1st Leibniz MMS Days

How could the formation of the Global Digital Mathematics Library support mathematical modelling and simulation?

Olaf Teschke

FIZ Karlsruhe

Berlin, January 28th, 2016

zbMATH
the first resource for mathematics

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

What is a digital mathematics library?

Traditionally, a library has been a dynamic collection of publications, arranged in an accessible form.

"Mathematics" obviously refers to a restriction to math content.

Digital? (Scans vs. sophisticated XML, access options,…)

Content? Organization?

Naively (~2000): " Make whole math literature digitally available"

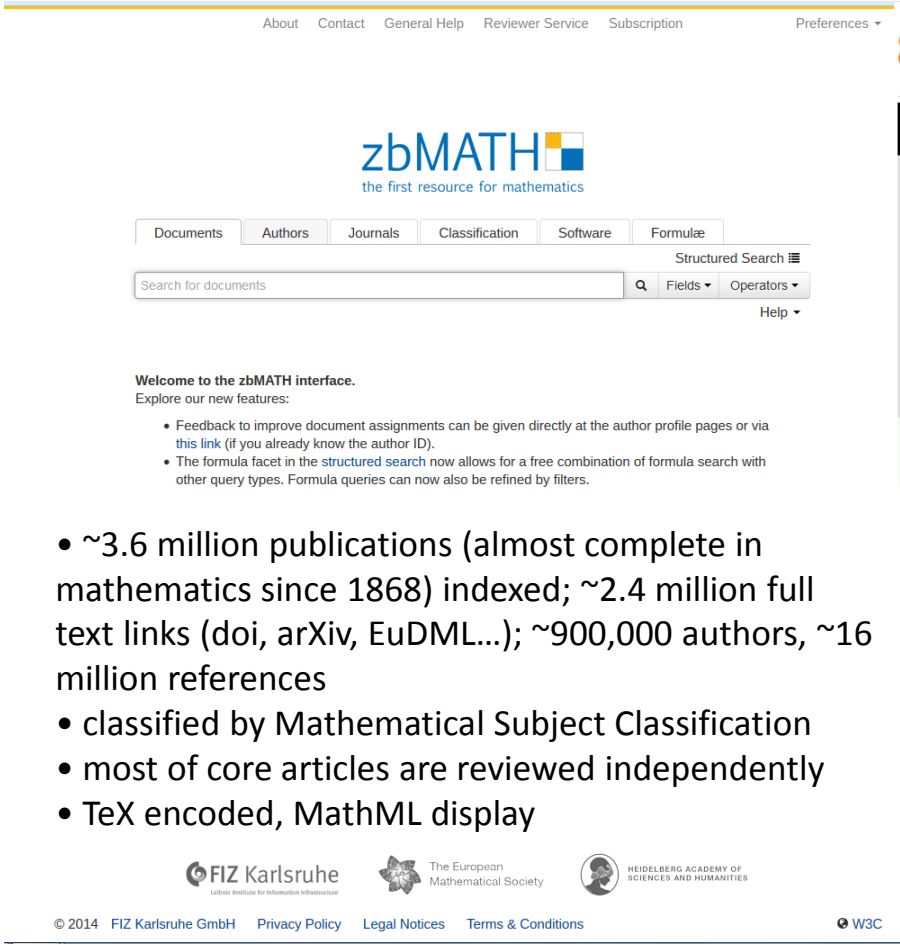
> Shaped into World Digital Mathematics Library (WDML): Digitization and accessibility projects, distributed and centralized frameworks

Components of digitization

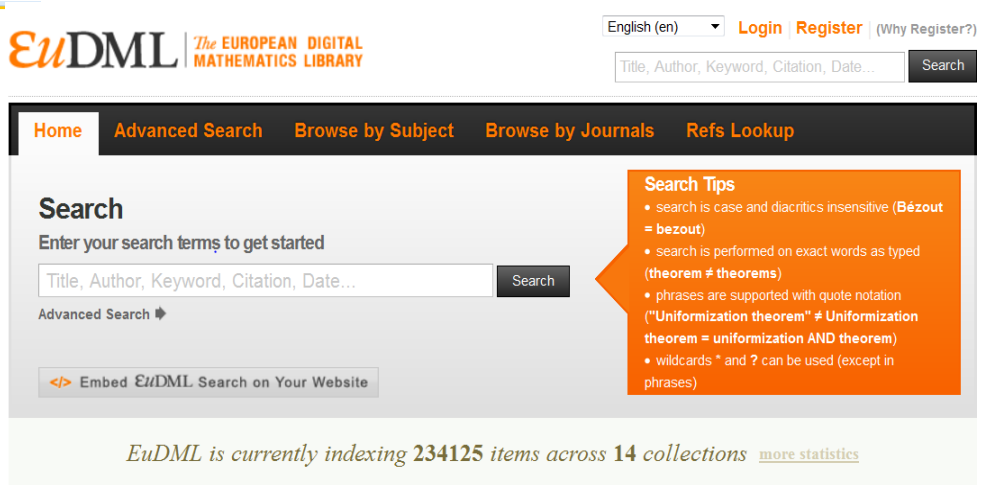
- Basic layers: HTML, (La)TeX, PDF, TCP/IP, XML,…
- (Retro-)Digitization efforts: Gallica, GDZ, JSTOR, Numdam, Publishers…
- E-born documents: arXiv, Electronic Publishing, ELibM, Euclid, EuDML…
- Superstructures: CrossRef, MathSciNet, MSC, zbMATH,…

By now, about 2.4 million math documents in zbMATH have links to electronic versions (out of 3.6 millions).

FIZ services in the classical DML framework



- ~3.6 million publications (almost complete in mathematics since 1868) indexed; ~2.4 million full text links (doi, arXiv, EuDML…); ~900,000 authors, ~16 million references
- classified by Mathematical Subject Classification
- most of core articles are reviewed independently
- TeX encoded, MathML display



- >240,000 math full-texts
- almost 300 journals available via (sometimes moving wall) OA
- Metadata TeX encoded, documents mostly pdf

zbMATH the first resource for mathematics

FIZ Karlsruhe Leibniz Institute for Information Infrastructure

What is a digital mathematics library? (II)

More recent digital libraries concept as collection of digital objects with a management -> more granular math entities may be contained, e.g.:

Let p be a prime factor with multiplicity n of the order of a finite group G, so that the order of G can be written as $p^n m$, where $n > 0$ and p does not divide m. Let p^n be the number of Sylow p-subgroups of G. Then the following hold:

(1) p^n divides m.
(2) $n_p \equiv 1 \bmod p$.
(3) $n_p = |G : N_G(P)|$, where P is any Sylow p-subgroup of G and N_G denotes the normalizer.

```
[/\ ∀ P, [max P | p.-subgroup(G) P] = p.-Sylow(G) P,
    [transitive G, on 'Syl_p(G) | 'JG],
    ∀ P, p.-Sylow(G) P → #|'Syl_p(G)| = #|G : 'N_G(P)|
& prime p → #|'Syl_p(G)| %% p = 1%N].
```

Coq formalization, by Freek Wiedijk

The shape of things to come?

NRC report “Developing a 21st Century Global Library for Mathematics Research”

(The National Academies Press, http://www.nap.edu/catalog.php?record_id=18619, http://arxiv.org/abs/1404.1905)

Some theses:

- Digitization in (classical) sense is already at a mature state (arXiv, EuDML, ...), diminishing returns from retrodigitization
- More value is expected from creating a quality digital mathematics information resource “of greater value than the sum of its contributing parts”
- Automated recognition of mathematical objects (e.g., theorems, proofs, sequences, groups) is not yet possible but should be seen as long-term goal
- Existing scalable methods and algorithms can be improved to assist intellectual analysis of mathematical concepts in the research literature

IMU GDML working group

2014 IMU GDML Panel (ICM Seoul; Bouche, Daubechies, Greuel, Ion, Zhang) related to the installment of a working group (Ion (chair), Bouche, Buchberger, Kohlhase, Pitman, Teschke, Watt, Weisstein) to identify and pursue realizable components of a GDML architecture

Mission statement: The GDML should

- enhance openness and accessibility of all mathematical knowledge world-wide
- serve research mathematics, education and the scientific and technological use of mathematics.
- be a resource for developing tools to promote use and development of mathematics.
- facilitate creation, dissemination and archiving of semantically annotated mathematical material.
- encourage the collaborative development of services based on semantic annotation.
- be a truly global resource, shaped as a distributed service which matches the highest possible standards of independence from national interests, of reliability, and of data protection

Disadvantages of classical publications for mathematical modelling and simulation

- Inadequate presentation of software
- Inadequate presentation of data
- Inadequate presentation of models
- Incomplete presentation of simulation results
- …

Necessary improvements in a GDML framework

Linked open data concepts should facilitate

- Software and data repositories in appropriate formats
- Sustainable links
- Semantic layers to enable appropriate organization and retrieval

Measures to gain and improve semantic information:

- Appropriate model annotation
- Link-generated enrichment
- Formats for human-computer interaction

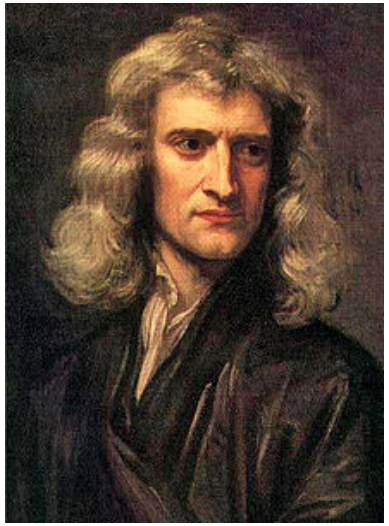
Most important:

Community should define appropriate policies for research data

User-driven (bottom-up) approaches complement heuristic (top-down) information

x

Semantics as modelling layer



$\dot{y}$ $\frac{dy}{dt}$

Transport of Heat

$$\partial_t(\rho^{\text{bulk}} c_p^{\text{bulk}} T) + \text{div}(\vec{u}\rho^{\text{fluid}} c_p^{\text{fluid}} T) - \text{div}(\lambda \text{grad} T) - S^T = 0$$

Hydraulics

$$\rho S \partial_t p - \text{div}(\rho \frac{K}{\mu} \text{grad}(p)) + \text{div}(\rho^2 \frac{K}{\mu} \vec{g}) - S^m = 0$$

Mechanics

$$\sigma_{kl} = D_{ijkl} \epsilon_{kl}$$

Chemistry

$$\partial_t(C^j \rho^{\text{fluid}}) + \text{div}(\vec{u}\rho^{\text{fluid}} C^j) - \text{div}(D^j \text{grad} C^j) - S^{C^j} = 0$$

The shape of things to come? (II) - Algorithms, data, and semantics

- Increasingly powerful generic algorithms
- Increasing value of the appropriate data fed into these algorithms
- Even more increasing value of the appropriate model

zbMATH the first resource for mathematics

FIZ Karlsruhe Leibniz Institute for Information Infrastructure

Thank you !

zbMATH
the first resource for mathematics

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

A top-down view on semantics: Content Analysis in Mathematics – Why?

- Embedding a publication in its general mathematical context
- Standardized content description

→ Easy searching for content also if no further content description like abstracts available

→ Possibility for visualization of embedded content

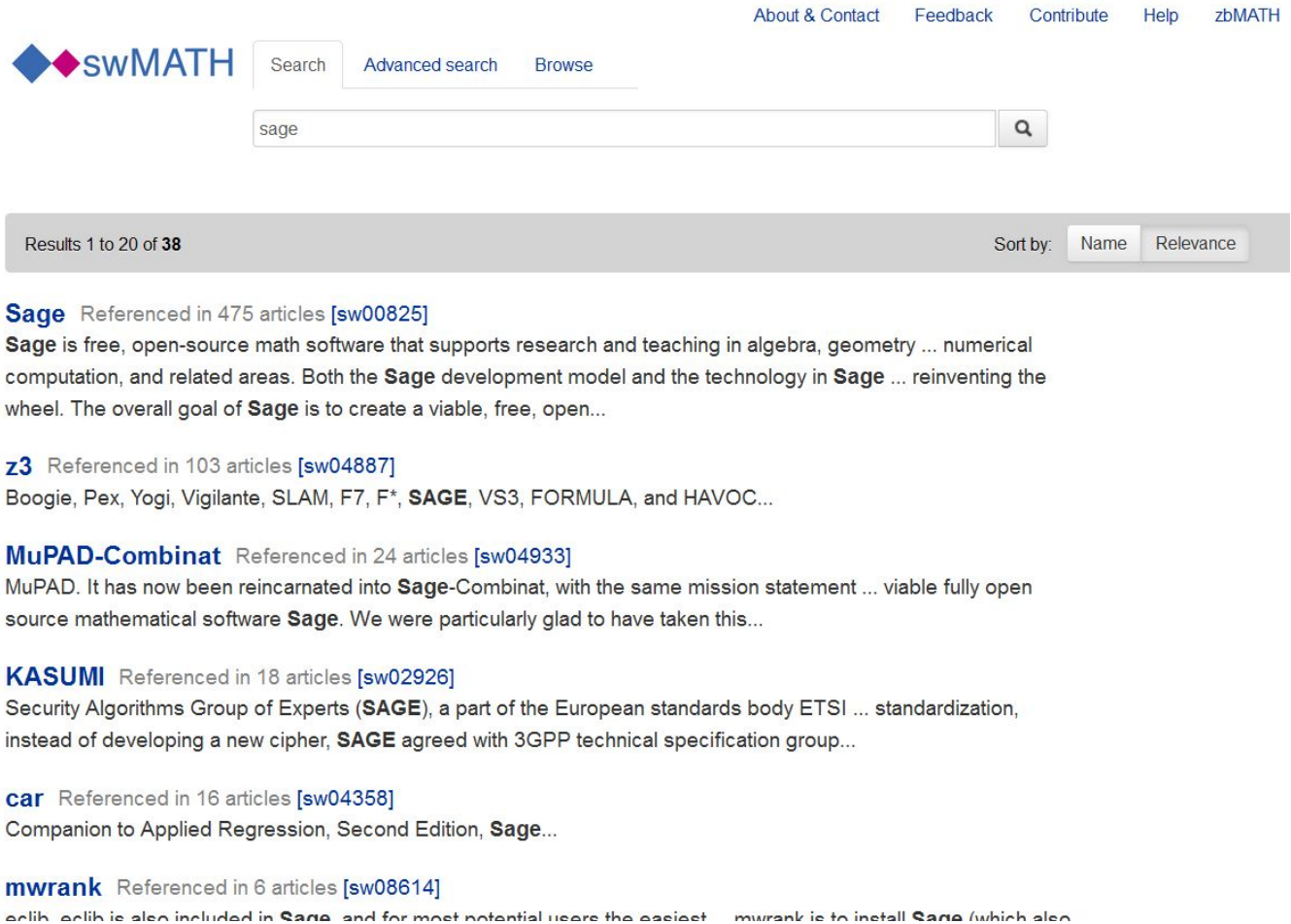
→ Possibility for data analysis

→ Possibility for search interaction (proposal of search terms)

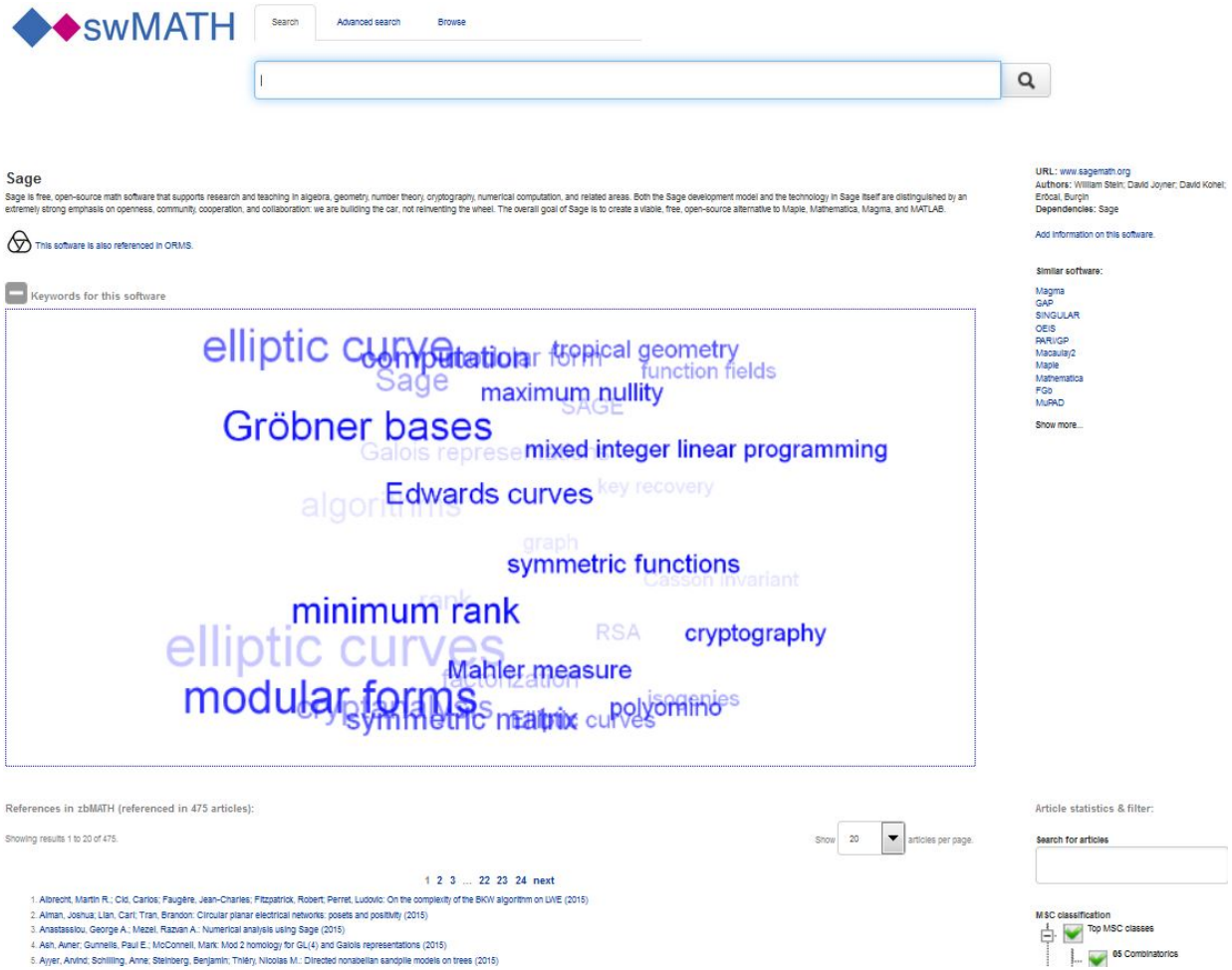
Mathematical software as additional layer (I)

Since 2011, software information is extracted from the corpus mainly by heuristics + intellectual refinement

As of Jun 2015, 10.064 systems and packages have been linked to 89.124 publications



Mathematical software as additional layer (II)



While the area of application is well reflected by corresponding documents, there is up to now no software classification/ontology represential typical facets like dependencies or functions (e.g., computer algebra systems, solvers, interfaces…)

zbMATH
the first resource for mathematics

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Formula search (I) - Overview

Motivation: Enable Access to formula content

Steps:

- Extract formula information from corpora
- Statistical analysis
- XML conversion (MathML 2.0 W3C standard), normalisation

➢ realized for zbMATH data to enable first prototype

Further challenges:

- Semantic analysis, disambiguation (example: ``what does dx/dy mean?´´)
- Integration with textual information

Major tasks are the digitisation of math content, normalisation, and semantic enrichment

Formula search (II) - Example

zbMATH
the first resource for mathematics

Documents | Authors | Journals | Classification | Software | Formulæ

\int_?a^?b (?f(x))^2 dx=?r | Examples | Help

$$\int_a^b (f(x))^2\,dx = r$$

Nowak, Werner Georg

Lattice points in a circle: an improved mean-square asymptotics. (English) Zbl 1092.11039

Acta Arith. 113, No. 3, 259-272 (2004).

Summary: Let $P(x)$ denote the lattice point discrepancy (number of integer points minus area) of an origin-centered compact circular disc of radius $\sqrt{x}$. The question for the order of magnitude of $P(x)$, for large x, is known as the Gaussian circle problem. In fact, in this most classic topic, the sharpest known results have been established only quite recently [*M. N. Huxley*, Proc. Lond. Math. Soc. (3) 87, 591–609 (2003; Zbl 1065.11079)]: Huxley proved that

$$P(x) \ll x^{131/416}(\log x)^{18637/8320}, \qquad \left(\frac{131}{416} = 0\cdot 3149\cdots\right)$$

while *K. Soundararajan* [Int. Math. Res. Not. 2003, No. 36, 1987–1998; Zbl 1130.11329)] showed that

$$P(x) \neq o\left((x\log x)^{1/4}(\log\log x)^{(3/4)\left(2^{1/3}-1\right)}(\log\log\log x)^{-5/8}\right).$$

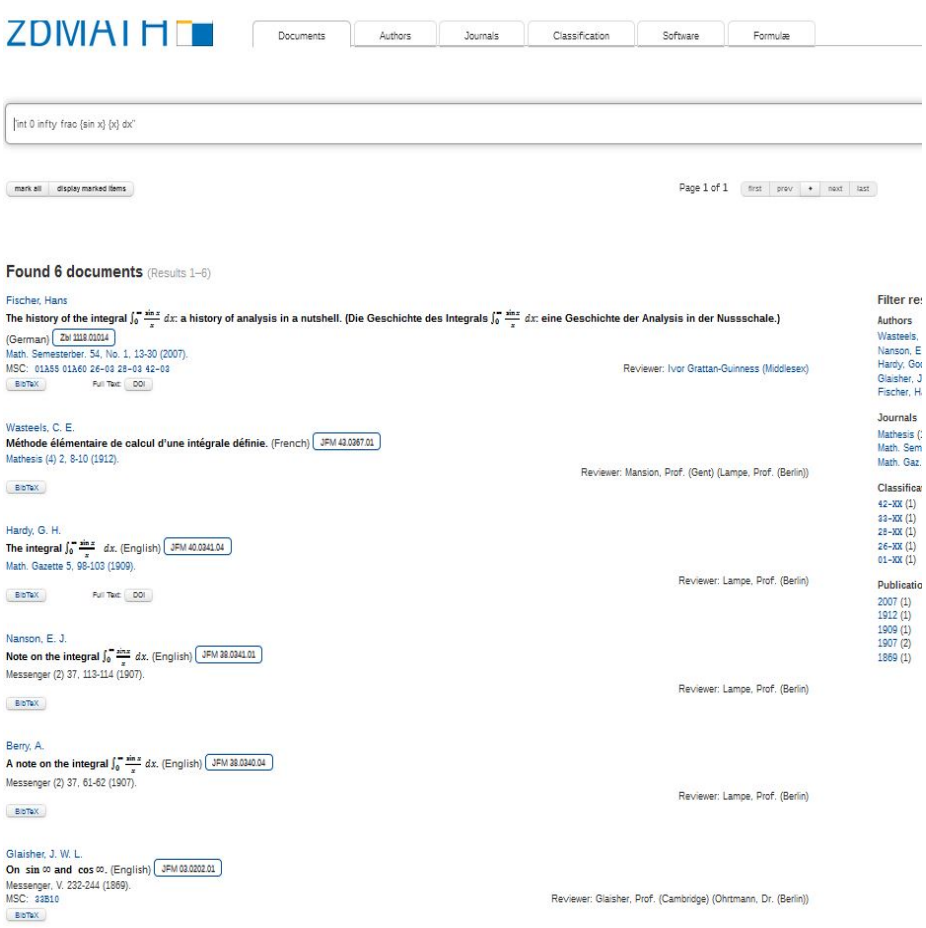
It is known that $P(x) \ll x^{1/4}$ *in mean-square*. More precisely,

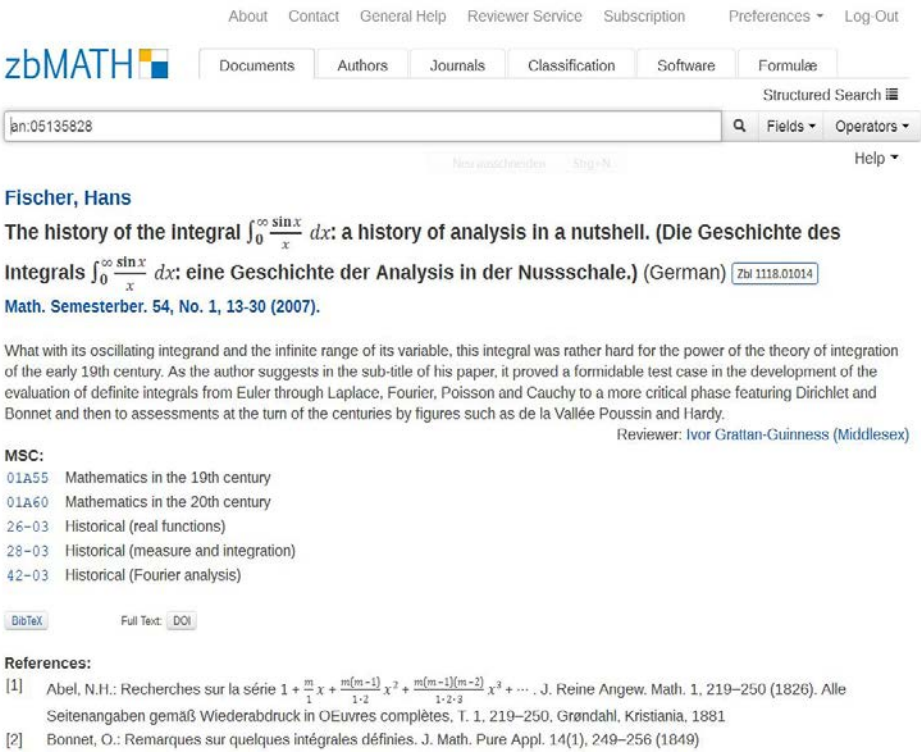
$$\int_0^X (P(x))^2\,dx = CX^{3/2} + Q(X), \qquad C \approx 1\cdot 69396\cdot$$

For this new error term $Q(X)$, the present article provides the refined upper bound

$$Q(X) \ll X(\log X)^{3/2}\log\log X,$$

Formula search (III) – Usability





Example:

Naive pure-tex-searches may still provide the best results.

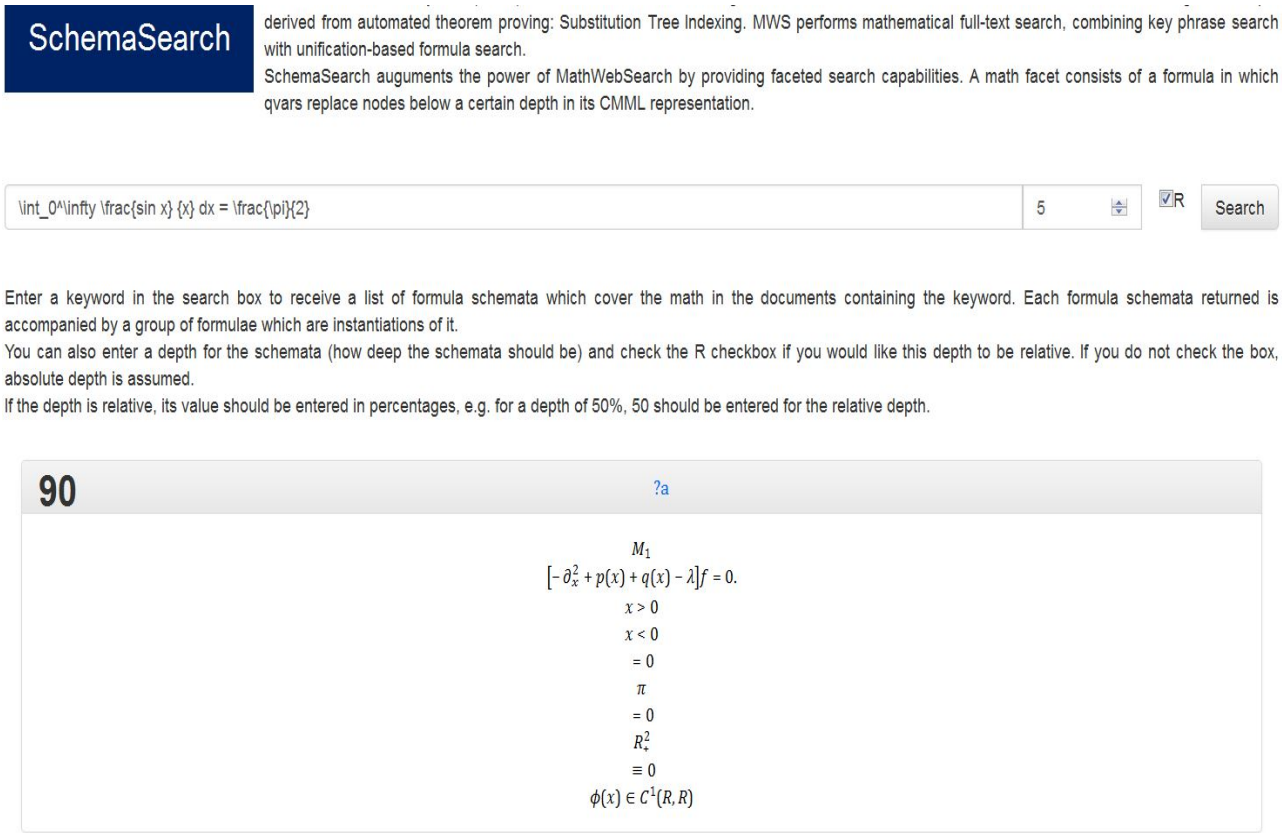
zbMATH
the first resource for mathematics

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Formula search (IV) – usability (II)

Extra challenge: find a way to create a user-friendly gateway to formula retrieval!

Unsuccessful approach (so far): Enable formula browsing based on statistics.



Thank you !

zbMATH
the first resource for mathematics

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

FIZ Karlsruhe

Leibniz Institute for Information Infrastructure