Integrating Mathematics into the Web of Data

Christoph Lange

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

Abstract. Mathematics, a ubiquitous foundation of science, technology, and engineering, is currently missing from the Web of Data. We argue why future internet applications would benefit from mathematical Linked Open Data. Due to the predominance of semantic XML-based markup languages in mathematics – which has good reasons –, contributing mathematical knowledge to the Weeb of Data cannot exclusively rely on RDF. While that is achievable within the current Linked Data architecture, better application-layer support would be desirable. This is a summary of our position; we refer to [26] for full background.

1 A Short (Pre-)History of Mathematics on the Web

Understanding the potential benefits of the Web of Data and mathematics for each other requires a brief review of the huge amount of mathematical knowledge and its applications on the Web: The Web 1.0 arχiv [43] hosts 600,000 mathematical/scientific pre-prints, mostly encoded in LaTeX – excellent for *publishing* but incomprehensible to automated computation or reasoning engines. The mathematical community has collaboratively developed or reviewed proofs in the Polymath Web 2.0 wiki/blog [21], e.g. recently for a claim of $P \neq NP$ [17]. Mathematical articles in the Wikipedia and PlanetMath [35] Web 2.0 encyclopediæ also contain LaTeX formulæ. The Connexions courseware repository [12] recommends Content MathML instead; it *could* enable machine services such as interactive computer algebra (see, e.g., [14]), but Connexions merely uses it for a limited adaptation of the presentation of formulæ. Finally, formalized and machine-verified mathematical libraries, e.g. Mizar [28], are available on the Web 1.0 but only fully supported by specialized tools requiring local installation.

The Mathematical Knowledge Management (MKM) community was an early adopter of the Semantic Web. However, the immaturity of the technology at that time – lacking reliable querying support and Web 2.0 user interfaces – frustrated the hopes set into the MathNet [24] distributed information and communication infrastructure, the HELM digital library [23], and the MONET computation web service architecture [29]. The MKM community has given up using Semantic Web technologies on a large scale after 2004, and the Semantic Web community and funding bodies have focused on different use cases.

2 Why Mathematics on the Web of Data?

Mathematics is a ubiquitous foundation of science, technology, and engineering. Having not only application areas of mathematics but also their mathematical

foundations well represented on the Web of Data would enable the following applications:

General-Purpose Mathematical Knowledge: Wikipedia states the Pythagorean theorem as $a^2 + b^2 = c^2$ (in LaTeX) and categorizes it as "Article containing proofs" and "Mathematical theorem". A search for the semantically equivalent expression $z = \sqrt{x^2 + y^2}$ would fail. From the categorization it is not clear for a machine whether the article contains a [correct] proof of *that* theorem. The same restrictions apply to DBpedia, the Linked Dataset obtained from Wikipedia [16]. That forced the Polymath collaborators to search for previous publications of refutations of $P \neq NP$ "proofs" by keyword.

Statistics: Public sector information, increasingly being published as Linked Data by governments, has been used to provide, e.g., localized information retrieval about crime statistics and hospital waiting list statistics [32]. Statistical datasets contain values derived from ground values, or from other derived values using mathematical functions. As planning data collection and interpreting collected data requires mathematical models, statistical datasets need a notion of mathematical semantics [45, 27].

Publication Databases: The RKB Explorer Linked Dataset [2] classifies ACM's scientific publications according to the ACM Computing Classification System [42]. Still, it is impossible for a Linked Data agent to understand that a publication merely classified as "F.1.3 Complexity Measures and Classes" deals with the P and NP complexity classes, and how they are defined.

Enterprise Applications: Linked Data do not have to be open; the architecture also works in enterprise intranets. Renault has used them for retrieving information about spare car parts [37]. Now consider decisions to be made when designing whole cars: They ultimately require mathematical understanding. An engineer looking for an efficient engine for a projected city car might feed inputs such as the weight of the car, the average length and duration of a trip, the most widely available type of fuel and the average environment temperature when starting the engine into a mathematical model of the engine in order to predict its fuel consumption under these constraints.

e-Science: The above use case is actually about reproducing an experiment – one of the key principles of e-Science [5]. Fine-grained reproducibility once more demands a representation of the mathematical models. Some e-science datasets include them, e.g. the SysMO SEEK "'assets catalogue' describing data, models, [...], workflows and experiment[s]" [5] from systems biology of microorganisms [41]. Publishing that as Linked Data is in progress (cf. [5]). Currently, the mathematical models are given as Content MathML formulæ deeply nested into XML files and thus not directly accessible via URIs.

3 Representing Mathematical Knowledge on the Web

Where mathematical knowledge is currently represented machine-comprehensibly, it is usually done in the native languages of computer algebra systems or proof

assistants, or in Content MathML (the "semantic" sublanguage of MathML [4]) or OpenMath [11] semantic XML markup languages. Translations from the native languages of many mathematical systems to these exchange languages and vice versa are available. Content MathML represents formulæ as functional trees. It has a built-in vocabulary of symbols (operators, functions, etc.) from high-school and introductory university mathematics and relies on the OpenMath Content Dictionary (CD) extension mechanism otherwise. OpenMath CDs are mathematical domain ontologies. A canonical URI format for symbols in CDs ensures minimum Semantic Web compatibility; e.g., the addition operator – the *plus* symbol in the *arith1* CD – has the URI `http://www.openmath.org/cd/arith1#plus`. The official peer-reviewed CD collection defines 260 symbols from arithmetics, set theory, first-order logic, algebra, calculus, as well as transcendental and statistical functions [34].[1]

Ontologies for representing mathematical knowledge in RDF exist, sometimes derived from these markup languages [26]. Despite standard URI formats for mathematical symbols, *formulæ* have always constituted a problem. Their n-ary ordered tree structure precludes a straightforward RDF representation. Linked lists or ordered sets – RDF's built-in collections or sequences [6] or self-made remakes – are unavoidable. These are, however, badly supported by software and do not go well along with querying[2] and DL reasoning[3]. Full RDF representations of formulæ can be found in the N3 language supported, e.g., by the cwm [8] reasoner; however, the syntactic sugar that makes these representations relatively intuitive is not available in plain RDF. Encodings of Content MathML, using RDF collections, have been suggested (see, e.g., [36]), but not adopted in practice.

An alternative – suggested once [10], but never done in practice – is representing formulæ as Content MathML/OpenMath XML literals in RDF while representing anything else in pure RDF. However, XML literals are largely opaque to contemporary RDF tools. Virtuoso [33] allows for filtering XML literals matched by a SPARQL graph pattern by XPath node tests [7]. Corese can additionally reuse variables from the proper SPARQL part of a query in XPath expressions [13]. None of these extensions has made it into SPARQL yet.

RDFa [1] would allow for leaving the representation of n-ary and ordered structures to the embedding XML. The RDFa 1.1 API [39], once implemented by browsers, will at least give in-browser scripts similar means of accessing embedded RDF as the Document Object Model (DOM) does for XML. The XSPARQL [3] query language combines SPARQL and XQuery; however, queries would still rely

[1] The OpenMath CD language with its support for semi-formal descriptions of symbols and their mathematical properties is sufficiently expressive to allow for simple computations and to create requirements specifications for implementors of the above-mentioned translations from and to mathematical systems, but not to support fully automated inference. The OMDoc language [31, 25] adds the latter layer to OpenMath and also builds a bridge to LaTeX by covering informal text.

[2] At least support for querying RDF collections, which some query processors already support by non-standard extensions, will be standardized in SPARQL 1.1 [22].

[3] In OWL one has to avoid RDF collections, as the RDF encoding of OWL uses them internally to represent n-ary expressions. Self-made linked lists work around that [18].

on a separate service that makes the RDFa annotations available as queryable RDF. RDFa has not yet been used for representing formulæ either. Neither the MathML nor the OpenMath developers are currently planning to support RDFa.[4]

The only approach that has actually been employed in practice is standoff markup. An RDF graph points to XML fragments and adds information to them, e.g. additional metadata or links not supported by the respective XML language, whereas n-ary structures and order are only represented in XML. Most of the knowledge is represented redundantly in RDF and XML – one of them possibly automatically generated from the other one – to provide maximum information to agents that only understand one representation. For example, the HELM and MONET RDF representations of formulæ have focused on symbols in key positions, e.g. the root of the assumption part of an inference rule, for the purpose of, e.g., finding applicable theorems, or for matching mathematical problems (e.g. definite integration) to web services solving them.

4 Challenges to Publishing Mathematical Linked Data

This section summarizes the ways of publishing mathematical Linked Data that we have explored so far and the challenges that we have encountered in doing so.

In [45, 27], we describe a scenario where an agent accesses both RDF datasets and OpenMath CDs by dereferencing URIs and using content negotiation. The rules for computing derived values in statistical datasets are represented as RDF annotations pointing to a function – a symbol from an OpenMath CD – and other values from the dataset that should be passed as arguments to the function. An agent that wants to verify a derived value has to construct an OpenMath formula from this RDF representation and send it to an OpenMath computation service. For functions called using named arguments, the agent has to consult the XML representation of the CD to get their order right.

In a different setting, we have published human-readable XHTML+MathML documents, where semantic annotations – OpenMath for formulæ and RDFa for the rest – act as anchors for assistive services that provide additional information on demand or adapt the presentation of a document [15]. This can be combined with XML/RDF content negotiation. We have realized interactive declaration lookup for symbols in formulæ by dereferencing their canonical URIs (pointing to a symbol declaration in a CD) from the formula's annotation and transforming the OpenMath declarations thus obtained to human-readable Presentation MathML.

Besides the above-mentioned restrictions in querying XML/RDF combinations, we have identified three challenges to publishing mathematical Linked Data:

Missing MIME Types: Content negotiation distinguishes representation formats by MIME type. MathML 3 has officially registered MIME types [4], whereas an OpenMath MIME type has merely been proposed so far [27].

[4] This may be due to the fact that MathML is already at least as expressive as RDFa. Besides Content MathML for functional trees, it has a fine-grained general-purpose annotation mechanism, which has actually existed long before RDFa. OpenMath has a similar annotation syntax, albeit without URI support.

Bad Authoring Practices – from a Linked Data point of view – result from authors of mathematical XML markup using URIs wrongly or not at all. Hardly any community-contributed OpenMath CDs specifies a base URI or references symbols by absolute URIs, which indicates a lack of awareness. The fallback base URI `http://www.openmath.org/` is, even independently from Linked Data considerations, not suitable for CDs from developers who do not own the `openmath.org` domain. Finally, authors who are aware of CDs and symbols having URIs usually merely consider them globally unique *names* without relevance for retrieving information about these resources [27].

URI Format Restrictions: Thirdly, while RDF publishers can freely choose URIs [9], the URI formats of non-RDF languages often impose restrictions that complicate Linked Data publishing and have to be worked around. OpenMath's *base/cd#symbol* URI schema complies well with Linked Data practices – unless CDs grow large. As resolving fragments is up to the client, consequently using hash URIs forces clients to always download a complete CD, in which they could then locate the desired symbol. Publishers of large CDs could work around that by serving, upon an initial request, an RDF graph that merely redirects, via *rdfs:seeAlso* links, hash to slash URIs, from which the client would then be able to retrieve the desired fine-grained information.[5] A final problem with old languages such as the OpenMath CD language, is that certain entities – e.g. properties of symbols – cannot be given IDs. An XML→RDF translator might generate some, but an agent interested in retrieving XML representations would also need them. As a use case for such fine-grained links, consider the (currently Web 1.0) Digital Library of Mathematical Functions (DLMF [30]). It contains a large number of equations describing or defining mathematical functions, which could be linked to the corresponding properties in OpenMath CDs.

5 Conclusion and Future Work

We have made the case for mathematical knowledge on the Web of Data by outlining potential applications, and mentioned current challenges to publishing mathematical Linked Data. Doing so requires combining XML and RDF, particularly due to the inherent structural complexity of mathematical formulæ and the key role of XML exchange languages in mathematics. Publishing mathematical knowledge is feasible within the current Linked Data architecture, but it would benefit from better application-layer support for dealing with combinations of XML and RDF, and with impractically restricted legacy URI formats.

We conclude with an agenda towards publishing relevant mathematical datasets. The probably most foundational dataset needed to get mathematics on the Web of Data started is the official OpenMath CDs. The initial publication, planned for spring 2011, is, however, only the first step in making the knowledge from the CDs accessible; the second step is linking mathematics-related existing datasets to the OpenMath CDs, so that services for these existing datasets can

[5] See [26] for a discussion of further problems, also in related languages.

be extended by mathematical functionality – as outlined for statistical datasets above. DBpedia is a further candidate, as that would offer its large audience a more formal perspective on mathematics.

Mathematical knowledge collections that are already available on the Web, but not currently in a semantic representation, should also be triplified, at least with shallow mathematical metadata and links to relevant OpenMath CDs. The DLMF could, e.g., benefit from access to OpenMath computation services, whereas the benefit for PlanetMath would be similar as for DBpedia. We should also take a serious view on the April fool's joke "Linked Open Numbers", a huge dataset describing billions of natural numbers [44]. It provides descriptions as trivial as the name of each number in natural language, and its successor. But how about a dataset of non-trivial properties of numbers? Accessing, e.g., prime factor decompositions of large numbers – an information relevant for cryptography – in a linked dataset, could be much faster than computing it once more, provided a supercomputer has already done the computation once and published the results. From an RDF reasoning and querying point of view, such a dataset could serve as an *oracle*, providing information whose original computation would by far exceed, e.g., description logic reasoning. Another such source of non-trivial knowledge about numbers is the Web 1.0 Online Encyclopedia of Integer Sequences [38].

Related to information retrieval and computation is the development of suitable query languages and reasoners. N3 reasoners already support a basic set of mathematical functions; SPARQL 1.1 supports basic arithmetics. Additionally, many query processors allow for defining extension functions; a path for supplying arbitrary functions to query processors via OpenMath should be investigated. Such an extension could even be specified formally as an entailment regime [20][6].

The arχiv offers a path towards publishing mathematical structures of scientific publications. A long-term effort to automatically recover their mathematical structure from the LaTeX sources is in progress [19], the translation of 300,000 publications to somewhat more semantically structured XHTML+MathML being a first success [40]. The publications have stable URIs, and their metadata are available as XML, which makes a Linked Data publication feasible right now. Next, much harder steps would be interlinking with existing publication datasets, e.g. DBLP, and identifying mathematical symbols that could be linked to the OpenMath CDs. With an identification of inference structures, this would ultimately enable machine support for the next collaborative Web-based review of a $P \neq NP$ proof.

References

[1] *RDFa Core 1.1. Syntax and processing rules for embedding RDF through attributes.* Working Draft. URL: http://www.w3.org/TR/rdfa-core/.

[2] *acm.rkbexplorer.com.* URL: http://acm.rkbexplorer.com.

[3] W. Akhtar et al. "XSPARQL: Traveling between the XML and RDF worlds – and avoiding the XSLT pilgrimage". In: *ESWC*. Springer, 2008.

[6] originally suggested by Denny Vrandečić

[4] *Mathematical Markup Language (MathML) 3.0*. Recommendation. URL: http://www.w3.org/TR/MathML/.

[5] S. Bechhofer et al. "Why Linked Data is Not Enough for Scientists". In: *6th IEEE e-Science conference*. 2010.

[6] *RDF/XML Syntax Specification (Revised)*. Recommendation. URL: http://www.w3.org/TR/rdf-syntax-grammar/.

[7] *XML Path Language (XPath) 2.0*. Recommendation. URL: http://www.w3.org/TR/xpath20/.

[8] T. Berners-Lee. *cwm. a general purpose data processor for the semantic web*. URL: http://www.w3.org/2000/10/swap/doc/cwm.html.

[9] *Best Practice Recipes for Publishing RDF Vocabularies*. Working Group Note. URL: http://www.w3.org/TR/swbp-vocab-pub/.

[10] S. Buswell. *RDF/XML Test cases for RDF Logic, Web Ontology and Maths content*. Deliverable 5.3b. SWAD-Europe, 2001. URL: http://www.w3.org/2001/sw/Europe/reports/xml_test_cases/wp53.html.

[11] S. Buswell et al. *The Open Math Standard 2.0*. Tech. rep. The OpenMath Society, 2004. URL: http://www.openmath.org/standard/om20.

[12] *Connexions*. URL: http://cnx.org.

[13] O. Corby et al. *Querying the Semantic Web of Data using SPARQL, RDF and XML*. Tech. rep. 6847. INRIA Sophia Antipolis, Feb. 2009. URL: http://hal.inria.fr/docs/00/36/23/81/PDF/RR-6847.pdf.

[14] H. Cuypers et al. "MathDox, a system for interactive Mathematics". In: *ED-MEDIA*. AACE, June 2008. URL: http://go.editlib.org/p/29092.

[15] C. David et al. "Publishing Math Lecture Notes as Linked Data". In: *ESWC (Part II)*. Springer, 2010.

[16] *DBpedia*. URL: http://dbpedia.org.

[17] *Deolalikar P vs NP paper*. URL: http://michaelnielsen.org/polymath1/index.php?title=Deolalikar_P_vs_NP_paper&oldid=3654.

[18] N. Drummond et al. "Putting OWL in Order: Patterns for sequences in OWL". In: *OWL: Experiences and Directions (OWLED)*. Nov. 2006.

[19] D. Ginev et al. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *AST Workshop at Informatik*. 2009. URL: http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf.

[20] *SPARQL 1.1 Entailment Regimes*. Working Draft. URL: http://www.w3.org/TR/sparql11-entailment/.

[21] T. Gowers and M. Nielsen. "Massively collaborative mathematics". In: *Nature* 461.15 (2009).

[22] *SPARQL 1.1 Query Language*. Working Draft. URL: http://www.w3.org/TR/sparql11-query/.

[23] *HELM. Hypertextual Electronic Library of Mathematics*. URL: http://helm.cs.unibo.it.

[24] *Math-Net. an International Information and Communication System*. URL: http://www.math-net.org.

[25] M. Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[26] C. Lange. "Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web". submitted to Semantic Web Journal. 2011. URL: http://www.semantic-web-journal.net/content/new-submission-ontologies-and-languages-representing-mathematical-knowledge-semantic-web.

[27] C. Lange. "Towards OpenMath Content Dictionaries as Linked Data". In: *OpenMath Workshop*. July 2010. arXiv:1006.4057v1 [cs.DL].

[28] *Mizar Mathematical Library*. URL: http://www.mizar.org/library.

[29] *MONET – Mathematics on the net*. URL: http://monet.nag.co.uk.

[30] *Digital Library of Mathematical Functions*. URL: http://dlmf.nist.gov.

[31] *OMDoc*. URL: http://omdoc.org.

[32] T. Omitola et al. "Put in Your Postcode, Out Comes the Data: A Case Study". In: *ESWC (Part I)*. Springer, 2010.

[33] OpenLink Software. *OpenLink Universal Integration Middleware – Virtuoso Product Family*. URL: http://virtuoso.openlinksw.com.

[34] OpenMath *CDs*. URL: http://www.openmath.org/cd/.

[35] *PlanetMath.org – Math for the people, by the people*. URL: http://planetmath.org (visited on 2010-07-14).

[36] A. Robbins. *Semantic MathML*. URL: http://straymindcough.blogspot.com/2009/06/semantic-mathml.html.

[37] F.-P. Servant. "Linking Enterprise Data". In: *LDOW*. CEUR-WS 369. Apr. 2008. URL: http://CEUR-WS.org/Vol-369/.

[38] N. J. A. Sloane. "The On-Line Encyclopedia of Integer Sequences". In: *Notices of the AMS* 50.8 (2003).

[39] *RDFa API. An API for extracting structured data from Web documents*. Working Draft. URL: http://www.w3.org/TR/rdfa-api/.

[40] H. Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. URL: http://kwarc.info/kohlhase/papers/mcs09.pdf.

[41] *SysMO-DB SEEK*. URL: http://www.sysmo-db.org/seek/.

[42] *The 1998 ACM Computing Classification System*. 1998. URL: http://www.acm.org/about/class/ccs98.

[43] arXiv.org *e-Print archive*. URL: http://www.arxiv.org.

[44] D. Vrandečić et al. "Leveraging Non-Lexical Knowledge for the Linked Open Data Web". In: *RAFT*. 2010. URL: http://km.aifb.kit.edu/projects/numbers/linked_open_numbers.pdf.

[45] D. Vrandečić et al. "Semantics of Governmental Statistics Data". In: *WebSci*. 2010. URL: http://journal.webscience.org/400/.