OpenDreamKit Deliverable D6.1 Full-text Search (Formulae + Keywords) over LaTeX-based Documents (e.g. the arXiv subset)

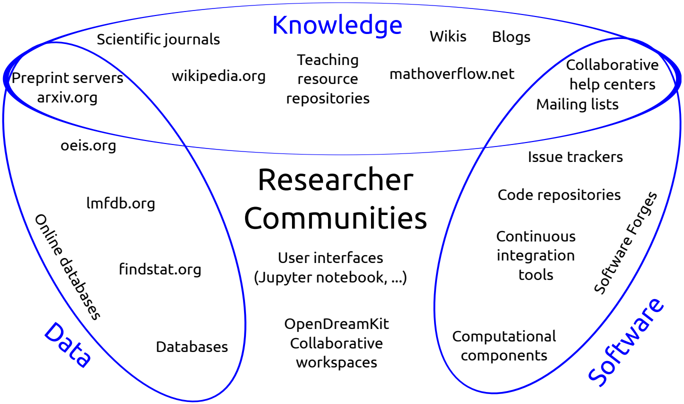
Michael Kohlhase

http://kwarc.info/kohlhase
Computer Science
Jacobs University Bremen, Germany

OpenDreamKit Workshop, Bremen, 28. June 2016

Background: WP6 (Data/Knowledge/Software-Bases)

- The Big Picture:



Knowledge

Scientific journals

Wikis

Blogs

Preprint servers
arxiv.org

wikipedia.org

Teaching
resource
repositories

mathoverflow.net

Collaborative
help centers
Mailing lists

oeis.org

Issue trackers

Researcher
Communities

lmfdb.org

Code repositories

Online databases

Software Forges

Continuous
integration
tools

findstat.org

User interfaces
(Jupyter notebook, ...)

Data

Databases

OpenDreamKit
Collaborative
workspaces

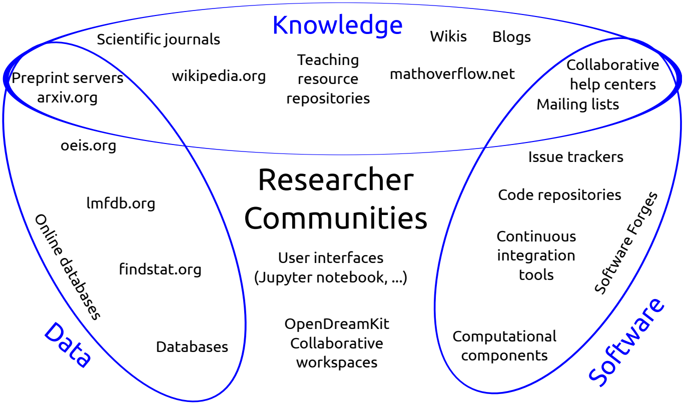
Computational
components

Software

Storage resources (local, shared folder, cloud)	Computational resources (local, super computer, cloud)

Background: WP6 (Data/Knowledge/Software-Bases)

- The Big Picture:



Knowledge

Scientific journals

Wikis

Blogs

Preprint servers
arxiv.org

wikipedia.org

Teaching
resource
repositories

mathoverflow.net

Collaborative
help centers
Mailing lists

oeis.org

Issue trackers

Researcher
Communities

lmfdb.org

Code repositories

Online databases

Software Forges

Continuous
integration
tools

findstat.org

User interfaces
(Jupyter notebook, ...)

OpenDreamKit
Collaborative
workspaces

Computational
components

Data

Databases

Software

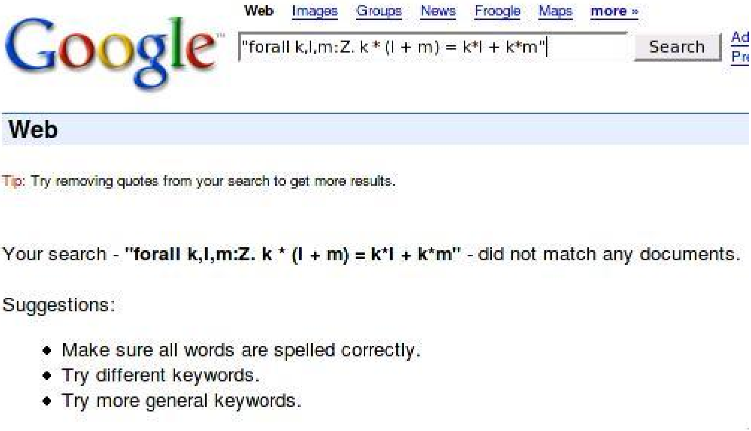
Storage resources (local, shared folder, cloud)	Computational resources (local, super computer, cloud)

- What do do with all this data/knowldege/software?: We need search!

More Mathematics on the Web

- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- ...Engineering Company Intranets, ...
- Question: How will we find content that is relevant to our needs
- Idea: try Google (like we always do)
- Sicenario: Try finding the distributivity property for $\mathbb{Z}$ ($\forall k, l, m \in \mathbb{Z}. k \cdot (l + m) = (k \cdot l) + (k + m)$)

Searching for Distributivity



"forall k,l,m:Z. k * (l + m) = k*l + k*m"

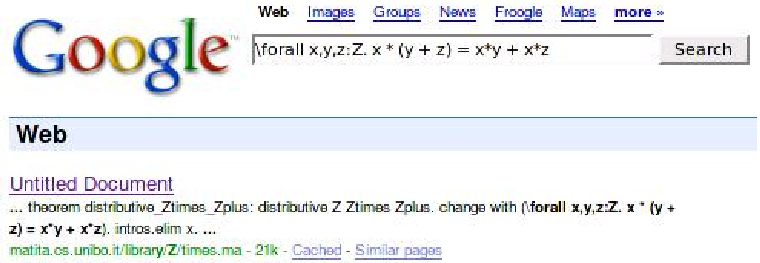
Tip: Try removing quotes from your search to get more results.

Your search - **"forall k,l,m:Z. k * (l + m) = k*l + k*m"** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

Searching for Distributivity



\forall x,y,z:Z. x * (y + z) = x*y + x*z

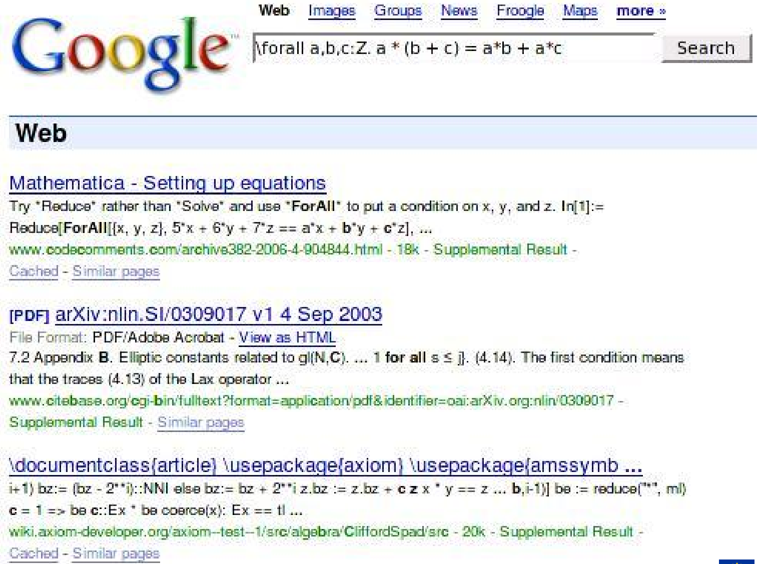
Web

Untitled Document

... theorem distributive_Ztimes_Zplus: distributive Z Ztimes Zplus. change with (**\forall x,y,z:Z. x * (y + z) = x*y + x*z**). intros.elim x. ...

matita.cs.unibo.it/library/Z/times.ma - 21k - Cached - Similar pages

Searching for Distributivity



\forall a,b,c:Z. a * (b + c) = a*b + a*c

Web

Mathematica - Setting up equations

Try "Reduce" rather than "Solve" and use "**ForAll**" to put a condition on x, y, and z. In[1]:= Reduce[**ForAll**[{x, y, z}, 5*x + 6*y + 7*z == a*x + **b***y + **c***z], ...

www.codecomments.com/archive382-2006-4-904844.html - 18k - Supplemental Result - Cached - Similar pages

[PDF] arXiv:nlin.SI/0309017 v1 4 Sep 2003

File Format: PDF/Adobe Acrobat - View as HTML

7.2 Appendix **B**. Elliptic constants related to gl(N,**C**). ... 1 **for all** s ≤ j}. (4.14). The first condition means that the traces (4.13) of the Lax operator ...

www.citebase.org/cgi-**b**in/fulltext?format=application/pdf&identifier=oai:arXiv.org:nlin/0309017 - Supplemental Result - Similar pages

\documentclass{article} \usepackage{axiom} \usepackage{amssymb ...

i+1) bz:= (bz - 2**i)::NNI else bz:= bz + 2**i z.bz := z.bz + **c z** x * y == z ... **b**,i-1)] be := reduce("*", ml) **c** = 1 => be **c**::Ex * be coerce(x): Ex == tl ...

wiki.axiom-developer.org/axiom--test--1/src/algebra/CliffordSpad/src - 20k - Supplemental Result - Cached - Similar pages

Does Image Search help?

- Math formulae are visual objects, after all (let's try it)

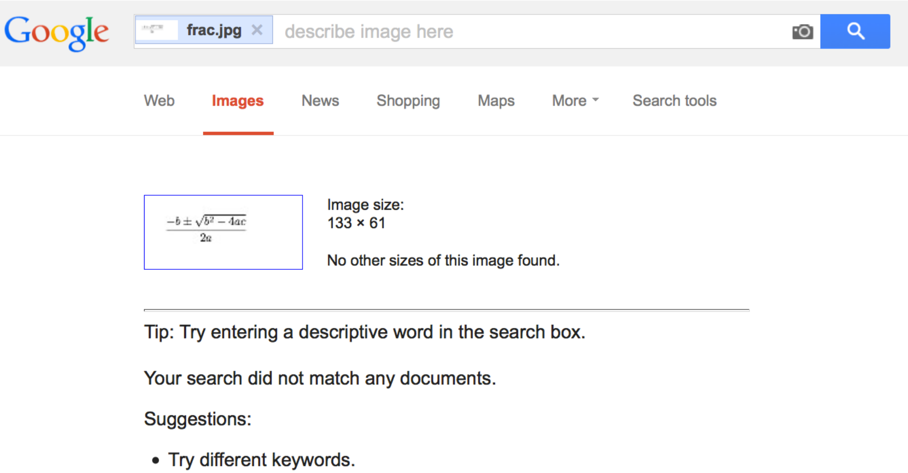


Image size:
133 × 61

No other sizes of this image found.

Tip: Try entering a descriptive word in the search box.

Your search did not match any documents.

Suggestions:

- Try different keywords.

Of course Google cannot work out of the box

- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables. (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Formulae are not images either: They have internal (recursive) structure and compositional meaning
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, MG11])
  . . . and works surprisingly well (using e.g. Lucene as an indexing engine)
- Idea: Use database techniques (extract metadata and index it)
  Indeed this is done for the `Coq`/`HELM` corpus ([AGC$^+$06])
- Our Idea: Use Automated Reasoning Techniques (free term indexing from theorem prover jails)
- Demo: MathWebSearch on Zentralblatt Math, the arXiv Data Set

Instantiation Queries

- Application: Find partially remembered formulae
- **Example 0.1** An engineer might face the problem remembering the energy of a given signal $f(x)$
  - Problem: hmmmm, have to square it and integrate
  - Query Term: $\int_{\boxed{min}}^{\boxed{max}} \boxed{f}(x)^2 dx$ ($\boxed{i}$ are search variables)
  - One Hit: Parseval's Theorem $\frac{1}{T}\int_0^T s^2(t)dt = \sum_{k=-\infty}^{\infty} \|c_k\|^2$ (nice, I can compute it)
- This works out of the box (has ween working in MathWebSearch for some time)
- Another Application: Underspecified Conjectures/Theorem Proving
  - during theory exploration we often have some freedom
  - express that using metavariables in conjectures
  - instantiate the conjecture metavariables as the proof as the proof dictates

  applied e.g. in Alan Bundy's "middle-out reasoning" in proof planing

Generalization Queries

- Application: Find (possibly) applicable theorems
- **Example 0.2** A researcher wants to estimate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)| dt$ from above
  - Idea: Find inequation such that $\int_{\mathbb{R}^2} |\sin(t)\cos(t)| dt$ matches left hand side.
  - Query: $\int_{\mathbb{R}^2} |\sin(x)\cos(x)|\, dx \leq \boxed{rhs}$
  - matches e.g. Hölder's Inequality in the index: ($\boxed{i}$ are universal variables)

$$\int_{\boxed{D}} \left|\boxed{f}(x)\boxed{g}(x)\right| dx \leq \left(\int_{\boxed{D}} \left|\boxed{f}(x)\right|^p dx\right)^{\frac{1}{p}} \left(\int_{\boxed{D}} \left|\boxed{g}(x)\right|^q dx\right)^{\frac{1}{q}}$$

  - Solution: Instantiate query accordingly and get

$$\int_{\mathbb{R}^2} |\sin(x)\cos(x)|\, dx \leq \left(\int_{\mathbb{R}^2} |\sin(x)|^p\, dx\right)^{\frac{1}{p}} \left(\int_{\mathbb{R}^2} |\cos(x)|^q\, dx\right)^{\frac{1}{q}}$$

Problem: Where do the index formulae come from in particular the universal variables (we'll come back to that later)

Where do the universal variables come from

- Problem: we need to have e.g. Hölder's Inequality in the index:

$$\int_{\boxed{D}} \left|\boxed{f}(x)\boxed{g}(x)\right| dx \leq \left(\int_{\boxed{D}} \left|\boxed{f}(x)\right|^p dx\right)^{\frac{1}{p}} \left(\int_{\boxed{D}} \left|\boxed{g}(x)\right|^q dx\right)^{\frac{1}{q}}$$

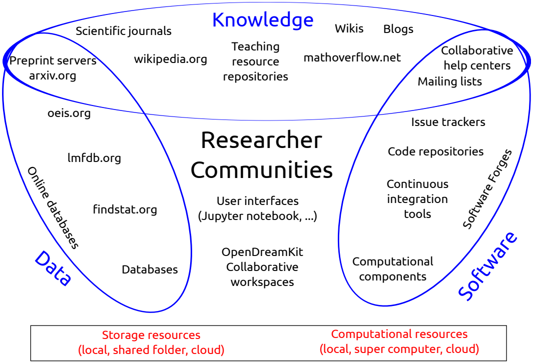
  - How do we know what symbols are "universal" (to be instantiated?)
  - what is their scope (when are different occurrences of f different?)
  - we have no sources with explicit quantifiers, but ([Wikipedia])

    Let (D, Σ, μ) be a measure space and let $1 \leq p,q \leq \infty$ with $1/p + 1/q = 1$. Then, for all measurable real- or complex-valued functions f and g on D, ...

- Solution: Use techniques from computational linguistics and integrate them into the indexing pipeline. (we have started a bit on the arXiv)
- Another Solution: Use born-formal representations (e.g. theorem prover libraries, computer algebra data, knowledge bases)

Back to OpenDreamKit as a VRE

- OpenDreamKit builds on an ecosystem of Data/Knowledge/Software



Knowledge

Scientific journals

Wikis

Blogs

Preprint servers
arxiv.org

wikipedia.org

Teaching
resource
repositories

mathoverflow.net

Collaborative
help centers

Mailing lists

oeis.org

Issue trackers

Researcher
Communities

lmfdb.org

Code repositories

Online databases

Software Forges

User interfaces
(Jupyter notebook, ...)

findstat.org

Continuous
integration
tools

Data

Databases

OpenDreamKit
Collaborative
workspaces

Computational
components

Software

Storage resources
(local, shared folder, cloud)

Computational resources
(local, super computer, cloud)

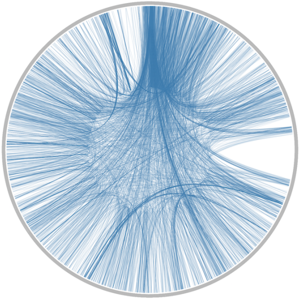
Joint search is a global service that binds them together into a math VRE

- Call for Action: export text + content MathML from all ODK components
- Preview: OEIS Search

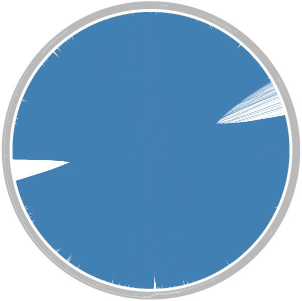
Auto-Discovering Relations between OEIS Sequences

- Idea: use the data that is already in the OEIS (see [LK16])
  - parse the (ASCII-art) formulae in the OEIS (⇝ content MathML)
  - find relations between the "generating functions" of sequence
  - submit back to the OEIS
- Results:

current realations



with ODK (one B.Sc.)



Andrea Asperti, Ferruccio Guidi, Claudio Sacerdoti Coen, Enrico Tassi, and Stefano Zacchiroli.
A content based mathematical search engine: Whelp.
In Jean-Christophe Filliâtre, Christine Paulin-Mohring, and Benjamin Werner, editors, *Types for Proofs and Programs, International Workshop, TYPES 2004, revised selected papers*, number 3839 in LNCS, pages 17–32. Springer Verlag, 2006.

Enxhell Luzhnica and Michael Kohlhase.
Formula semantification and automated relation finding in the OEIS.
In *Mathematical Software - ICMS 2016 - 5th International Congress*, LNCS. Springer, 2016.
accepted.

Paul Libbrecht and Erica Melis.
Methods for Access and Retrieval of Mathematical Content in ActiveMath.
In N. Takayama and A. Iglesias, editors, *Proceedings of ICMS-2006*, number 4151 in LNAI, pages 331–342. Springer Verlag, 2006.
`http://www.activemath.org/publications/Libbrecht-Melis-Access-and-Retrieval-ActiveMath-ICMS-2006.pdf`.

Jozef Misutka and Leo Galambos.

System description: Egomath2 as a tool for mathematical searching on wikipedia.org.
In James Davenport, William Farmer, Florian Rabe, and Josef Urban, editors, *Calculemus/MKM*, number 6824 in LNAI, pages 307–309. Springer Verlag, 2011.

Rajesh Munavalli and Robert Miner.
Mathfind: a math-aware search engine.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–735, New York, NY, USA, 2006. ACM Press.

Bruce R. Miller and Abdou Youssef.
Technical aspects of the digital library of mathematical functions.
Annals of Mathematics and Artificial Intelligence, 38(1-3):121–136, 2003.