac{1}{\sqrt{2\pi}} \exp\{B\}\right)$ WHERE B=variation$(x^2 + jy^2)$
- `MathWebSearch` preprocessor compiles subqueries into one unification query for efficiency.

Where do the universal variables come from

- Problem: we need to have e.g. Hölder's Inequality in the index:

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

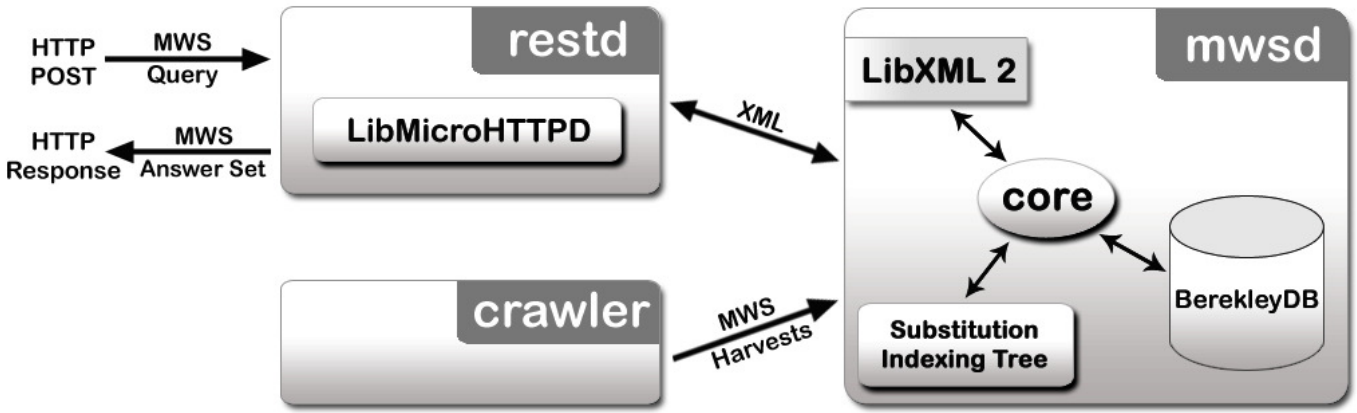
  - How do we know what symbols are "universal" (to be instantiated?)
  - what is their scope (when are different occurrences of f different?)
  - we have no sources with explicit quantifiers, but ([Wikipedia])

    Let (D, Σ, μ) be a measure space and let $1 \leq p, q \leq \infty$ with $1/p + 1/q = 1$. Then, for all measurable real- or complex-valued functions f and g on D, ...

- Solution: Use techniques from computational linguistics and integrate them into the indexing pipeline. (we have started a bit on the arXiv)

The MathWebSearch System

System Architecture

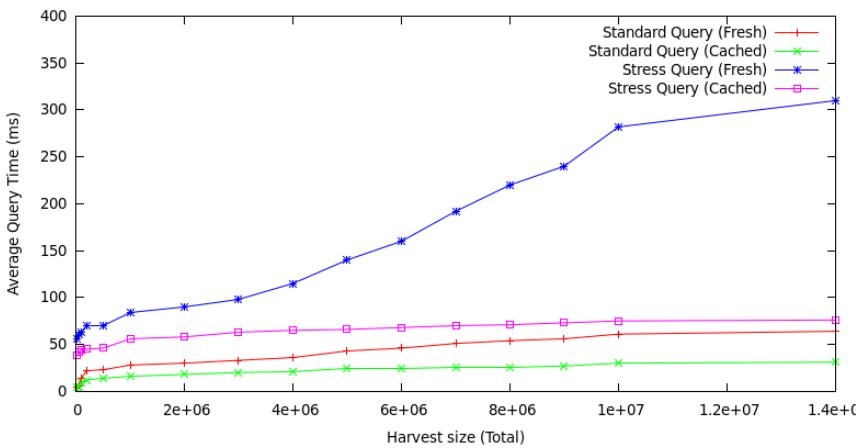


- crawlers for MathML, `OpenMath`, and OAI repositories. (convert your's?)
- multiple search servers based substitution tree indexing (formula search)
- a RESTful server that acts as a front-end for multiple search servers.
- various front ends tailored to specific applications (search appliances)
  - a Google-like web front end for human users (`search.mathweb.org`)
  - a LaTeX-based front-end for the arXiv (`http://arxivdemo.mathweb.org`)
  - special integrations for theorem prover libraries (MizarWiki, TPTP)

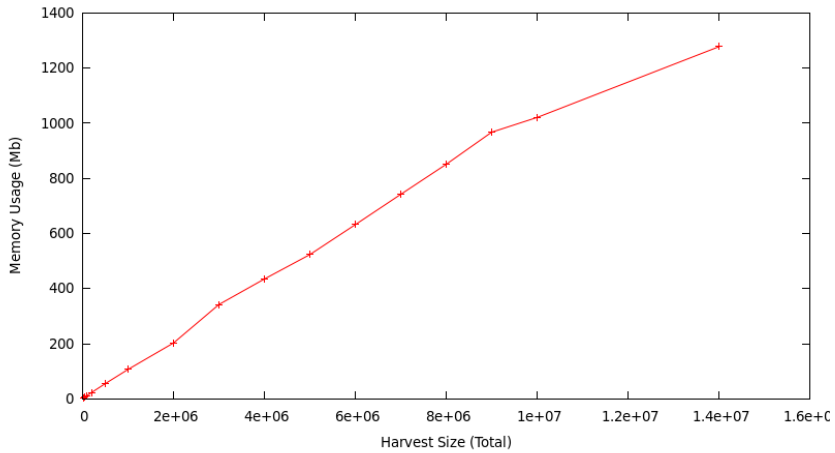
Index statistics

- Experiment: Indexing the arXiv (700k documents, $\sim 10^8$ non-trivial formulae)
- Results: indexing up to 15 M formulae on a standard laptop

Query Times

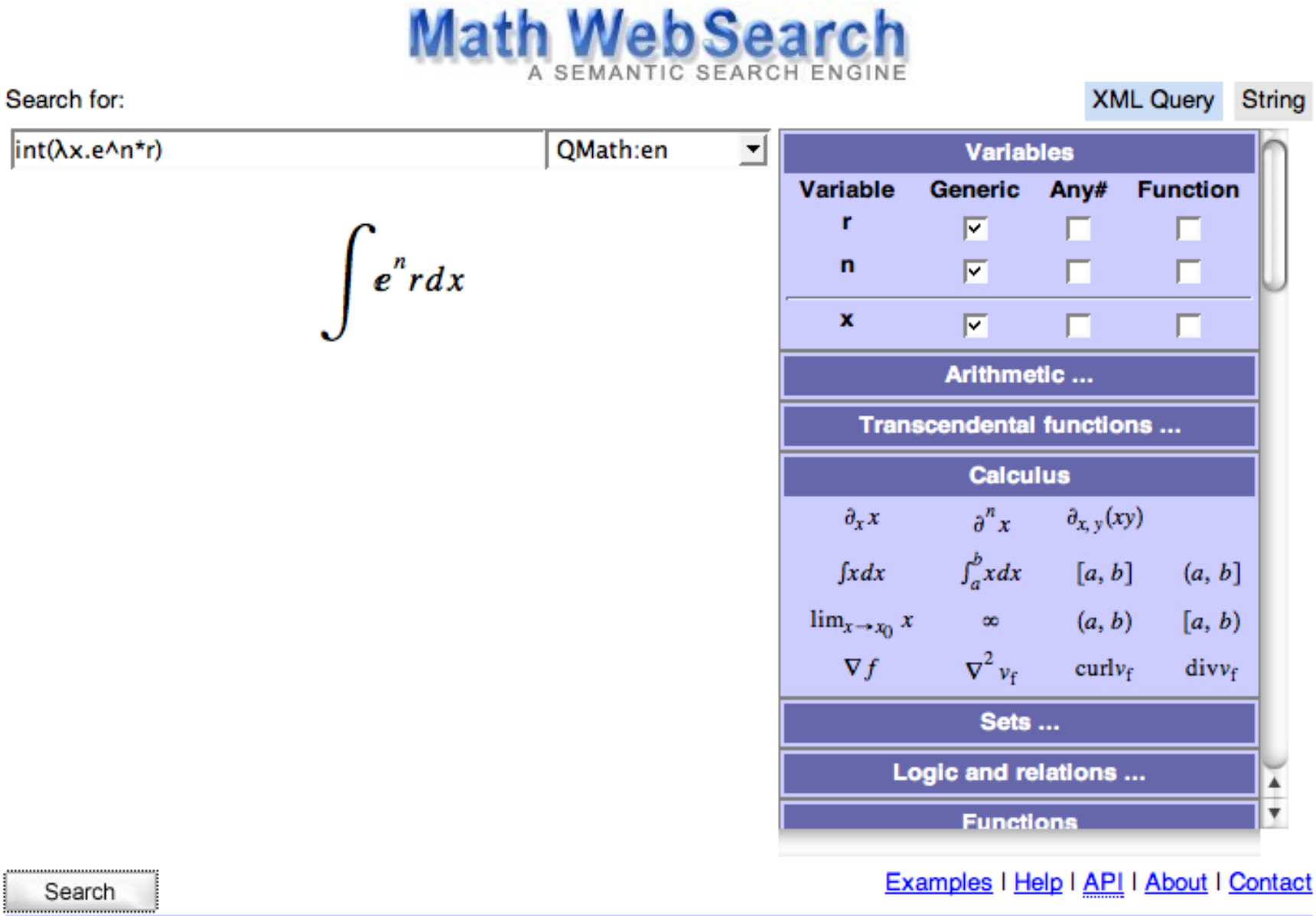


Memory Footprint



  - query time is constant ($\sim$ 50 ms) (as expected; goes by depth $\times$ symbols)
  - memory footprint seems linear ($\sim 100 \frac{B}{\text{formula}}$) (expected more duplicates)
- So we need ca 15 *GB* RAM for indexing the whole arXiv.
- Can index all published Math ($\hat{=}$ 5 $\times$ arXiv) on a large server. (ZBL $\hat{=}$ 3M art.)

Instead of a Demo: Searching for Signal Power



Instead of a Demo: Search Results

Other integrals (5 formulas) (Source)

Other integrals (5 formulas)

Matched term:

$$\int \frac{e^{3z/4}}{\left(-2+e^{3z/4}\right)\sqrt{-2+e^{3z/4}+e^{3z/2}}}\,dz == \frac{2}{3}\left(\log\left(-2+e^{3z/4}\right)-\log\left(4\sqrt{-2+e^{3z/4}+e^{3z/2}}+5e^{3z/4}-2\right)\right)$$

Rank: 100%

XML Source

Used substitution:

$$\mathbf{n} \to 3z4^{-1}$$

$$\mathbf{r} \to \left(\left((-2)+e^{3z4^{-1}}\right)\left((-2)+e^{3z4^{-1}}+e^{3z2^{-1}}\right)^{1/2}\right)^{-1}$$

$$\mathbf{x} \to z$$

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

Questions Activity Sign In Books Articles MWS Engine BETA

\lim_{\qvar{x}\rightarrow 0}\qvar{y}

$\lim_{x \to 0} y$

```
<m:apply>
      <m:apply>
       <m:csymbol
cd="ambiguous">subscript</m:csymbol>
       <m:limit/>
       <m:apply>
        <m:ci>→</m:ci>
```

Search

Examples - LaTeX queries
Generic subscript search
Specific subscript search
Specific integral search
Physical constant search
All limits approaching zero
Text in math search

1 2 next

$$\chi(t,t_w) = \lim_{h_0 \to 0} \frac{m[h](t)}{h_0}.$$

Generalized off-equilibrium fluctuation-dissipation relations in random Ising systems

Author: Federico Ricci-Tersenghi <ricci@chimera.roma1.infn.it>

$$\lim_{\mu,\mu_0 \to 0} I_1^t(\mu,\mu_0,\phi-\phi_0) = \frac{aF_0}{4(c+1)},$$

Behavior of the reflection function of a plane-parallel medium for directions of incidence and reflection tending to horizontal directions

Author: Daphne Stam <d.m.stam@sron.nl>

$$\lim_{\mu,\mu_0 \to 0} I_1^t(\mu,\mu_0,\phi-\phi_0)$$

Behavior of the reflection function of a plane-parallel medium for directions of incidence and reflection tending to horizontal directions

Instead of a Demo: Appliccable 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;
```

But Math consists of more than Formulae

- Idea: Text and Formulae co-constrain each other in search
  would like to search for formulae as well als words
- Problem: turnkey text search engines exist, but are incompatible with `MathWebSearch`
- Solution: MaTeSearch combines text and formula search results.
  1. compute text/formula search individually, rank them
  2. compute ordered intersection of both and display.
- Problem: Ranking for formula search is still in its infancy

Conclusions & Future Work

- The time is ripe for eMath 3.0
  - We have the necessary building blocks (integration needed)
  - we have first case studies for eMath 3.0 (more to come)
  - General Metaphor: Math Document as an interface to Math Knowledge

Conclusions & Future Work

- The time is ripe for eMath 3.0
  - We have the necessary building blocks (integration needed)
  - we have first case studies for eMath 3.0 (more to come)
  - General Metaphor: Math Document as an interface to Math Knowledge
- Active Documents
  - as interfaces to a structured knowledge base (background ontology)
  - via semantic annotation of documents (referencing it)

Conclusions & Future Work

- The time is ripe for eMath 3.0
  - We have the necessary building blocks (integration needed)
  - we have first case studies for eMath 3.0 (more to come)
  - General Metaphor: Math Document as an interface to Math Knowledge
- Active Documents
  - as interfaces to a structured knowledge base (background ontology)
  - via semantic annotation of documents (referencing it)
- The Planetary System as an integrated interaction platform.
  - semantic services for added value
  - semantification support for documents

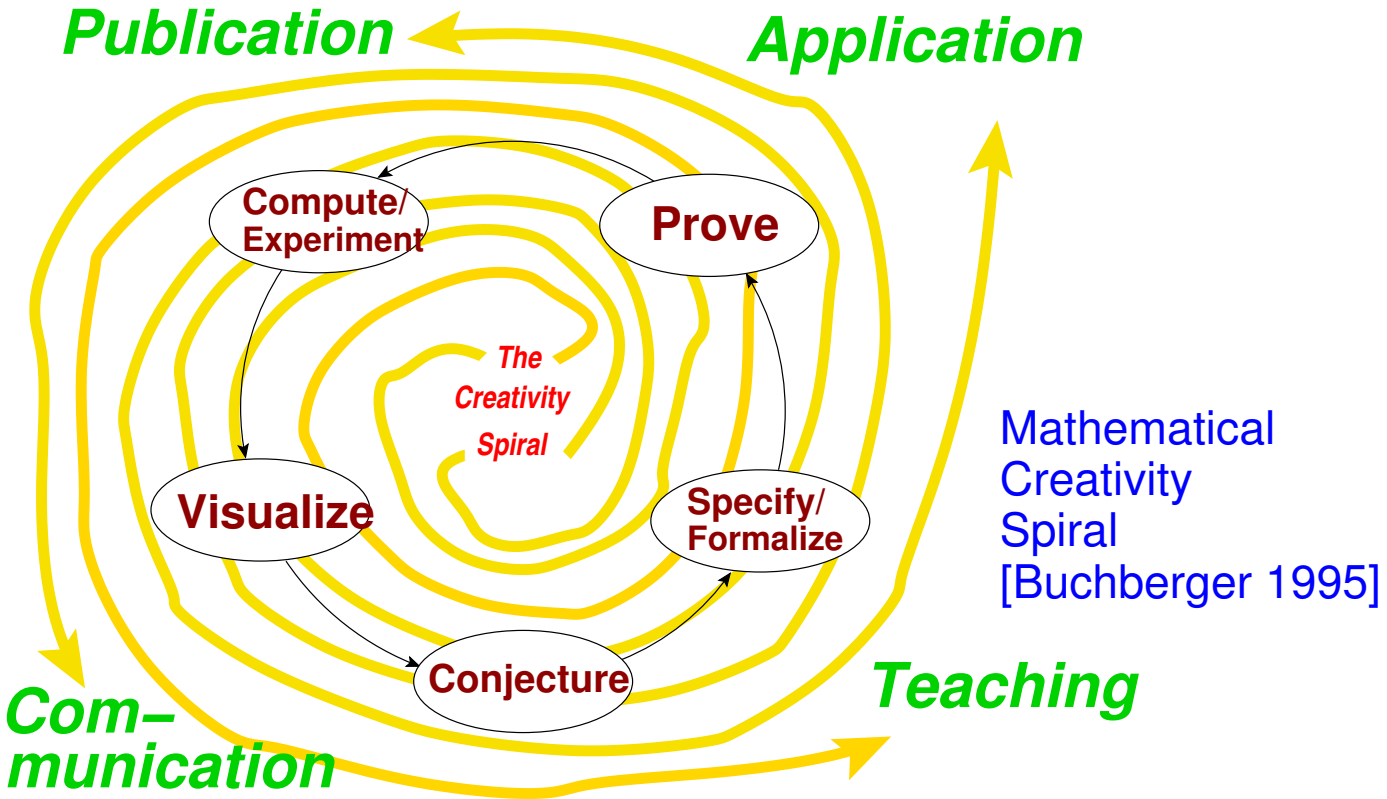
Conclusions & Future Work

- The time is ripe for eMath 3.0
  - We have the necessary building blocks (integration needed)
  - we have first case studies for eMath 3.0 (more to come)
  - General Metaphor: Math Document as an interface to Math Knowledge
- Active Documents
  - as interfaces to a structured knowledge base (background ontology)
  - via semantic annotation of documents (referencing it)
- The Planetary System as an integrated interaction platform.
  - semantic services for added value
  - semantification support for documents
- Management of Change (to keep mutable collections consistent)
  - Change Impact Analysis based on the existing semantic relations
  - Flexible impact resolution workflows in Planetary

Conclusions & Future Work

- The time is ripe for eMath 3.0
  - We have the necessary building blocks (integration needed)
  - we have first case studies for eMath 3.0 (more to come)
  - General Metaphor: Math Document as an interface to Math Knowledge
- Active Documents
  - as interfaces to a structured knowledge base (background ontology)
  - via semantic annotation of documents (referencing it)
- The Planetary System as an integrated interaction platform.
  - semantic services for added value
  - semantification support for documents
- Management of Change (to keep mutable collections consistent)
  - Change Impact Analysis based on the existing semantic relations
  - Flexible impact resolution workflows in Planetary
- Research Interest: Apply this to semi-formal STEM Documents

The way we do math will change dramatically



- Every step will be supported by mathematical software systems
- Towards an infrastructure for web-based mathematics!


Serge Autexier, Catalin David, Dominik Dietrich, Michael Kohlhase, and Vyacheslav Zholudev.
Workflows for the management of change in science, technologies, engineering and mathematics.
In Davenport et al. [DFRU11], pages 164–179.


Mark Birbeck, Ben Adida, Ivan Herman, and Manu Sporny.
RDFa 1.1 primer.
W3C Working Draft, World Wide Web Consortium (W3C).


James Davenport, William Farmer, Florian Rabe, and Josef Urban, editors.
Intelligent Computer Mathematics, number 6824 in LNAI. Springer Verlag, 2011.


Constantin Jucovschi and Michael Kohlhase.
sTeXIDE: An integrated development environment for sTeX collections.
In Serge Autexier, Jacques Calmet, David Delahaye, Patrick D. F. Ion, Laurence Rideau, Renaud Rioboo, and Alan P. Sexton, editors, *Intelligent Computer Mathematics*, number 6167 in LNAI. Springer Verlag, 2010.


Graham Klyne and Jeremy J. Carroll.
Resource Description Framework (RDF): Concepts and abstract syntax.
W3C recommendation, World Wide Web Consortium (W3C), 2004.


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 Davenport et al. [DFRU11], pages 307–309.


Bruce Miller.
LaTeXML: A LaTeX to XML converter.
Web Manual at http://dlmf.nist.gov/LaTeXML/.
seen September2011.


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.


OWL 2 web ontology language: Document overview.
W3C recommendation, World Wide Web Consortium (W3C), October 2009.