

Krextor – An Extensible Framework for Contributing Content Math to the Web of Data

Christoph Lange

Computer Science, Jacobs University Bremen,
ch.lange@jacobs-university.de

1 Problem Statement: No Math on the Web of Data Yet

An increasing amount of scientific knowledge is being contributed to the emerging Web of [Linked] Data, where it is made available in a machine-comprehensible way and interlinked with other related datasets. This already powers distributed query answering engines and intelligent semantic mashups enriching web publications – however, it still largely lacks mathematical functionality.¹ There are e-science datasets – with mathematical model descriptions opaque to machines. There are statistical datasets, e.g. from e-government – without explicit descriptions of how values have been derived. There are digital libraries and databases of scientific publications – with information about who cited your paper, but not who is building on your mathematical *ideas*.

2 The Krextor XML→RDF Extraction Library

In contrast to the document-oriented, often XML-based content markup of MKM, the graph-based RDF data model is most widely used for representing knowledge on the Web of Data. Therefore, in order to contribute mathematical knowledge to the Web of Data, we have developed the Krextor [3] library, and, on top of that, extraction modules that translate the structural outlines of OpenMath, OMDoc, and other content markup to RDF. Krextor is an XSLT library that aims at facilitating the repetitive task of implementing translations from several XML input languages. It does so by offering convenience templates and functions for frequently occurring patterns in XML→RDF translation, such as creating RDF resources for things represented by XML elements, generating (“minting”) linked data compliant URIs for these resources, and translating XML text nodes or attribute to properties of these resources. Krextor allows for flexible integration into applications by supporting multiple output serializations of the RDF extracted, including callbacks to Java application code – whereas traditional hard-coded XSLT implementations would rather translate from exactly one XML

¹ For a review of the state of the art of linked data, we refer to [2], and for further background about the potential of *mathematical* linked data to [6].

input language to exactly one RDF output serialization (e.g. RDF/XML). Besides OpenMath and OMDoc², we have developed extraction modules for special MKM applications, such as encoding semantic web ontologies in OMDoc, and external developers have adopted Krextor outside of MKM [3].

3 Publishing the OpenMath CDs as Linked Data

In contrast to previous work [5], the current focus of Krextor development is on expanding the coverage of the OpenMath 2 CD language (and proposed extensions beyond that), in order to prepare the publication of the official CDs at openmath.org as linked open data [7]. The official OpenMath CDs have a great potential for bootstrapping a mathematical Web of Data, as they are widely in use, e.g. in that they define the semantics of Content MathML 3 [1].

Krextor has been used with OpenMath CDs before, but specifically for maintaining the (then) experimental collection of “OpenMath/MathML 3 CDs” *inside* a closed semantic wiki [8], which pre-processed them in a special way. The current focus is on making most out of the official OpenMath CDs *as they are*, which means:³ (i) Supporting the maintenance of links from concepts in the OpenMath CDs to semantically equivalent concepts in related datasets – such as the Digital Library of Mathematical Functions (DLMF [9]) or the PlanetMath encyclopedia [11]. As the reference encoding of the OpenMath CD model [12] does not currently have annotation facilities, this is done as standoff markup in separate RDF files next to the CDs.⁴ An example is the definition of the sine function in terms of the exponential function ($\sin z = \frac{e^{iz} - e^{-iz}}{2i}$); the correspondence between its OpenMath and DLMF representations is expressed by the RDF triple `<http://dlmf.nist.gov/4.14.E1> owl:sameAs <http://www.openmath.org/cd/transcl#sin.prop0>`. (ii) A prerequisite for that: Giving stable identifiers to mathematical properties of symbols – even though the reference CD encoding does not provide such identifiers. This is important as, for example, the DLMF entries mainly correspond to OpenMath mathematical properties [7]. (iii) Utilizing existing XSLT code for translating OpenMath objects into Content MathML and other machine-comprehensible representations [10], so that interested applications can retrieve them right from the same dataset.

² cf. [4, chapter 3] for a detailed description of the target RDF vocabularies/ontologies that we have developed for capturing mathematical knowledge, or for mappings to existing RDF vocabularies that we reused, e.g. for metadata

³ See <http://trac.kwarc.info/krextor/wiki/OpenMathExtractionModule> for a technical documentation and examples.

⁴ These links have to be maintained manually for now; automatically identifying such correspondences between would require advanced linguistic methods.

4 Coverage of the System Demo

The demo will focus on (i) the OpenMath CD extraction module, but also on (ii) Krextor’s possibilities for implementing extraction modules for new MKM languages. Regarding (i), I will particularly explain how to create new links between the OpenMath CDs and external datasets, and how RDF- and/or OpenMath-aware client applications can utilize the OpenMath CD linked dataset. Regarding (ii), I am prepared for a “hacking session” with any visitors who are interested in extracting RDF from their XML-based MKM language, in order to contribute their mathematical knowledge collections to the Web of Data.

References

- [1] *Mathematical Markup Language (MathML) Version 3.0*. W3C Recommendation. 2010. URL: <http://www.w3.org/TR/MathML3>.
- [2] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011. URL: <http://linkeddatabook.com>.
- [3] *Krextor – The KWARC RDF Extractor*. URL: <http://kwarc.info/projects/krextor/> (visited on 12/06/2010).
- [4] C. Lange. “Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration”. submitted January 31, defended March 11. PhD thesis. Jacobs University Bremen, 2011. URL: <https://svn.kwarc.info/repos/swim/doc/phd/phd.pdf>.
- [5] C. Lange. “Krextor – An Extensible XML→RDF Extraction Framework”. In: *Scripting and Development for the Semantic Web (SFSW)*. CEUR Workshop Proc. 449. 2009. URL: <http://CEUR-WS.org/Vol-449/>.
- [6] C. Lange. “Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web”. *Semantic Web Journal* (accepted). 2011. URL: <http://www.semantic-web-journal.net/content/new-submission-ontologies-and-languages-representing-mathematical-knowledge-semantic-web>.
- [7] C. Lange. “Towards OpenMath Content Dictionaries as Linked Data”. In: *23rd OpenMath Workshop*. 2010. arXiv:1006.4057v1 [cs.DL].
- [8] C. Lange and A. González Palomo. “Easily Editing and Browsing Complex OpenMath Markup with SWiM”. In: *Mathematical User Interfaces Workshop*. 2008. URL: <http://www.activemath.org/workshops/MathUI/08/proceedings/LangeGonzales-OMEdit.html>.
- [9] National Institute of Standards and Technology, ed. *Digital Library of Mathematical Functions*. May 7, 2010. URL: <http://dlmf.nist.gov>.
- [10] OpenMath Society. URL: <http://www.openmath.org/standard/omxsl/> (visited on 08/09/2010).
- [11] *PlanetMath.org – Math for the people, by the people*. URL: <http://planetmath.org> (visited on 01/06/2011).
- [12] *The Open Math Standard, Version 2.0*. Tech. rep. The OpenMath Society, 2004. URL: <http://www.openmath.org/standard/om20>.