Publishing Math Lecture Notes as Linked Data

Catalin David, Michael Kohlhase, Christoph Lange, Florian Rabe, and Vyacheslav Zholudev

Computer Science, Jacobs University Bremen,
{c.david,m.kohlhase,ch.lange,f.rabe,v.zholudev}@jacobs-university.de

Abstract. Lecture notes, particularly such with mathematical formulæ, are often written non-semantically in LaTeX and thus only really useful for reading and printing. Important questions of learners (e. g. "where is an example for the concept of structural induction") and lecturers (e. g. "what content from the repository can I reuse this year") cannot be answered automatically that way, but only by manual lookup. We convert a corpus of LaTeX lecture notes to semantic markup and expose them as Linked Data in XHTML+MathML+RDFa. Our demo application makes the resulting documents interactively browsable, and our ontology enables query answering and paves the path towards an integration of our corpus with external data sources.

1 Application: Computer Science Lecture Notes

Over the last seven years, the second author has accumulated a large corpus of teaching materials, comprising more than 2,000 slides, about 1,000 homework problems, and hundreds of pages of course notes, all written in LaTeX. The material covers a general first-year introduction to computer science, graduate lectures on logics, and research talks on mathematical knowledge management. This situation is typical for university educators and researchers and represents the state of the art in mathematical sciences (math, physics, computer science, some parts of engineering): LaTeX has proven suitable for writing high-quality lecture notes and publishing them as PDF, especially in our setting with a lot of mathematical formulæ. However, in our educational setting, we would like more support from the generated documents, given that screen reading and e-books support a much larger degree of interactivity.

For example, while reading notes students want to directly look up the definition or meaning of a symbol (e. g., $\models$) in a formula or examples for a difficult concept (e. g., structural induction). During a course they may want to query the whole body of lecture notes for a selection of advanced material that is appropriate for self-study based on the topics covered in the lecture. They want to use a semantic search engine to find related material in other universities' online course notes, on mathematical web sites, or Wikipedia. Lecturers want to query their repository for document parts that can be reused in an upcoming lecture given the prerequisites that students are expected to meet and the material that has already been covered. When giving a course for a special audience such as

mathematics for physicists, they want to draw examples from that domain even though they are less familiar with it. They also want to identify didactic gaps, such as concepts without examples, or unjustified proof steps. These services require semantic annotations in the lecture notes that are understandable for external search engines.

However, plain LaTeX is barely usable for anything *beyond* on-screen reading and printing. Even simple semantic annotations are uncommon, rare exceptions are the `\title` command making its meaning explicit or `\frac{a}{b}` focusing on functional structure instead of visual layout. This is especially problematic for symbols in formulæ, as they are often overloaded with multiple definitions or presentable using different notations. $\binom{n}{k}$ can be a vector or a binomial coefficient, and a French or Russian would rather write the latter as $\mathcal{C}_n^k$.

Therefore, we have developed a semantic representation of mathematical knowledge in LaTeX and a presentation process that preserves these semantic structures as Linked Data in the output, making them amenable to mashups that offer interactive exploration, as well as semantic searching and querying. These are based on an ontology for mathematical knowledge so that mathematical content can be linked across very different repositories.

2 Research Background and Related Work

The importance of LaTeX in scientific authoring and its extensibility by macros has motivated research on semantic extensions enabling modern publishing workflows. SALT (Semantically annotated LaTeX [7]) marks up rhetorical structures and fine-grained citations in scientific documents; however, its vocabulary is not extensible, whereas our own sTeX offers macros for introducing new mathematical symbols and reusing arbitrary metadata vocabularies. In mathematics e-learning systems like ActiveMath [1] and MathDox [15], students can explore lecture notes adapted to their previous knowledge and interactively solve exercises. These systems draw on a semantic representation of mathematical formulæ and higher-level structures, e. g. proof steps or course module dependencies, in standardized XML languages, e. g. OpenMath [22] and OMDoc [9]. They utilize semantic structures of mathematical knowledge but do not publish them in a standard representation like RDF, which would enable general-purpose queries beyond the built-in services, as well as integration with other systems on the web. The Linking Open Data movement promotes best practices for publishing semantic data on the web [13]; they can be published as standalone RDF, or embedded into HTML documents as RDFa [17]. Sindice is an example of a search engine that crawls and indexes Linked Data on the web [19]; the Sparks Ozone Browser is an example of a mashup that utilizes RDFa annotations in HTML documents for interactive browsing [20]. The design of our interactive documents is similar but additionally supports annotations in MathML formulæ. MathML has pioneered embedded annotations long before RDFa. Its *parallel markup* interlinks both the rendered appearance and the semantic structure of mathematical expressions, where the meaning of mathematical symbols is usually

defined in lightweight ontologies called OpenMath Content Dictionaries [16]. However, hardly any mathematical software that uses a semantic representation of mathematical formulæ internally also exposes it on the web. HELM (Hypertext Electronic Library of Mathematics [2]) pioneered the use of RDF for representing structures of mathematical knowledge, e. g. in what mathematical theory a symbol is introduced, what of its properties have been declared or asserted, and how the latter are proved. The HELM ontology, however, has not gained wide acceptance, and at the time of its development, there was no RDFa-like standard for embedding RDF into published documents.

3 Architecture and Demo

Our architecture publishes semantically enriched LaTeX lecture notes as XHTML+MathML+RDFa Linked Data. We kept LaTeX as an input language, as it is familiar to authors and well supported by editors, and as high-quality PDF can be obtained from it. With sTeX (semantically enhanced TeX), we have introduced LaTeX macros for marking up the semantic structure of mathematical formulæ and documents [10]. One can, e. g., declare a symbol *union*, formally define it, and make its semantic representation `\union{A,B}` expand to `A\cup B` for human-readable rendering. There are environments for mathematical statements and theories, e. g. `\begin{example}[for=union]`. We transform this into a semantically equivalent intermediate XML representation using LaTeXML. We use the standard XML languages OpenMath for formulæ [22] and OMDoc for higher-level structures [9]. Finally, our JOMDoc rendering library [8] generates human-readable output from this XML – an output that still contains the full semantic structure as annotations. A custom Java implementation renders formulæ as parallel markup of Presentation MathML annotated with OpenMath[1]; rendering higher-level structures as XHTML+RDFa [17] is implemented in XSLT. RDF, both serialized as RDF/XML and the machine-friendly RXR [3], is extracted from XML by our Krextor XML→RDF library [12]. It uses our OMDoc ontology as a vocabulary for representing mathematical structures (e. g. "d is a definition, e is an example for d"), inspired by HELM and designed as a more expressive counterpart of the OMDoc XML schema.

The whole transformation process (fig. 1) is integrated into our versioned XML database TNTBase [23, 24] and available at `http://kwarc.info/LinkedLectures`. TNTBase has a Subversion-compatible interface making it suitable as a lecture notes repository. The TeX→XML and XML→RDF transformations are automatically triggered by a hook upon committing a new revision of an sTeX lecture module. If the generated OMDoc+OpenMath is not schema-valid, the commit is rejected. On the other hand, it follows Linked Data best practices and, depending on the MIME type an HTTP client requests, serves a document as

[1] A proposal for fully representing formulæ in RDF [14] has not gained wide acceptance. RDF-based reasoners often limited to decidable first order logic subsets, which is insufficient for mathematical applications, and XML has a more straightforward notion of order (e. g. of the arguments of a mathematical function or of a set constructor).

OMDoc, as RDF (only a structural outline, not the full text and formulæ), or as XHTML+MathML+RDFa. The latter contains JavaScript code from our JOBAD library for interactive documents [11, 5], which operationalizes the annotations – Linked Data and other – in the rendered documents. JOBAD's definition lookup determines the OpenMath annotation of the Presentation MathML symbol the user clicked on, from that obtains the URI of the symbol, and then requests XHTML from that URI (resulting in the symbol's declaration and definition), which is then displayed in a popup. The RDFa annotations are currently used for making parts of a document (e. g. steps of a structured proof) foldable, and for displaying the local neighborhood in the RDF graph (e. g. related examples) in popups; this is implemented using the rdfQuery library [18], relying on the Linked Data structure in the latter case. Further third-party services can be integrated in a mashup style; we have demonstrated this for a unit conversion service [11, 5]. Besides JOBAD's services, the linked RDF data can also be crawled and put into a triple store for global queries, e. g. "find examples for all concepts from graph theory (about which I'm planning a lecture), assuming as prerequisites the concepts from formal languages (and their prerequisites)". This would yield the parse tree of a context-free language as an example for the concept "tree" – as operating systems were not among the prerequisites.

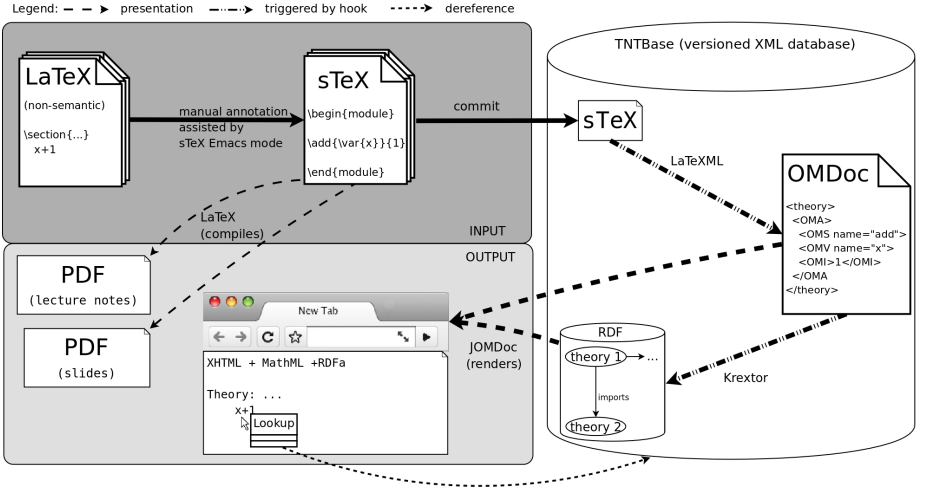

Fig. 1. The complete LaTeX→OMDoc→RDF→Linked Data pipeline

Our demo shows the complete pipeline in action: (i) annotating a document with our sTeX Emacs mode, (ii) committing it to TNTBase, (iii) automatic translation to OMDoc, schema validation, and RDF extraction, (iv) retrieving the document in different representations, (v) browsing the XHTML+MathML+RDFa rendering, (vi) and interacting with the Linked Data in it. Additionally, we will

demonstrate the generation of PDF from the sTEX sources, as well as possibilities for querying the extracted RDF using SPARQL.

4 Conclusion and Outlook

Our architecture makes legacy TEX-based lecture notes available as Linked Data. We expose these data to external clients but have also implemented services for interactively exploring the XHTML+MathML+RDFa presentation of our data. We are also working on preserving some of the semantics in the PDF output, as SALT does. Evaluation of our enriched lecture notes by the student end users is still pending.

To the best of our knowledge, we are the first provider of RDF-based Linked Open Data in the domain of mathematics and among the first to operationalize the Linked Data structures of formula markup. Having successfully transformed more than 300,000 normal, non-semantic LATEX documents from `arxiv.org` to XHTML+Presentation MathML [21] and working on machinery for automatically annotating them using natural language processing [6], we will soon be able to expose an even larger collection of mathematical knowledge as Linked Open Data. The planned exposition of the RDF extracted from OMDoc via a SPARQL endpoint will facilitate federated queries and thus integration with other data sources, such as DBpedia [4], whose mathematical knowledge does not have a semantics as strong as ours, but which provides abundant informal background knowledge, e. g. about the originators of mathematical theories. On the other hand, hardly any well-known mathematical site (e. g. `planetmath.org` and `mathworld.wolfram.com`) currently exposes machine-understandable metadata. We promote our technology, starting with lightweight RDFa annotation using the OMDoc ontology, as a migration path towards their integration into a true mathematical semantic web.

References

1. ActiveMath. URL: `http://www.activemath.org`.
2. A. Asperti, L. Padovani, C. Sacerdoti Coen, F. Guidi, and I. Schena. Mathematical Knowledge Management in HELM. In: *Annals of Mathematics and Artificial Intelligence* 38.1–3 (2003).
3. D. Beckett. Modernising Semantic Web Markup. In: *XML Europe*. 2004.
4. *DBpedia*. URL: `http://dbpedia.org`.
5. J. Giceva, C. Lange, and F. Rabe. Integrating Web Services into Active Mathematical Documents. In: *MKM/Calculemus 2009*. LNAI 5625. 2009.
6. D. Ginev, C. Jucovschi, S. Anca, M. Grigore, C. David, and M. Kohlhase. An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus. In: *AST Workshop at Informatik*. 2009.
7. T. Groza, S. Handschuh, K. Möller, and S. Decker. SALT – Semantically Annotated LATEX for Scientific Publications. In: *ESWC*. LNCS 4519. 2007.
8. *Java Library for OMDoc documents*. URL: `http://jomdoc.omdoc.org`.

9. M. Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. 2006.
10. M. Kohlhase. Using LaTeX as a Semantic Markup Format. In: *Mathematics in Computer Science* (2008).
11. M. Kohlhase, J. Giceva, C. Lange, and V. Zholudev. JOBAD – Interactive Mathematical Documents. In: *AI Mashup Challenge*. 2009.
12. C. Lange. Krextor – An Extensible XML→RDF Extraction Framework. In: *Scripting and Development for the Semantic Web (SFSW)*. 2009.
13. *Linked Data*. URL: `http://linkeddata.org/guides-and-tutorials`.
14. M. Marchiori. The Mathematical Semantic Web. In: *Mathematical Knowledge Management*. LNCS 2594. 2003.
15. *MathDox – Interactive Mathematics*. URL: `http://www.mathdox.org`.
16. *MathML Version 3.0*. Candidate Recommendation. W3C, 2009.
17. *RDFa in XHTML: Syntax and Processing*. Recommendation. W3C, 2008.
18. *rdfQuery*. URL: `http://code.google.com/p/rdfquery/`.
19. *Sindice – The Semantic Web Index*. URL: `http://sindice.com`.
20. *Sparks O_3 Browser*. URL: `http://oak.dcs.shef.ac.uk/sparks/`.
21. H. Stamerjohanns, M. Kohlhase, D. Ginev, C. David, and B. Miller. Transforming large collections of scientific publications to XML. In: *Mathematics in Computer Science* (2009).
22. *The Open Math Standard, Version 2.0*. Tech. rep. Open Math Society, 2004. URL: `http://www.openmath.org/standard/om20`.
23. V. Zholudev and M. Kohlhase. TNTBase: a Versioned Storage for XML. In: *Balisage Markup Conference*. 2009.
24. V. Zholudev, M. Kohlhase, and F. Rabe. A [insert XML Format] Database for [insert cool application]. In: *XML Prague*. 2010.