

Transforming the arXiv to XML

MICHAEL KOHLHASE

Computer Science
School of Engineering & Science
Jacobs University Bremen, Germany
<http://kwarc.info/kohlhase>

SCOOP, 27. June 2008

The arXMLiv Project: arXiv \rightsquigarrow semantic XML

- ▶ **Idea:** Develop a large corpus of knowledge in *OMDoc/PhysXML*
 - ▶ to get around the chicken-and-egg problem of MKM
 - ▶ corpus-linguistic methods for semantics recovery (linguists interested)
- ▶ **The Cornell Preprint arXiv:** (<http://www.arxiv.org>)
Open access to ca. 500.000 e-prints in Physics, Mathematics, Computer Science and Quantitative Biology.
- ▶ **The arXMLiv Project:** (<http://arxmliv.kwarc.info>)
 - ▶ use Bruce Miller's \LaTeX XML to transform to XHTML+MathML
 - ▶ we have an automated, distributed build system (ca. 1 CPU-year)
 - ▶ create ca. 8000 \LaTeX XML 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 papers by model assumptions (expanding, stationary, or contracting universe)
 - ▶ use linguistic techniques to add the necessary semantics

The MKM Authoring/Migration Problem

- ▶ Very interesting systems for Mathematical Knowledge Management (MKM)
- ▶ They promise to navigate/index/search/adapt/. . . large corpora of MK
- ▶ **Problem:** where is the beef?
- ▶ **Possible sources:**
 - ▶ libraries from theorem proving- and program verification and computer algebra systems (most of us do that)
 - ▶ Write your own in MathML/OpenMath/OMDoc/. . . (very tedious)
 - ▶ convert from SGML/Office engineering documents (difficult to get)
 - ▶ adapt from MS PowerPoint documents (works surprisingly well)
 - ▶ migrate from existing $\text{T}_\text{E}\text{X}/\text{L}\text{A}\text{T}_\text{E}\text{X}$ documents (There's the beef)
- ▶ $\text{T}_\text{E}\text{X}/\text{L}\text{A}\text{T}_\text{E}\text{X}$ is a power-user's interface to mathematics!

T_EX/L_AT_EX 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: $\binom{n}{k}$, ${}_n C^k$, C_n^n , and C_n^k 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 $f: S \rightarrow S \dots$ ” (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 T_EX/L_AT_EX

T_EX/L_AT_EX 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 T_EX/L^AT_EX 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 T_EX macro mechanism for this (well established)
 - ▶ the author knows the semantics best (at least she understands)
 - ▶ the burden is alleviated by manageability savings (MKM on T_EX/L^AT_EX)

sTeX Approach: **Semantic pre-loading** of T_EX/L^AT_EX documents.

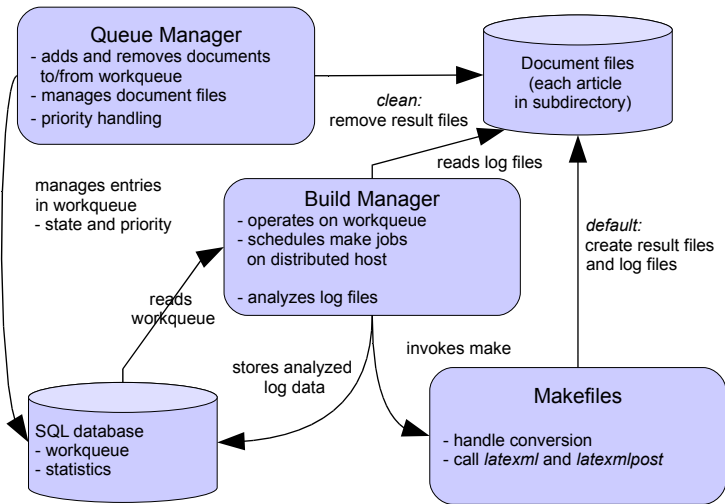
- ▶ Introduce semantic macros: e.g. `\union{a,b,c}` \rightsquigarrow `a ∪ b ∪ c`
- ▶ Mark up discourse structure: (largely invisible)
e.g. `\begin{proof}[id=Wiles,for=Fermat]... \end{sproof}`
- ▶ Generate PDF and XML from that (via L^AT_EXML [Miller])

The arXMLiv Build System [Stamerjohanns]

- ▶ **Problem: Size Matters!**: We do not have a good overview over the conversion task
 - ▶ over 6000 \LaTeX packages are used in the arXiv (which ones matter most?)
 - ▶ full conversion run takes multiple processor years. (turnaround time)
 - ▶ our intuitions about “ \LaTeX in the wild” are all wrong
- Idea**: distribute conversion task over computer cluster, crawl log files, aggregate results (build a conversion harness)

▶

The arXMLiv Build System [Stamerjohanns]



Converting T_EX/L_AT_EX Documents to XML

- ▶ Hermes [Anghelache] and T_EX4HT [Gurari] use the T_EX parser, seed the DVI file with semantic information, parse DVI for transformation.
- ▶ L_AT_EXML [Miller] and SGLR/Elan4 [van den Brand, Stuber] re-implement the T_EX parser. (do not expand semantic macros)
- ▶ **Case Study:** Converting Intro Computer Science to *OMDoc* via semantic pre-loading and L_AT_EXML
- ▶ **L_AT_EXML workflow:** (used in our case study)
 - ▶ L_AT_EXML $\hat{=}$ T_EX parser + XML emitter + post-processing pipeline.
 - ▶ L_AT_EXML bindings for the XML emitter (for all L_AT_EX packages as well)

```
DefConstructor("\Reals", "<XMTok name='Reals' />");  
DefConstructor("\SmoothFunctionsOn{}", "  
    <XMApp><XMTok name='SmoothFunctionsOn' />#1</XMApp>");  
DefMacro("\SmoothFunctionsOnReals", "\SmoothFunctionsOn\Reals");
```

Future Plans for arXMLiv

- ▶ **State:** \LaTeX -to-XHTML+MathML Format Conversion works (60% success)
- ▶ **Over the summer:** Bump up success rate to 70%, download 2007/2008 (another 85 Kpapers),...
- ▶ **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

Possible Applications

- ▶ 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 15 years; drink from the source)
- ▶ finding scientific communities of practice. (SCOOP)