Math Searching and MathML in the NSDL

Robert Miner
Director of New Product Development
Design Science, Inc.

Top 10 Reasons to Care about Math on the Web

1. Access!
2. Access!
3. Access!
4. Access!
5. Access!
6. Access!
7. Access!
8. Access!
9. Sex appeal
10. Other

Qu'est-ce que c'est Access?

Access means different things to different people:

- Eliminating barriers of time and space
- Accessibility for the visually impaired
- Improving discovery of relevant information and communities

Accessibility of Math on the Web, Part I

Though math content lags behind many other disciplines due to the complexity of authoring and viewing math for the web, a great deal of material is nonetheless accessible in the sense of *available*:

- research abstracts -- good coverage, especially of recent publications
- research articles -- good, varies by publisher and age
- preprints -- depends on subject and community
- curriculum & enrichment materials -- still spotty

Accessibility of Math on the Web, Part II

Unfortunately, much of the material that is available on the web is *not* accessible in the sense of *easily discoverable*:

- Virtually all searches are text-based, thus requiring one to know good search terms in advance. Unless you know [*X*,*Y*] is called a Lie bracket, finding that out from the 13M hits for "[X,Y]" on Google is hard.
- Because so much of the math that is online is locked up in PDFs, images, etc, text-based searches often fail, or cover only the abstract material.
- What little metadata is available is mostly at the document level and restricted to cataloging info.

Envisioning Better Math Searching, Part I

Rob Corliss tells of finding the solution to a problem arising from a dynamical system in a Hungarian combinatorics journal. The problem involved a series, which he entered in to the The On-Line Encyclopedia of Integer Sequences

This is a special case of a math-aware keyword search: *Enter a math expression and see where it appears in the literature.*

Challenges: differing notations, detecting mathematical equivalence, etc.

Envisioning Better Math Searching, Part II

A parent wants to help a child do a project on fractals. There are 518,000 Google hits on "fractals", and while a few of them were appropriate, most were just pictures, or too advanced, and it was heavy going with a 5th grader.

By switching over to the Eisenhower National Clearinghouse, and searching for reviewed internet lessons for grades 3-6, they located 10 excellent resources in just a few minutes.

This is an example of a math-aware metadata search: *Enter a search term along with criteria, and see what documents meet them.*

Challenges: creating the metadata and controlled vocabularies

How Does MathML Fit In?

MathML is an XML encoding for mathematical notation and basic mathematical semantics. It was originally developed ('96-'98) for three main reasons:

- To address problems with displaying and authoring math for web pages.
  [Partly successful]
- To fit in better with an XML/Web-centric technologies
  [Quite successful, and picking up speed]
- To support greater functionality and reuse
  [Just coming into its own]

MathML and Searching: An Opportunity

The confluence of three factors has created a rare opportunity for improving access to math and science literature:

- There is a broad trend toward XML-based workflows on the part of content providers and toolmakers. With the shift, best practices are in flux.
- MathML is highly-structured, and information-rich, supporting greater functionality and reuse.
- For a change, there is a credible market incentive for improving math support.

How Does NSDL Fit In?

The National Science Digital Library is an NSF-funded "digital library of exemplary resource collections and services, organized in support of science education at all levels." The NSDL strives to be a center of innovation for digital library infrastructure and practice.

In recognition of the window of opportunity to improve math searching, NSDL awarded Design Science a grant with three objectives:

1. To identify a framework for activity leading to the adoption and deployment of improved mathematical searching.
2. To develop a testbed document collection for useability testing
3. To identify promising ways of better searching mathematical content.

The Framework Activity

Many people at many organizations are working on aspects of enhancing math searching. A framework is required so that these disparate efforts ultimately fit together to achieve a lasting, industry-wide improvement.

To this end, the grant funds two workshops to bring together researchers, tool makers and content providers to

- identify areas requiring cooperation/standardization
- establish best practices for workflows being put in place today

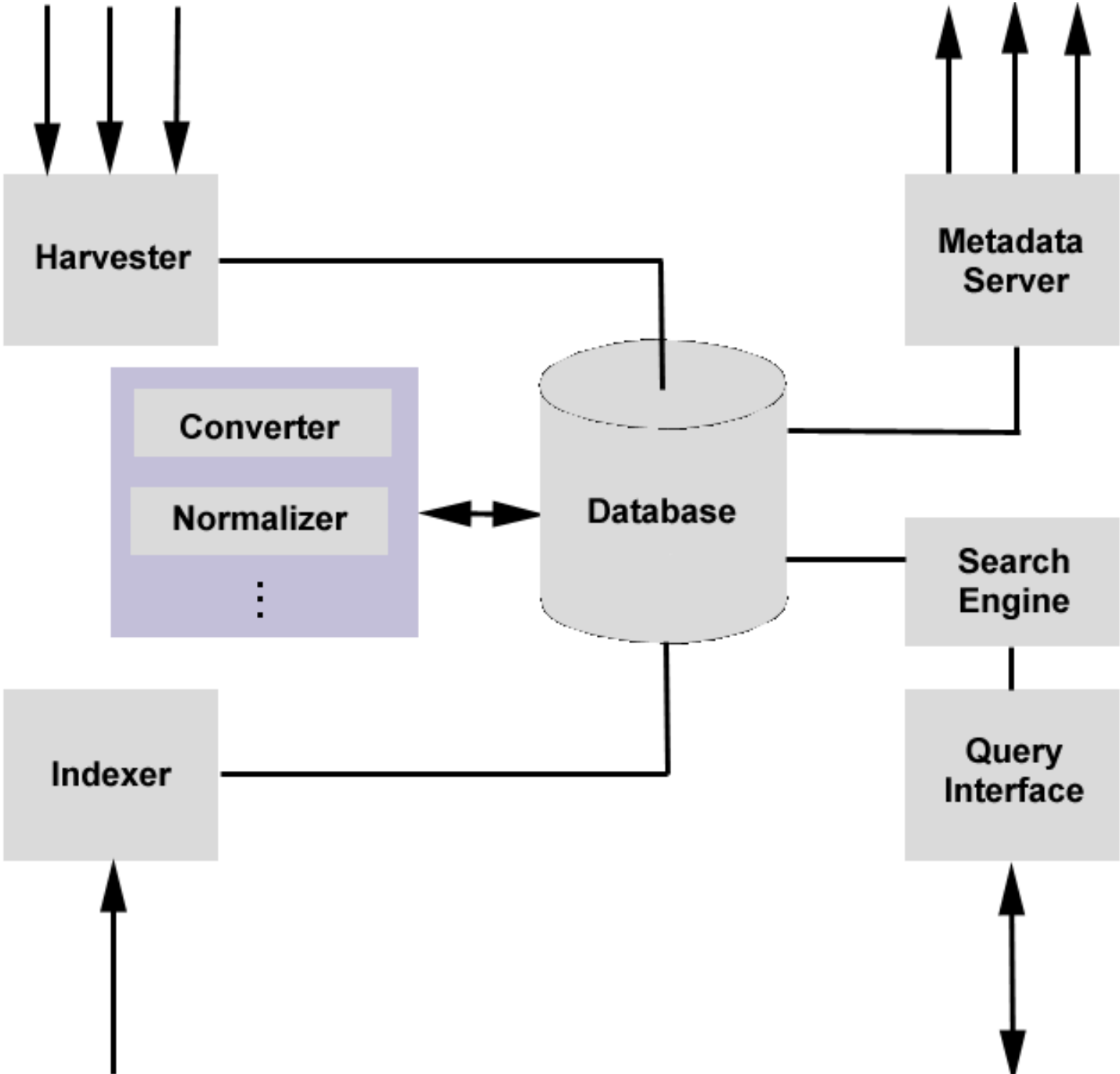
The first workshop will be April 26-27 at IMA.

Architectural Framework Requirements

Although the framework is yet to be devised, some requirements for an equation search architecture are relatively clear:

- Some common notion of an "equation record":
- A means of sharing equations records between distributed collections, including digital rights management capabilities
- Neutrality with respect to data formats, controlled vocabularies, metadata schemes, etc.
- Means by which metadata can be accumulated by various parties over time

A Hypothetical Architecture



The Testbed Activity

The testbed activity seeks to

- develop and/or integrate open source software to implement the framework architecture
- create a collection of documents on which to conduct useability testing.

The software framework will provide no algorithms, but will establish APIs via which can be implemented by a variety of algorithm modules.

The Algorithm Activity

Algorithms are required at several points in the framework architecture:

At Authoring Time

Smart authoring tools could improve metadata in new documents using equation libraries, heuristics, profiles, etc.

At Indexing Time

Detection of math in PostScript/PDF, math-aware OCR, TeX translation, hashing signatures, automatic generation of metadata.

At Query Time

Math-aware algorithms might employ canonicalization, knowledge of formal properties (e.g. commutivity), invariant hashing signatures, etc.

Proposed Algorithm Research

Three families of algorithms will be investigated as part of the NSDL grant:

Pattern Matching with Normalization
- Strategy: normalize MathML, canonicalization of variable names, standard forms for classes such as polynomials, etc.

Formal Evaluation
- Strategy: convert to content MathML where possible, formally evaluate expressions at test points, use pullback metrics

Per-equation Metadata Methods
- Strategy: automatic indexing using context and heuristics, various use of taxonomic methods in conjuction with other techniques

Getting Involved

Math searching promises to be an exciting area. It enjoys not only an abundance of meaty problems at the intersection of math and computer science, but also interest from the public and private sector.

Ways to get involved include:

- Keeping abreast of news and opportunities by visiting http://www.dessci.com/searching.
- Attending one of the workshops
- Setting up an equation metadata server
- Building a searchable document collection