

Lecture Notes in Artificial Intelligence 4180

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Michael Kohlhase

OMDoc – An Open Markup Format for Mathematical Documents

[version 1.2]

Foreword by Alan Bundy

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Author

Michael Kohlhase
International University Bremen
Computer Science
Campus Ring 1, 28759 Bremen, Germany
E-mail: m.kohlhase@iu-bremen.de

Library of Congress Control Number: 2006931135

CR Subject Classification (1998): I.2, F.4, F.3.1, G.4, H.3, I.1, I.7

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-37897-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-37897-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version. This work is licensed by the Creative Commons Share-Alike license <http://creativecommons.org/licenses/by-sa/2.5/>. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© by author 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Markus Richter, Heidelberg
Printed on acid-free paper SPIN: 11826095 06/3142 5 4 3 2 1 0

To Andrea — my wife, collaborator, and best friend — for all her support

Foreword

Computers are changing the way we think. Of course, nearly all desk-workers have access to computers and use them to email their colleagues, search the Web for information and prepare documents. But I'm not referring to that. I mean that people have begun to think about what they do in computational terms and to exploit the power of computers to do things that would previously have been unimaginable.

This observation is especially true of mathematicians. Arithmetic computation is one of the roots of mathematics. Since Euclid's algorithm for finding greatest common divisors, many seminal mathematical contributions have consisted of new procedures. But powerful computer graphics have now enabled mathematicians to envisage the behaviour of these procedures and, thereby, gain new insights, make new conjectures and explore new avenues of research. Think of the explosive interest in fractals, for instance. This has been driven primarily by our new-found ability rapidly to visualise fractal shapes, such as the Mandelbrot set. Taking advantage of these new opportunities has required the learning of new skills, such as using computer algebra and graphics packages.

The argument is even stronger. It is not just that computational skills are a useful adjunct to a mathematician's arsenal, but that they are becoming essential. Mathematical knowledge is growing exponentially: following its own version of Moore's Law. Without computer-based information retrieval techniques it will be impossible to locate relevant theories and theorems, leading to a fragmentation and slowing down of the field as each research area rediscovers knowledge that is already well-known in other areas. Moreover, without the use of computers, there are potentially interesting theorems that will remain unproved. It is an immediate corollary of Gödel's Incompleteness Theorem that, however huge a proof you think of, there is a short theorem whose smallest proof *is* that huge. Without a computer to automate the discovery of the bulk of these huge proofs, we have no hope of proving these simple-stated theorems. We have already seen early examples of this phenomenon in the Four-Colour Theorem and Kepler's Conjecture on sphere

packing. Perhaps computers can also help us to navigate, abstract and, hence, understand these huge proofs.

Realising this dream of computer access to a world repository of mathematical knowledge, visualising and understanding this knowledge, and reusing and combining it to discover new knowledge, presents a major challenge to mathematicians and informaticians. The first part of this challenge arises because mathematical knowledge will be distributed across multiple sources and represented in diverse ways. We need a lingua franca that will enable this babel of mathematical languages to communicate with each other. This is why this book — proposing just such a lingua franca — is so important. It lays the foundations for realising the rest of the dream.

OMDOC is an open markup language for mathematical documents. The ‘markup’ aspect of OMDOC means that we can take existing knowledge and annotate it with the information required to retrieve and combine it automatically. The ‘open’ aspect of OMDOC means that it is extensible, so future-proofed against new developments in mathematics, which is essential in such a rapidly growing and complex field of knowledge. These are both essential features. Mathematical knowledge is growing too fast and is too distributed for any centrally controlled solution to its management. Control must be distributed to the mathematical communities that produce it. We must provide lightweight mechanisms under local control that will enable those communities to put the produce of their labours into the commonwealth with minimal effort. Standards are required to enable interaction between these diverse knowledge sources, but they must be flexible and simple to use. These requirements have informed OMDOC’s development. This book will explain to the international mathematics community what they need to do to contribute to and to exploit this growing body of distributed mathematical knowledge. It will become essential reading for all working mathematicians and mathematics students aspiring to take part in this new world of shared mathematical knowledge.

OMDOC is one of the first fruits of the Mathematical Knowledge Management (MKM) Network (<http://www.mkm-ig.org/>). This network combines researchers in mathematics, informatics and library science. It is attempting to realise the dream of creating a universal digital mathematics library of all mathematical knowledge accessible to all via the World-Wide-Web. Of course, this is one of those dreams that is never fully realised, but remains as a source of inspiration. Nevertheless, even its partial realisation would transform the way that mathematics is practised and learned. It would be a dynamic library, providing not just text, but allowing users to run computer software that would provide visualisations, calculate solutions, reveal counter-examples and prove theorems. It would not just be a passive source of knowledge but a partner in mathematical discovery. One major application of this library will be to teaching. Many of the participants in the MKM Network are building teaching aids that exploit the initial versions of the library.

There will be a seamless transition between teaching aids and research assistants — as the library adjusts its contribution to match the mathematical user's current needs. The library will be freely available to all: all nations, all age groups and all ability levels.

I'm delighted to write this foreword to one of the first steps in realising this vision.

May 2006

Alan Bundy

Preface

Mathematics is one of the oldest areas of human knowledge¹. It forms the basis of most modern sciences, technology and engineering disciplines. Mathematics provides them with modeling tools such as statistical analysis or differential equations. Inventions like public-key cryptography show that no part of mathematics is fundamentally inapplicable. Last, but not least, we teach mathematics to our students to develop abstract thinking and hone their reasoning skills.

However, mathematical knowledge is far too vast to be understood by one person, moreover, it has been estimated that the total amount of published mathematics doubles every ten to fifteen years [Odl95]. Thus the question of supporting the management and dissemination of mathematical knowledge is becoming ever more pressing but remains difficult. Even though mathematical knowledge can vary greatly in its presentation, level of formality and rigor, there is a level of deep semantic structure that is common to all forms of mathematics and that must be represented to capture the essence of the knowledge.

At the same time it is plausible to expect that the way we do (i.e., conceive, develop, communicate about, and publish) mathematics will change considerably in the years to come. The Internet plays an ever-increasing role in our everyday life, and most of the mathematical activities will be supported by mathematical software systems connected by a commonly accepted distribution architecture, which makes the combined systems appear to the user as one homogeneous application. They will communicate with human users and amongst themselves by exchanging structured mathematical documents, whose document format makes the context of the communication and the meaning of the mathematical objects unambiguous.

Thus the inter-operation of mathematical services can be seen as a knowledge management task between software systems. On the other hand, math-

¹ We find mathematical knowledge written down on Sumerian clay tablets, and even Euclid's *Elements*, an early rigorous development of a larger body of mathematics, is over 2000 years old.

ematical knowledge management will almost certainly be web-based, distributed, modular, and integrated into the emerging math services architecture. So the two fields constrain and cross-fertilize each other at the same time. A shared fundamental task that has to be solved for the vision of a “web of mathematical knowledge” (MATHWEB) to become reality is to define an open markup language for the mathematical objects and knowledge exchanged between mathematical services. The OMDOC format (Open Mathematical Documents) presented here is an answer to this challenge, it attempts to provide an infrastructure for the communication and storage of mathematical knowledge.

Mathematics – with its long tradition in the pursuit of conceptual clarity and representational rigor – is an interesting test case for general knowledge management, since it abstracts from vagueness of other knowledge without limiting its inherent complexity. The concentration on mathematics in OMDOC and this book does not preclude applications in other areas. On the contrary, all the material directly extends to the STEM (science, technology, education, and mathematics) fields, once a certain level of conceptualization has been reached.

This book tries to be a one-stop information source about the OMDOC format, its applications, and best practices. It is intended for authors of mathematical documents and for application developers. The book is divided into four parts: an introduction to markup for mathematics (Part I), an OMDOC primer with paradigmatic examples for many kinds of mathematical documents (Part II), the rigorous specification of the OMDOC document format (Part III), and an XML document type definition and schema (Part IV).

The book can be read in multiple ways:

- for users that only need a casual exposure to the format, or authors that have a specific text category in mind, it may be best to look at the examples in the OMDOC primer (Part II of this book),
- for an in-depth account of the format and all the possibilities of modeling mathematical documents, the rigorous specification in Part III is indispensable. This is particularly true for application developers, who will also want to study the external resources, existing OMDOC applications and projects, in Part IV.
- Application developers will also need to familiarize themselves with the OMDOC Schema in the Appendix.

Acknowledgments

Of course the OMDOC format has not been developed by one person alone. The original proposal was taken up by several research groups, most notably the Ω MEGA group at Saarland University, the MAYA and ACTIVEMATH projects at the German Research Center of Artificial Intelligence (DFKI), the MoWGLI EU Project, the RIACA group at the Technical University of Eindhoven, and the COURSECAPSULES project at Carnegie Mellon University. They discussed the initial proposals, represented their materials in OMDOC and in the process refined the format with numerous suggestions and discussions.

The author specifically would like to thank Serge Autexier, Bernd Krieg-Brückner, Olga Caprotti, David Carlisle, Claudio Sacerdoti Coen, Arjeh Cohen, Armin Fiedler, Andreas Franke, George Goguadze, Alberto González Palomo, Dieter Hutter, Andrea Kohlhase, Christoph Lange, Paul Libbrecht, Erica Melis, Till Mossakowski, Normen Müller, Immanuel Normann, Martijn Oostdijk, Martin Pollet, Julian Richardson, Manfred Riem, and Michel Vollebregt for their input, discussions, and feedback from implementations and applications.

Special thanks are due to Alan Bundy and Jörg Siekmann. The first triggered the work on OMDOC, has lent valuable insight over the years, and has graciously consented to write the foreword to this book. Jörg continually supported the OMDOC idea with his abundant and unwavering enthusiasm. In fact the very aim of the OMDOC format: openness, cooperation, and philosophic adequateness came from the spirit in his Ω MEGA group, which the author has had the privilege to belong to for more than 10 years.

The work presented in this book was supported by the “Deutsche Forschungsgemeinschaft” in the special research action “Resource-adaptive cognitive processes” (SFB 378), and a three-year Heisenberg Stipend to the author. Carnegie Mellon University, SRI International, and the International University Bremen have supported the author while working on revisions for versions 1.1 and 1.2.

Table of Contents

Part I Setting the Stage for Open Mathematical Documents

1	Document Markup for the Web	3
1.1	Structure vs. Appearance in Markup	3
1.2	Markup for the World Wide Web	5
1.3	XML, the eXtensible Markup Language.....	6
2	Markup for Mathematical Knowledge	13
2.1	Mathematical Objects and Formulae	14
2.2	Mathematical Texts and Statements.....	20
2.3	Large-Scale Structure and Context in Mathematics.....	21
3	Open Mathematical Documents	25
3.1	A Brief History of the OMDoc Format	25
3.2	Three Levels of Markup.....	28
3.3	Situating the OMDoc Format	29
3.4	The Future: An Active Web of (Mathematical) Knowledge .	31

Part II An OMDoc Primer

4	Textbooks and Articles	35
4.1	Minimal OMDoc Markup	36
4.2	Structure and Statements	39
4.3	Marking up the Formulae	41
4.4	Full Formalization	45
5	OpenMath Content Dictionaries	49
6	Structured and Parametrized Theories	55
7	A Development Graph for Elementary Algebra	59

8	Courseware and the Narrative/Content Distinction	65
8.1	A Knowledge-Centered View	67
8.2	A Narrative-Structured View	71
8.3	Choreographing Narrative and Content OMDoc	73
8.4	Summary	74

9	Communication Between Systems	75
----------	--	----

Part III The OMDoc Document Format

10	OMDoc as a Modular Format	83
10.1	The OMDoc Namespaces	83
10.2	Common Attributes in OMDoc	85
11	Document Infrastructure	89
11.1	The Document Root	90
11.2	Metadata	91
11.3	Document Comments	92
11.4	Document Structure	93
11.5	Sharing Document Parts	94
12	Metadata	97
12.1	The Dublin Core Elements (Module DC)	98
12.2	Roles in Dublin Core Elements	101
12.3	Managing Rights	102
12.4	Inheritance of Metadata	104
13	Mathematical Objects	107
13.1	OpenMath	107
13.2	Content MathML	114
13.3	Representing Types in Content-MATHML and OPENMATH	116
13.4	Semantics of Variables	119
13.5	Legacy Representation for Migration	120
14	Mathematical Text	121
14.1	Multilingual Mathematical Vernacular	121
14.2	Formal Mathematical Properties	123
14.3	Text Fragments and Their Rhetoric/Mathematical Roles	124
14.4	Phrase-Level Markup of Mathematical Vernacular	126
14.5	Technical Terms	127
14.6	Rich Text Structure	128

15	Mathematical Statements	133
15.1	Types of Statements in Mathematics	133
15.2	Theory-Constitutive Statements in OMDoc	136
15.3	The Unassuming Rest	141
15.4	Mathematical Examples in OMDoc	146
15.5	Inline Statements	148
15.6	Theories as Structured Contexts	149
16	Abstract Data Types	155
17	Representing Proofs	159
17.1	Proof Structure	160
17.2	Proof Step Justifications	163
17.3	Scoping and Context in a Proof	167
17.4	Formal Proofs as Mathematical Objects	169
18	Complex Theories	173
18.1	Inheritance via Translations	174
18.2	Postulated Theory Inclusions	177
18.3	Local/Required Theory Inclusions	179
18.4	Induced Assertions	180
18.5	Development Graphs	182
19	Notation and Presentation	187
19.1	Styling OMDoc Elements	188
19.2	A Restricted Style Language	190
19.3	Notation of Symbols	191
19.4	Presenting Bound Variables	198
20	Auxiliary Elements	201
20.1	Non-XML Data and Program Code in OMDoc	202
20.2	Applets and External Objects in OMDoc	205
21	Exercises	209
22	Document Models for OMDoc	213
22.1	XML Document Models	213
22.2	The OMDoc Document Model	215
22.3	OMDoc Sub-Languages	217

Part IV OMDoc Applications, Tools, and Projects

23	OMDoc Resources	223
23.1	The OMDoc Web Site, Wiki, and Mailing List	223
23.2	The OMDoc Distribution	223
23.3	The OMDoc Bug Tracker	224
23.4	An XML Catalog for OMDoc	225
23.5	External Resources	225
24	Validating OMDoc Documents	227
24.1	Validation with Document Type Definitions	228
24.2	Validation with RelaxNG Schemata	232
24.3	Validation with XML Schema	232
25	Transforming OMDoc	235
25.1	Extracting and Linking XSLT Templates	235
25.2	Interfaces for Systems	237
25.3	Presenting OMDoc to Humans	240
26	Applications and Projects	241
26.1	Introduction	241
26.2	QMath Parser	244
26.3	Sentido Integrated Environment	247
26.4	MBase	252
26.5	A Search Engine for Mathematical Formulae	254
26.6	Semantic Interrelation and Change Management	258
26.7	MathDox	262
26.8	ActiveMath	266
26.9	Authoring Tools for ACTIVEMATH	272
26.10	SWIM – An OMDoc-Based Semantic Wiki	274
26.11	Induction Challenge Problems	278
26.12	MAYA	281
26.13	HETS	286
26.14	CPoint	290
26.15	$\text{\LaTeX}$: A $\text{\LaTeX}$ -Based Workflow for OMDoc	294
26.16	An Emacs Mode for Editing OMDoc Documents	298
26.17	Converting Mathematica Notebooks to OMDoc	301
26.18	Standardizing Context in System Interoperability	304
26.19	Proof Assistants in Scientific Editors	309
26.20	VeriFun	313

Part V Appendix

A	Changes to the Specification	319
A.1	Changes from 1.1 to 1.2.....	320
A.2	Changes from 1.0 to 1.1.....	327
B	Quick-Reference	333
C	Table of Attributes	339
D	The RelaxNG Schema for OMDoc	345
D.1	The Sub-language Drivers	345
D.2	Common Attributes	347
D.3	Module MOBJ: Mathematical Objects and Text	347
D.4	Module MTXT: Mathematical Text	348
D.5	Module DOC: Document Infrastructure	349
D.6	Module DC: Dublin Core Metadata	350
D.7	Module ST: Mathematical Statements	351
D.8	Module ADT: Abstract Data Types	353
D.9	Module PF: Proofs and Proof objects.....	354
D.10	Module CTH: Complex Theories	355
D.11	Module RT: Rich Text Structure	356
D.12	Module EXT: Applets and Non-XML Data.....	356
D.13	Module PRES: Adding Presentation Information.....	357
D.14	Module QUIZ: Infrastructure for Assessments	359
E	The RelaxNG Schemata for Mathematical Objects	361
E.1	The RelaxNG Schema for OpenMath	361
E.2	The RelaxNG Schema for MathML.....	363
	Bibliography	375
	Index	389