# Towards a Global Digital Mathematics Library

Jim Pitman

Mathematica and Statistics, University of California, Berkeley

CICM 2015, Washington DC, July 13 2015

# NRC Report

**Developing a 21st Century Global Library for Mathematics Research**

Committee report for the National Reseach Council of the US National Academies.

The National Academies Press
`http://www.nap.edu/catalog.php?record_id=18619`

`http://arxiv.org/abs/1404.1905`

Advertisement and summary (with Clifford Lynch)
`http://www.ams.org/notices/201407/rnoti-p776.pdf`

# IMU GDML Working Group

Members:

- Patrick Ion (chair)
- Thierry Bouche
- Bruno Buchberger
- Michael Kohlhase
- Jim Pitman
- Olav Teuchke
- Stephen Watt
- Eric Weisstein

Current focus: two organizational tasks:

- Create a legal foundation to set up
  IMKT = International Mathematical Knowledge Trust
- Workshop on requirements for a Semantic Capture Language

# History

- QED Manifesto (Robert Boyer 1993). RIP 1996.
  - Wiedijk (2007)
  - Gulf between formal/actual math
  - UI limitations for Mizar/HOL/Cog
- DML Digitization efforts (EuDML, Project Euclid, ... )
- Wolfram Alpha, Computational Continued Fractions
- OEIS (Encyclopedia of Integer Sequences)
- Wikipedia/Mathematics, Springer Encyclopedia of Math.
- Mathematic/Sage
- Vastly improved search and indexing.
- Linked Open Data

# NRC Report: Charge

- evaluation of the potential value of a GDML
- appropriate scope
- issues and alliances involved in establishing such a library
- range of desired capabilities for such a system
- which capabilities within reach soon
- resource needs and a way forward

Non-charge

- copyright and open access

Pragmatic view of operating in the current landscape

# NRC Report: Findings

- Construction of mathematical libraries through centralized aggregation of resources has reached a point of diminishing returns. e.g. retrospective digitization efforts, EuDML, Project Euclid, ...
- Better create a comprehensive digital mathematics information resource of greater value than the sum of its contributing parts.
- Fully automated recognition of mathematical concepts and ideas (e.g., theorems, proofs, sequences, groups) is not yet possible, though something to aspire to;
- Significant benefit coult be realized by using existing scalable methods and algorithms to assist human agents in identifying important mathematical concepts in the research literature

# NRC Report: Way forward

Establish an organization (IMKT)
IMKT = International Mathematical Knowledge Trust
to support new functionalities and services over mathematical concepts
(theorems / proofs / formulas / identities /... )

- listing
- searching
- browsing
- navigating
- annotating
- linking

GDML = Union of such efforts supported by IMKT

# Roles for IMKT

- oversee GDML development
- develop platforms, tools, and services for curation, annotation and navigation of mathematical information
- encourage agents to make mathematical information available through APIs as Linked Open Data
- mobilize and coordinate the mathematical community to engage with these capabilities
- support an ongoing applied research program in mathematical information management

# Approaches to semantic capture

Provide identifiers for mathematical concepts to facilitate linking and navigation, by some combination of

- computational machine learning
- textual analysis methods
- community-based editorial efforts

Need for a community supported *Semantic Capture Language*.

Significant first steps by Wolfram Reseach project

*Computational Knowledge of Continued Fractions*

supported by Sloan Foundation, now part of WolframAlpha

# NRC Report: Recommendations (I)

- A primary role of the IMKT should be to provide a platform that engages the mathematical community in enriching the GDML knowledge base and identifies connections in the data.
- The GDML should rely on citation indexing, community sourcing, and a combination of other computationally based methods for linking among articles, concepts, authors, and so on.
- Community engagement and the success of community-sourced efforts need to be continuously evaluated to ensure that IMKT missions continue to align with community needs and that community engagement efforts are effective.

# NRC Report: Recommendations (II)

- The GDML should be open and built to cooperate with both researchers and existing services. In particular, the content (knowledge structures) of the library, at least for vocabularies, tags, and links, should also be open, although the library will link to both open and copyright-restricted literature.
- The IMKT should serve as a nexus for the coordination of research and research outcomes, including
  - management and encouragement of the creation of a knowledge-based library of mathematical concepts such as theorems and proofs
  - community endorsements
  - development and funding of open information systems of use to mathematicians,
  - encouragement of best practices to facilitate knowledge management in research mathematics.

# NRC Report: Recommendations (III)

- The initial GDML planning group should set up a task force of suitable experts to produce a realistic plan, timeline, and prioritization of components, using this report as a high-level blueprint, to present to potential funding agencies (both public and private).
- The IMKT needs to build an ongoing relationship with the research communities spanning mathematics, computer science, information science, and related areas concerned with knowledge extraction and structuring in the context of mathematics and to help translation of developments in these areas from research to large-scale application.

# Broader Issues

Ownership and control of mathematical information:

- licensing: can you machine-prcess the data? ...
- version control / provenance /archiving
- data quality control (at least syntactically correct LaTeX/JSON/XML/RDF/Wolfram/...)
- by what authority is a particular result known to be mathematically true or false?
- how is that authority represented in the data?
- reconciliation of data from different sources

# Existing systems of ownership and control

- journals
- books (especially handbooks: tables of integrals, ...)
- MR/ZbMATH
- Wikipedia
- OEIS = Online Encyclopedia of Integer Sequences
- WolframAlpha
- Mizar/Coq/HAL/...

Can we do better? I hope so. e.g.

- Standardize mathematical data formats (Semantic Capture Language)
- Open APIs to allow bulk checking of data to help identify errors
- Make it easy for users to propose corrections e.g.
- use Git for version control, ....