Enriching Metadata for XML Journal Articles Through Extraction of MathML and Function Names

Timothy W. Cole (t-cole3@uiuc.edu), William H. Mischo, Thomas G. Habing, & Ying-ping Chen
University of Illinois at Urbana-Champaign

Enhancing the Searching of Mathematics
IMA Hot Topics Workshop, Univ. of Minnesota, April 2004
Presentation Overview

Librarian bias – using math content to search & link literature

- Context & objectives of current research
- Legacy journal article testbed used
- Metadata schemas used
- Enriching article metadata
  - Finding occurrences of Functions metadata in article literature
  - Extracting, selecting, normalizing, & including MathML
- Future work
Objectives of Current Project

Part of grant to integrate Wolfram Functions & MathWorld sites into NSDL (see also Michael Trott’s presentation)

- Generate Dublin Core metadata for these resources
- Export metadata using OAI

- Investigate ways to enrich journal article metadata to facilitate search & discovery & linking to reference literature (e.g., Functions & MathWorld sites)
  - Main focus of rest of this presentation
Square root

Sqrt

Mathematica Notation: Sqrt[z]
Traditional Notation: $\sqrt{z}$

http://functions.wolfram.com/01.01.02.0001.01

MathML Form

<math xmlns='http://www.w3.org/1998/Math/MathML' mathematicaForm='TraditionalForm' xmlns:mathematica='http://www.wolfram.com/XML/'><semantics><mrow><msqrt><mi>z</mi></msqrt></mrow></semantics></math>

Date Added to functions.wolfram.com (modification date)

2001-10-29

© 2002 Wolfram Research, Inc.
Sample of metadata exported in OAI

```xml
<oai_dc:dc ...
    <dc:title>Square root: Primary definition (formula ...</dc:title>
    <dc:subject>Sqrt</dc:subject>
    <dc:subject>square root</dc:subject>

    <dc:description>Primary definition (2 formulas)</dc:description>
    <dc:description>&lt;math ... &lt;/math&gt;</dc:description>
    <dc:date>2001-10-29</dc:date>
    <dc:publisher>Wolfram Research, Inc.</dc:publisher>
    <dc:type>Text</dc:type>
    <dc:format>text/html; charset=iso-8859-1</dc:format>
    <dc:identifier>http://functions.wolfram.com/.../Sqrt/02/0001/</dc:identifier>
    <dc:identifier>http://functions.wolfram/.../01.01.02.0001.01</dc:identifier>
    <dc:language>en</dc:language>
    <dc:rights>&#169; 2002 Wolfram Research, Inc.</dc:rights>
</oai_dc:dc>
```
Enriching Metadata for Journal Literature

- **Hypotheses:**
  - Readers of a journal article may want to know more about mathematical functions used in article … And vice-versa
  - Functions site keywords/phrases & mathematics in body of articles may be useful for linking, search & discovery

- **Approach:**
  - Examine journal article testbed for presence of Functions site keywords; add links & terminology to article metadata
  - Extract selected MathML occurrences from article body; “normalize”; add to article metadata
Article Testbed: Illinois DLI-I / D-Lib Test Suite


- Featured:
  - Multi-publisher, Markup-Based Full-Text Journal Testbed.

- Testbed contains 75,000+ Articles from 60 Journal Titles
  - Received as SGML (various DTDs); converted to XML
  - Content from AIP, APS, ASCE, IEE, ASM, ACM, Elsevier
  - Additional support from IEEE, NRL, NTT Learning Systems
Converting Legacy Markup to MathML

- **Goal:** Convert publisher-specific SGML/XML math markup to presentation MathML
- **Groundrules:**
  - Minimize need for human intervention
  - Utilize standards-based techniques (e.g., XSLT, DOM)
  - Embed MathML in full XML document
  - **Validate success based on quality of presentation**
  - Strive for consistency across MathML viewers
- **Outcome:**
  - 2.5 million instances of MathML
  - Limited quality & consistency (discussed more later)
Mathematics Markup Transformations

- Identify & translate mathematical character references
- Identify & tokenize mathematical content
- Recognize & transform mathematical markup (e.g., embellishments, script & limit schemtas, etc.)

ISO 12083 Math

\[ a_i^2 \]

Presentational MathML

\[
<math xmlns="http://www.w3.org/...">
  <msubsup>
    <mrow><mi>&alpha;</mi></mrow>
    <mrow><mi>i</mi></mrow>
    <mrow><mn>2</mn></mrow>
  </msubsup>
</math>

Original Metadata Format

- Qualified Dublin Core stored in separate RDF/XML files
- Most fields extracted from article XML via XSLT
  - Titles, author names & affiliations, subject terms, journal issue information, abstract, bibliography/endnotes
  - Article table of contents derived from section headings
  - MathML embedded in title, abstract, etc. was preserved
  - Added local semantics for terminology encodings, etc.
- Links to earlier & later articles, A & I database entries, etc. maintained in relational DB
  - Added to metadata records via scripts using XML DOM
Sample Legacy Metadata in RDF/XML

```xml
<rdf:RDF namespaces declarations omitted >
  <rdf:Description
    rdf:about="http://…/~acm/TOMS/25_3/LECUYER/05_LECUYER_FULL.XML">
    <dc:title rdf:parseType="Literal">
      Beware of Linear Congruential Generators with Multipliers of the Form
    </dc:title>
    <math display="inline" altimg="math0001.png"
      xmlns="http://www.w3.org/1998/Math/MathML">
      \( a = \pm 2^q \pm 2^r \)
    </math>
    <dc:creator>
      <rdf:Seq>
      ...
    </dc:creator>
  </rdf:Description>
</rdf:RDF>
```
Changes to Metadata Strategy for Current Work

- Metadata schema now qualified DC, dropped RDF
  - Local XML schema imports DCMI schemas, adds local semantics & content models that allow child elements, embedded MathML

- Add metadata terms from & links to Wolfram Functions
  - Functions keywords found in full-text added as dc:subject
  - Links to Functions pages added as dc:relationship

- Add selected MathML from article body to metadata
  - Added as dc:subject
Searching for Occurrences of Functions Site Metadata Terms in Article Testbed

- 83,000 pages containing keywords in <meta> elements
  - Approximately 7,500 unique keywords

- Not all keywords useful for full-text search
  - Eliminated numerals, Mathematica expressions, single word phrases, phrases containing selected words
  - About 1,000 keywords left after automated filtering

- 367 of these keywords appear in journal article testbed
  - 44,000 articles contain at least one keyword
  - On average each of those 44,000 contain 2 keywords
Issues in Using Functions Keywords/phrases for Characterizing Articles

- Vocabulary switching
  - Keywords in mathematics, not necessarily terms used in physics, electrical engineering, & computer science
- Meaning of words in full-text less precise
  - Synonyms, less specific use, etc.
- Context – same descriptive keywords may map to many different branches of Functions site
  - E.g., 300 occurrences of “q-series” in Functions site; which are relevant to 18 articles containing this keyword?

show examples
Adding MathML Found in Article Body

- 2.5 million occurrences of MathML in testbed, but most not useful for searching & linking
  - MathML used for greek characters in text
  - Many instances short, inline fragments
  - Quality – many instances can’t be parsed by Mathematica

- Criteria
  - Explicitly equations of block display format
  - Minimum length, 200 bytes
  - Accepted by Mathematica kernel

- About 3% of occurrences meet criteria
Issues Extracting MathML

- Quality & Consistency issues
  - Automated transformations from SGML Math to MathML only partially successful
  - Math in testbed articles includes lots of formatting inside Math elements (equation numbers, thinspace, …)

- Normalization issues
  - Use of Mathematica kernel as normalization process
  - Ambiguities of trying to normalize presentation MathML

show examples
<uiLib:qualifiedDC>
<identifier xsi:type="dct:URI">http://…/25_3/LECUYER/05_LECUYER_FULL.XML</identifier>
<title>Beware of Linear Congruential Generators with Multipliers of the Form a = ± 2 q …</title>
<dct:alternative>Beware of Linear Congruential Generators</dct:alternative>
<creator>L'Ecuyer, Pierre Université de Montréal Département d'Informatique …</creator>
<subject xsi:type="uiLib:ACMCCS_CAT">G.4 Mathematics of Computing:Mathematic…</subject>
<subject xsi:type="uiLib:ACMCCS_TERM">Performance</subject>
<subject xsi:type="uiLib:ACMCCS_KEYWORD">linear congruential generators</subject>
<subject xsi:type="uiLib:WolframFunctions">q-series</subject>
<dct:abstract>Linear congruential random-number generators with Mersenne …</dct:abstract>
<publisher>Association for Computing Machinery</publisher>
<dct:issued xsi:type="dct:W3CDTF">1999-09</dct:issued>
<dc:relation xsi:type="uiLib:WebPage">
</dc:relation>
</uiLib:qualifiedDC>
Related & Future Work

Related work:
- Metasearch over multiple A & I services simultaneously
- Grainger Library OAI sci-tech metadata aggregation

Future work:
- Other approaches to normalizing & searching MathML extracted from journal articles
- Using other vocabularies (e.g., MCS) to link between bibliographic resources and Wolfram Functions site
- Evaluating usefulness of these approaches for end-users