

# Tapping Sources of Mathematical (Big) Data

Michael Kohlhase

Professur für Wissensrepräsentation und -verarbeitung  
Informatik, FAU Erlangen-Nürnberg  
<http://kwarc.info>

March 27. 2017, AITP Obergurgl

## Take-Home Message (I will probably run out of time)

- ▶ I only go GOFAI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)

## Take-Home Message (I will probably run out of time)

- ▶ I only go GOFAI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)
- ▶ no DLAI (applying Deep Learning to everything)

## Take-Home Message (I will probably run out of time)

- ▶ I only go GOF AI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)
- ▶ no DLAI (applying Deep Learning to everything)
- ▶ BUT we have a lot of interesting Data
  - ▶ arXiv preprints and ZMath Abstracts (licensing problems)
  - ▶ OAF: the Open Archive of Formalizations (<http://oaf.mathhub.info>)
  - ▶ OEIS: "Conjecturing relations between Sequences" ([https://github.com/eluzhnica/\\*](https://github.com/eluzhnica/*))

# Take-Home Message (I will probably run out of time)

- ▶ I only go GOF AI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)
- ▶ no DLAI (applying Deep Learning to everything)
- ▶ BUT we have a lot of interesting Data
  - ▶ arXMLiv preprints and ZBMath Abstracts (licensing problems)
  - ▶ OAF: the Open Archive of Formalizations (<http://oaf.mathhub.info>)
  - ▶ OEIS: "Conjecturing relations between Sequences" ([https://github.com/eluzhnica/\\*](https://github.com/eluzhnica/*))
- ▶ Could use DLAI help (but not in ATP improvements)

# Take-Home Message (I will probably run out of time)

- ▶ I only go GOFAI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)
- ▶ no DLAI (applying Deep Learning to everything)
- ▶ BUT we have a lot of interesting Data
  - ▶ arXMLiv preprints and ZBMath Abstracts (licensing problems)
  - ▶ OAF: the Open Archive of Formalizations (<http://oaf.mathhub.info>)
  - ▶ OEIS: "Conjecturing relations between Sequences" ([https://github.com/eluzhnica/\\*](https://github.com/eluzhnica/*))
- ▶ Could use DLAI help (but not in ATP improvements)
- ▶ I am looking for good GOFAI Ph.D. students (maybe even DLFAI)

# 1 Background: Towards a Math Digital Library

# Towards a World Digital Library of Mathematics

- ▶ **Mathematics** plays a fundamental role in Science, Technology, and Engineering  
(learn from Math, apply for STEM)
- ▶ Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!

# Towards a World Digital Library of Mathematics

- ▶ **Mathematics** plays a fundamental role in Science, Technology, and Engineering  
(learn from Math, apply for STEM)
- ▶ Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- ▶ There is a lot of documents with maths
  - ▶ there are **120.000 journal articles per year** in pure/applied math, **3.5 Million overall**
  - ▶ **50 million science articles** in 2010 [Jin10] with a **doubling time** of **8-15 years** [Lvl10]

# Towards a World Digital Library of Mathematics

- ▶ **Mathematics** plays a fundamental role in Science, Technology, and Engineering  
(learn from Math, apply for STEM)
- ▶ Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- ▶ There is a lot of documents with maths
  - ▶ there are **120.000 journal articles per year** in pure/applied math, **3.5 Million overall**
  - ▶ **50 million science articles** in 2010 [Jin10] with a **doubling time** of **8-15 years** [Lvl10]
- ▶ We need to preserve this heritage and make it accessible to working mathematicians!

# Towards a World Digital Library of Mathematics

- ▶ **Mathematics** plays a fundamental role in Science, Technology, and Engineering  
(learn from Math, apply for STEM)
- ▶ Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- ▶ There is a lot of documents with maths
  - ▶ there are **120.000 journal articles per year** in pure/applied math, **3.5 Million overall**
  - ▶ **50 million science articles** in 2010 [Jin10] with a **doubling time** of **8-15 years** [Lvl10]
- ▶ We need to preserve this heritage and make it accessible to working mathematicians!
- ▶ The EUDML Project digitized large amounts of European Journals
- ▶ The (US) National Research Council issued a Plan/Report for a “World Digital Heritage Library of Mathematics” [DLC<sup>+</sup>14].
  - ▶ Form a non-profit organization IMKT (Sloan grant for founding)
  - ▶ digitize, standardize, and **semanticize** math content (→ added value services)
  - ▶ Collaborate with Publishers/Organizations (to obtain rights)
- ▶ The International Mathematical Union (IMU) chartered a WG to bring this about.

# Background: Mathematical Documents

- ▶ **Mathematics** plays a fundamental role in Science, Technology, and Engineering  
(learn from Math, apply for STEM)
- ▶ Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation,
- ▶ its conservation, dissemination, and utilization constitutes a challenge for the community and an attractive line of inquiry.
- ▶ **Challenge**: How can/should we do mathematics in the 21<sup>st</sup> century?
- ▶ Mathematical knowledge and objects are transported by documents
- ▶ **Three levels of electronic documents**:
  0. **printed** (for archival purposes) (~90%)
  1. **digitized** (usually from print) (~50%)
  2. **presentational**: encoded text interspersed with presentation markup (~20%)
  3. **semantic**: encoded text with functional markup for the meaning ( $\leq 0.1\%$ )transforming down is simple, transforming up needs humans or AI.
- ▶ **Observation**: Computer support for access, aggregation, and application is (largely) restricted to the semantic level.
- ▶ **This talk**: How do we do maths and math documents at the semantic level?

# But there is is more Math Knowledge than Documents

- ▶ There are large mathematical data bases
    - ▶ Zentralblatt Math: the first resource in Maths (<http://zbmath.org>)
    - ▶ MathSciNet: Mathematical Reviews (<http://www.ams.org/mathscinet/>)
    - ▶ LMFDB: L-functions & Modular Forms (<http://lmfdb.org>)
    - ▶ OEIS: Open Encyclopedia of Integer Sequences (<http://oeis.org>)
    - ▶ FindStat: Combinatoria Statistics Finder (<http://findstat.org>)
    - ▶ MGP: Math Genealogy Project (<http://www.genealogy.math.ndsu.nodak.edu>)
- in various representations and licenses, at various states of maintenance/decay.

# But there is more Math Knowledge than Documents

- ▶ There are large mathematical data bases
  - ▶ Zentralblatt Math: the first resource in Maths (<http://zbmath.org>)
  - ▶ MathSciNet: Mathematical Reviews (<http://www.ams.org/mathscinet/>)
  - ▶ LMFDB: L-functions & Modular Forms (<http://lmfdb.org>)
  - ▶ OEIS: Open Encyclopedia of Integer Sequences (<http://oeis.org>)
  - ▶ FindStat: Combinatorial Statistics Finder (<http://findstat.org>)
  - ▶ MGP: Math Genealogy Project (<http://www.genealogy.math.ndsu.nodak.edu>)in various representations and licenses, at various states of maintenance/decay.
- ▶ **Idea:** Some of this information is already in a semantic/machine-actionable form.
- ▶ **Problems:** licenses, representations, versioning, GUIs, system APIs, ...
- ▶ **Idea:** To arrive at a core DML start at Math DBs and
  - ▶ specify open licenses  $\leadsto$  data commons
  - ▶ standardize representations  $\leadsto$  knowledge commons
  - ▶ even in maths, data changes  $\leadsto$  support versioning
  - ▶ system APIs  $\leadsto$  collaborate on content, compete on services

# But there is more Math Knowledge than Documents

- ▶ There are large mathematical data bases
  - ▶ Zentralblatt Math: the first resource in Maths (<http://zbmath.org>)
  - ▶ MathSciNet: Mathematical Reviews (<http://www.ams.org/mathscinet/>)
  - ▶ LMFDB: L-functions & Modular Forms (<http://lmfdb.org>)
  - ▶ OEIS: Open Encyclopedia of Integer Sequences (<http://oeis.org>)
  - ▶ FindStat: Combinatoria Statistics Finder (<http://findstat.org>)
  - ▶ MGP: Math Genealogy Project (<http://www.genealogy.math.ndsu.nodak.edu>)in various representations and licenses, at various states of maintenance/decay.
- ▶ **Idea**: Some of this information is already in a semantic/machine-actionable form.
- ▶ **Problems**: licenses, representations, versioning, GUIs, system APIs, ...
- ▶ **Idea**: To arrive at a core DML start at Math DBs and
  - ▶ specify open licenses  $\leadsto$  data commons
  - ▶ standardize representations  $\leadsto$  knowledge commons
  - ▶ even in maths, data changes  $\leadsto$  support versioning
  - ▶ system APIs  $\leadsto$  collaborate on content, compete on services
- ▶ **OpenDreamKit**: EU Project 2015-2019  $\leadsto$  Math Virtual Research Environment  
Computer Algebra, HPC, MathUI, KWARC (<http://opendreamkit.org>)

# Zentralblatt Math: the first resource in Maths

The screenshot shows the zbMATH website in a web browser. The address bar displays <https://zbmath.org> and the search bar contains the text "mathaci". The website's navigation menu includes links for About, Contact, General Help, FAQ, Reviewer Service, Subscription, and Preferences. The main heading features the zbMATH logo and the tagline "the first resource for mathematics". Below this, there are tabs for Documents, Authors, Journals, Classification, Software, and Formulae. A "Structured Search" section includes a search input field with the placeholder "Search for documents", a search button, and dropdown menus for "Fields" and "Operators". A "Help" link is also present. A welcome message states: "Welcome to the zbMATH interface. Explore our new features:". This is followed by a bulleted list of updates: feedback for document assignments can be given directly at author profile pages or via a link (if you already know the author ID); and the formula facet in the structured search now allows for a free combination of formula search with other query types, with formula queries now also being refined by filters. The footer contains logos for FIZ Karlsruhe, The European Mathematical Society, and the Heidelberg Academy of Sciences and Humanities.

https://zbmath.org

mathaci

Services News MathWeb Jacobs AG Lists TRAC Rotary Haus Most Visited SPP Semantic Arg...

About Contact General Help FAQ Reviewer Service Subscription Preferences

zbMATH  
the first resource for mathematics

Documents Authors Journals Classification Software Formulae

Structured Search

Search for documents

Fields Operators

Help

**Welcome to the zbMATH interface.**  
Explore our new features:

- Feedback to improve document assignments can be given directly at the author profile pages or via [this link](#) (if you already know the author ID).
- The formula facet in the [structured search](#) now allows for a free combination of formula search with other query types. Formula queries can now also be refined by filters.

FIZ Karlsruhe  
Leibniz Institute for Information Infrastructure

The European  
Mathematical Society

HEIDELBERG ACADEMY OF  
SCIENCES AND HUMANITIES

# MathSciNet: Mathematical Reviews

The screenshot shows the MathSciNet website in a web browser. The browser's address bar displays `www.ams.org/mathscinet/` and the search bar contains the text `mathaci`. The website header includes the MathSciNet logo with a '75' anniversary badge, the ISSN 2167-5163, and navigation links such as Home, Preferences, Free Tools, About, Librarians, and Terms of Use. A secondary navigation bar lists various categories like Services, News, MathWeb, and others. The main content area features a search interface with tabs for Publications, Authors, Journals, and Citations. The 'Publications' tab is active, showing search fields for Author, Title, MSC Primary, and Anywhere, each with a dropdown menu. To the right of these fields are 'and' operators. Below the search fields are sections for Time Frame (Entire Database, Year, Year Range), Publication Type (All, Books, Journals, Proceedings), and Review Format (PDF, HTML). A 'Search' button and a 'Clear' button are at the bottom of the search area. A 'REMOTE ACCESS' banner is visible on the right side of the search area. At the bottom of the page, there is a footer with the AMS logo, mirror sites information, copyright notice, and privacy statement.

AMERICAN MATHEMATICAL SOCIETY  
**MathSciNet** 75  
Mathematical Reviews  
ISSN 2167-5163

Home Preferences Free Tools About Librarians Terms of Use

Jacobs University Bremen

Publications Authors Journals Citations

Search Terms

Author and  
Title and  
MSC Primary and  
Anywhere and

Time Frame  
☒ Entire Database  
☐ =  Year   
☐ Year Range:  to

Publication Type  
☒ All  
☐ Books  
☐ Journals  
☐ Proceedings

Review Format  
☐ PDF  
☒ HTML

Search Clear

REMOTE ACCESS  
Access your AMS  
online content  
wherever you go!  
Learn more

Facts and Figures: 3,207,917 total publications

Help Support Mail

AMS  
AMERICAN MATHEMATICAL SOCIETY

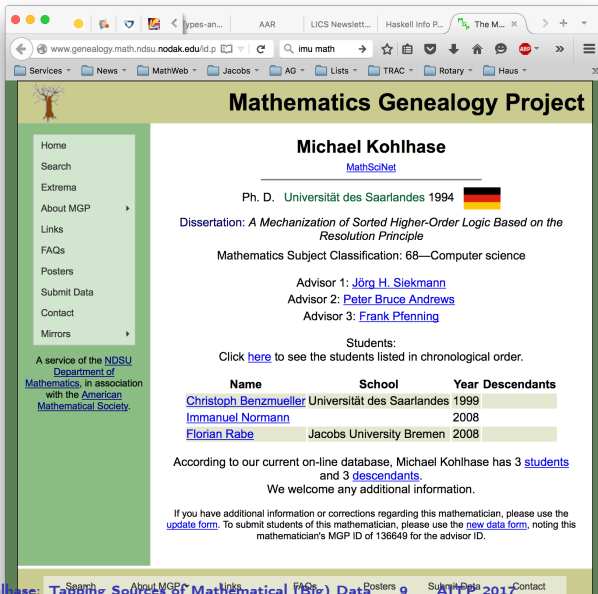
Mirror Sites Providence, RI USA  
© Copyright 2015, American Mathematical Society  
Privacy Statement

# LMFDB: L-functions & Modular Forms

The screenshot shows the LMFDB website with a green header and a sidebar on the left. The main content area is divided into several sections:

- Introduction and more**: Links to Introduction, Features, Tutorial, Map, Contact, Citation, and Future Plans.
- L-functions**: Degree: 1 2 3 4, First zeros,  $\zeta$  zeros, Operations.
- Modular Forms**: GL(2) Classical Maass Hilbert, GL(3) Maass, Other Siegel.
- Varieties**: Elliptic: /Q, /NumberFields, Genus 2: /Q, Curves.
- A Database**: The LMFDB is an extensive database of mathematical objects arising in Number Theory. Sample lists: L-functions, Elliptic curves, Maass forms, Tables of zeros, Number fields.
- Search and Browse**: Search for objects with specific properties, or browse categories. Browse: L-functions, Modular forms, Elliptic curves, Number fields.
- Hall of Fame**: Riemann zeta function, Cubic Field of discriminant -23, Ramanujan  $\Delta$  function and its L-function, First Rank 4 Elliptic curve and its L-function.
- Visualize Data**: Explore individual plots or view distributions of various objects. Examples: GL(4) Level one Maass forms, Isogeny graph of elliptic curve 102.c.
- Explore and Learn**: The LMFDB makes visible the connections predicted by the Langlands program. Knows offer background information when you need it.
- Code and Open Software**: Download the data, download the code, or see how the data were generated. Links to GitHub, SageMath, Pari/GP.


# MGP: Math Genealogy Project



The screenshot shows a web browser window displaying the Mathematics Genealogy Project website. The browser's address bar shows the URL [www.genealogy.math.ndsu.nodak.edu/id.php?id=136649](http://www.genealogy.math.ndsu.nodak.edu/id.php?id=136649). The website has a green header with the title "Mathematics Genealogy Project" and a tree icon. A left sidebar contains navigation links: Home, Search, Extrema, About MGP, Links, FAQs, Posters, Submit Data, Contact, and Mirrors. The main content area is for Michael Kohlhasse, a Ph.D. graduate from the University of Saarland in 1994. It lists his dissertation, subject classification, advisors (Jörg H. Siekmann, Peter Bruce Andrews, and Frank Pfenning), and students (Christoph Benzmueller, Immanuel Normann, and Florian Rabe). A table shows the names, schools, and years of his students. The page also includes a footer with search and navigation links.

Mathematics Genealogy Project

**Michael Kohlhasse**  
[MathSciNet](#)

Ph. D. Universität des Saarlandes 1994 

*Dissertation: A Mechanization of Sorted Higher-Order Logic Based on the Resolution Principle*

Mathematics Subject Classification: 68—Computer science

Advisor 1: [Jörg H. Siekmann](#)  
Advisor 2: [Peter Bruce Andrews](#)  
Advisor 3: [Frank Pfenning](#)

Students:  
Click [here](#) to see the students listed in chronological order.

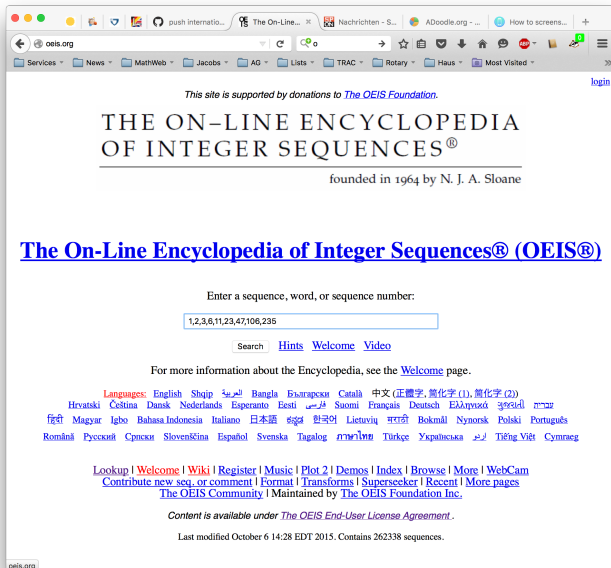
Name	School	Year	Descendants
<a href="#">Christoph Benzmueller</a>	Universität des Saarlandes	1999	
<a href="#">Immanuel Normann</a>		2008	
<a href="#">Florian Rabe</a>	Jacobs University Bremen	2008	

According to our current on-line database, Michael Kohlhasse has 3 [students](#) and 3 [descendants](#).  
We welcome any additional information.

If you have additional information or corrections regarding this mathematician, please use the [update form](#). To submit students of this mathematician, please use the [new data form](#), noting this mathematician's MGP ID of 136649 for the advisor ID.

Kohlhasse: [Search](#) [About MGP](#) [Links](#) [FAQs](#) [Posters](#) [Submit Data](#) [Contact](#)

# OEIS: Open Encyclopedia of Integer Sequences



The screenshot shows the OEIS website in a web browser. The browser's address bar shows 'oeis.org'. The page has a navigation bar with links like 'Services', 'News', 'MathWeb', 'Jacobs', 'AG', 'Lists', 'TRAC', 'Rotary', 'Haus', and 'Most Visited'. The main content area features the title 'THE ON-LINE ENCYCLOPEDIA OF INTEGER SEQUENCES®' and the text 'founded in 1964 by N. J. A. Sloane'. Below this is a search box with the text 'Enter a sequence, word, or sequence number:' and a search button. The search box contains the sequence '1,2,3,6,11,23,47,106,235'. There are also links for 'Hints', 'Welcome', and 'Video'. A paragraph of text follows, and then a list of languages. At the bottom, there are links for 'Lookup', 'Welcome', 'Wiki', 'Register', 'Music', 'Plot 2', 'Demos', 'Index', 'Browse', 'More', 'WebCam', 'Contribute new seq. or comment', 'Format', 'Transforms', 'Superseeker', 'Recent', 'More pages', 'The OEIS Community', and 'Maintained by The OEIS Foundation Inc.'. The footer of the page states 'Content is available under The OEIS End-User License Agreement.' and 'Last modified October 6 14:28 EDT 2015. Contains 262338 sequences.'

This site is supported by donations to [The OEIS Foundation](#). [login](#)

## THE ON-LINE ENCYCLOPEDIA OF INTEGER SEQUENCES®

founded in 1964 by N. J. A. Sloane

### [The On-Line Encyclopedia of Integer Sequences® \(OEIS®\)](#)

Enter a sequence, word, or sequence number:

[Hints](#) [Welcome](#) [Video](#)

For more information about the Encyclopedia, see the [Welcome](#) page.

**Languages:** English Shqip العربية Bangla Български Català 中文 (正體字, 简化字 (1), 简化字 (2))  
Hrvatski Čeština Dansk Nederlands Esperanto Eesti فارسی Suomi Français Deutsch Ελληνικά עברית  
हिंदी Magyar Igbo Bahasa Indonesia Italiano 日本語 ಕನ್ನಡ 한국어 Lietuvių मराठी Бокмål Nynorsk Polski Português  
Română Pycckий Cрпски Slovenščina Español Svenska Tagalog தமிழ் Türkçe Українська اردو Tiếng Việt Cymraeg

[Lookup](#) | [Welcome](#) | [Wiki](#) | [Register](#) | [Music](#) | [Plot 2](#) | [Demos](#) | [Index](#) | [Browse](#) | [More](#) | [WebCam](#)  
[Contribute new seq. or comment](#) | [Format](#) | [Transforms](#) | [Superseeker](#) | [Recent](#) | [More pages](#)  
[The OEIS Community](#) | Maintained by [The OEIS Foundation Inc.](#)

Content is available under [The OEIS End-User License Agreement](#).

Last modified October 6 14:28 EDT 2015. Contains 262338 sequences.

oeis.org

# Take-Home Message: Digital Libraries for Maths

---

- ▶ There is a lot of useful data in maths out there

# Take-Home Message: Digital Libraries for Maths

---

- ▶ There is a lot of useful data in maths out there
- ▶ But it needs integration, aggregation, and versioning

# Take-Home Message: Digital Libraries for Maths

---

- ▶ There is a lot of useful data in maths out there
- ▶ But it needs integration, aggregation, and versioning
- ▶ Licensing is a major stumbling block

## 2 Converting the arXiv

# The arXMLiv Project: arXiv to semantic XML

- ▶ **Idea:** Develop a large corpus of knowledge in *OMDoc/PhysML*
  - ▶ to get around the chicken-and-egg problem of MKM
  - ▶ corpus-linguistic methods for semantics recovery (linguists interested)
- ▶ **Definition 2.1 (The Cornell Preprint arXiv)** (<http://www.arxiv.org>)  
Open access to ca. 850K e-prints in Physics, Mathematics, Computer Science and Quantitative Biology.
- ▶ **Definition 2.2 (The arXMLiv Project)** (<http://arxmliv.kwarc.info>)
  - ▶ use Bruce Miller's  $\text{\LaTeX}$ XML to transform to XHTML+MathML
  - ▶ extend to  $\text{\LaTeX}$ XML daemon (RESTful web service) (<http://latexml.mathweb.org>)
  - ▶ we have an automated, distributed build system (ca. Q2CPU-years)
  - ▶ create ca. 12K  $\text{\LaTeX}$ XML binding files (8 Jacobs students help)
  - ▶ use MathWebSearch to index XML version (realistic search corpus)
- ▶ More semantic information will enable more added-value services, e.g.
  - ▶ filter hits by model assumptions (expanding, stationary, or contracting universe)
  - ▶ use linguistic techniques to add the necessary semantics

# Why reimplement the T<sub>E</sub>X parser? I

- ▶ **Problem:** The T<sub>E</sub>X parser can change the tokenizer while at runtime (`\catcode`)
- ▶ **Example 2.3 (Obfuscated T<sub>E</sub>X)** David Carlisle posted the following, when someone claimed that word counting is simple in T<sub>E</sub>X/L<sup>A</sup>T<sub>E</sub>X

```
\let~\catcode~'76~'A13~'F1~'j00~'P2jdefA71F~'7113jdefPALLF
PA' 'FwPA;;FPAZZFLaLPA//71F71iPAHHFLPAzzFenPASSFthP;A$$FevP
A@@FfPARR717273F737271P;ADDFRgniPAWW71FPATTFvePA**FstRsamP
AGGFRruoPAqq71.72.F717271PAY7172F727171PA??Fi*LmPA&&71jfi
Fjfi71PAVVfjbigskipRPWGAUU71727374 75,76Fjpar71727375Djfix
:76jel se&U76jfiPLAKK7172F71I7271PAXX71FVLnOSeL71SLRyadR@oL
RrhC?yLRurtKFeLPFovPgaTLtReRomL;PABB71 72,73:Fjif.73.jel se
B73:jfiXF71PU71 72,73:PWs;AMM71F71diPAJJFRdriPAQQFRsreLPAI
I71Fo71dPA!!FRgiePBt'el@ITLqdrYmu.Q.,Ke;vz vzLqip.Q.,tz;
;Lql.IrsZ.eap,qn.i.i.eLIMaesLdRcna;;!;h htLqm.MRasZ.il k,%
s$;z zLqs'.ansZ.Ymi,/sx;LYegseZRyal,@i;@TLRlogdLrDsW,@;G
LcYlaDLbJsW,SWXJW ree @rzchLhzsW;;WERceslnW qt.'oL.Rtrul;e
doTsW,Wk;Rri@stW aHAHHFndZPpqar.tridgeLinZpe.LtYer.W,:jbye
```

When formatted by TeX, this leads to the full lyrics of “The twelve days of christmas”. When formatted by L<sup>A</sup>T<sub>E</sub>XML, it gives

# Why reimplement the T<sub>E</sub>X parser? II

```
<song>
<verse>
  <line>On the first day of Christmas my true love gave to me</line>
  <line>a partridge in a pear tree.</line>
</verse>
<verse>
  <line>On the second day of Christmas my true love gave to me</line>
  <line>two turtle doves</line>
  <line>and a partridge in a pear tree.</line>
</verse>
<verse>
  <line>On the third day of Christmas my true love gave to me</line>
  <line>three french hens</line>
  <line>two turtle doves</line>
  <line>and a partridge in a pear tree.</line>
</verse>
<verse>
  <line>On the fourth day of Christmas my true love gave to me</line>
  <line>four calling birds</line>
  <line>three french hens</line>
  <line>two turtle doves</line>
  <line>and a partridge in a pear tree.</line>
</verse>
...
```

# Why reimplement the T<sub>E</sub>X parser? III

- ▶ But the real reason is: that we can take advantage of the semantics in the L<sup>A</sup>T<sub>E</sub>X.
- ▶ L<sup>A</sup>T<sub>E</sub>XML does not need to expand macros, we can tell it about XML equivalents.
- ▶ **Example 2.4 (Recovering the Semantics of Proofs)**

Add the following magic incantation to `amsthm.sty.ltxml` (L<sup>A</sup>T<sub>E</sub>XML binding)

```
DefEnvironment('{proof}', "<html:div class='proof'>#body</html:div>");
```

The arXMLiv approach: Try to cover most packages and classes in the arXiv  
(Jacobs undergrads' intro to research)

# Future Plans for arXMLiv

- ▶ **State:**  $\text{\LaTeX}$ -to-XHTML+MathML Format Conversion works (65% success)
- ▶ **Over the summer:** Bump up success rate to 75%, daily downloads, web site, instrumentation,...
- ▶ **Soon:** Integrate user-level quality control (integrate JS feedback into html)
- ▶ **starting Fall:** Extend post-processing by linguistic methods for semantic analysis
  - ▶ build semantics blackboard/database for linguistic information (rdf triples)
  - ▶ extend build system for arbitrary XML2BB processes
  - ▶ invite the linguists over (they leave semantics results in BB)
  - ▶ harvest the semantics BB to get OMDoc representations

# Current and Possible Applications

- ▶ the arXiv build system <http://arxmliv.kwarc.info>
- ▶ the transformation web service <http://tex2xml.kwarc.info>
- ▶ L<sup>A</sup>T<sub>E</sub>XML daemon to avoid perl and L<sup>A</sup>T<sub>E</sub>X startup times (Deyan Ginev)
  - ▶ keep L<sup>A</sup>T<sub>E</sub>XML alive as a daemon that can process multiple files/fragments (patch memory leaks)
  - ▶ a L<sup>A</sup>T<sub>E</sub>XML client just passes files/fragments along (10/s to 100/s)
- ▶ embedding/editing L<sup>A</sup>T<sub>E</sub>X in web pages <http://tex2xml.kwarc.info/test>
- ▶ a MathML version of the arXiv allows vision-impaired readers to understand the texts
- ▶ generalization search (need to know sentence structure for detecting universal variables)
- ▶ semantic search by academic discipline or theory assumption (need discourse structure)
- ▶ development of scientific vocabularies (over the past 18 years; drink from the source)

## Take-Home Message: arXMLiv (I am skipping the slides)

- ▶ We can create large XML/MathML document corpora that preserve  $\text{\LaTeX}$  semantics (good for DLAI)

# Take-Home Message: arXMLiv (I am skipping the slides)

- ▶ We can create large XML/MathML document corpora that preserve  $\text{\LaTeX}$  semantics (good for DLAI)
- ▶ We have problems re-distributing them (Licensing) (working on this)
- ▶ Lots of potential Applications
  - ▶ Formula Search ([arxivsearch.mathweb.info](https://arxivsearch.mathweb.info), <https://zbmath.org/formulae/>)
  - ▶ screenreaders for quantity expressions (semantics extraction, annotation)
  - ▶ applicable theorem search (need to identify the universal/existential/constant identifiers)
  - ▶ machine translation (need a handle on the math terminology (large, dynamic))

All of these are a mixture of DLAI and GOFAI methods

# Take-Home Message: arXMLiv (I am skipping the slides)

- ▶ We can create large XML/MathML document corpora that preserve  $\text{\LaTeX}$  semantics (good for DLAI)
- ▶ We have problems re-distributing them (Licensing) (working on this)
- ▶ Lots of potential Applications
  - ▶ Formula Search ([arxivsearch.mathweb.info](http://arxivsearch.mathweb.info), <https://zbmath.org/formulae/>)
  - ▶ screenreaders for quantity expressions (semantics extraction, annotation)
  - ▶ applicable theorem search (need to identify the universal/existential/constant identifiers)
  - ▶ machine translation (need a handle on the math terminology (large, dynamic))

All of these are a mixture of DLAI and GOFAI methods

- ▶ I am sceptical of DLAI autoformalization (surprise me!)

### 3 OAF: Assembling a Global Resource of Formalizations

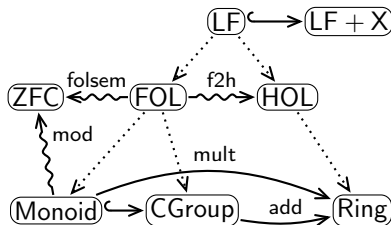
# OAF: Open Archive of Formalizations: Motivation

- ▶ **Idea1 (OAF):** Assemble all theorem prover libraries in a common synergy space
- ▶ **Observation:** Formal/symbolic systems and their libraries are non-interoperable
  - ▶ differing, mutually incompatible foundations (e.g., set theory, higher-order logic, constructive type theory, etc.),
  - ▶ library formats, and library structures,
- ▶ **Consequence:** Too much work is spent developing
  - ▶ basic libraries for mathematics in each system.
  - ▶ library organization features (e.g., distribution, browsing, search, change management) for each library format.
- ▶ **Problem:** All these investments bind resources that could be used to improve the core functionality of the systems and the scope of the libraries.
- ▶ **Idea2 (QAF = QED reloaded):** System and tool chain for all of formal maths!

- ▶ **Idea (OAF)**: Assemble all theorem prover libraries in a common synergy space
- ▶ **Problem**: Different systems have different, mutually incompatible logical/mathematical foundations (optimize different aspects)
- ▶ **Observation**: need a system with multiple foundations  $\leadsto$  foundational pluralism
- ▶ **Definition 3.1** A **foundation** (of mathematics) consists of
  - ▶ a **foundational language** (aka. logc, e.g. first-order logic or the CIC)
  - ▶ a **foundational theory** (e.g. axiomatic set theory)
- Idea1**: treat logics as mathematical theories themselves (metalogical frameworks)
- ▶ **Idea2**: relate logics in a theory graph via logic transformations (LATIN)

# Representing Logics and Foundations as Theories

- Logics and foundations represented as MMT theories (in the sample graph)

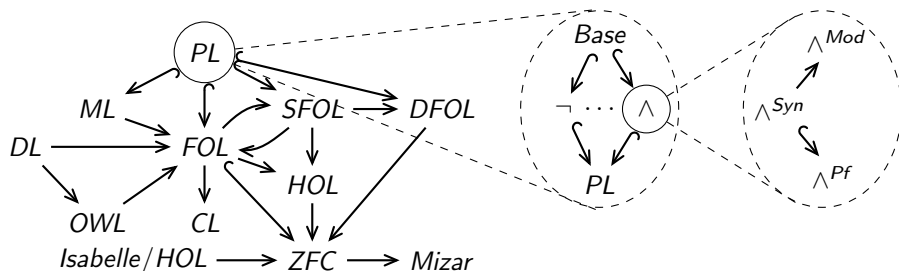


Meta-relation between theories – special case of inclusion (meta<sup>\*</sup>-level)

- Uniform Meaning Space:** morphisms between formalizations in different logics become possible via meta-morphisms.
- Remark 3.2** Semantics of logics as views into foundations, e.g., folsem.
- Remark 3.3** Models represented as views into foundations
- Example 3.4**  $\text{mod} := \{G \mapsto \mathbb{Z}, \circ \mapsto +, e \mapsto 0\}$  interprets Monoid in ZFC.

# The LATIN Logic Atlas

- **Definition 3.5** The LATIN project (Logic Atlas and Integrator) develops a logic atlas, its home page is at <http://latin.omdoc.org>.
- **Idea:** Provide a standardized, well-documented set of theories for logical languages, logic morphisms as theory morphisms.



- **Technically:** Use MMT as a representation language **logics-as-theories**
- Integrate logic-based software systems via views.
- **State:** About 1000 modules (theories and morphisms) written in MMT/LF [RS09]

# MMT a Module System for Mathematical Content

- ▶ MMT: Universal representation language for formal mathematical/logical content
- ▶ **Implementation:** MMT API with generic
  - ▶ module system for math libraries, logics, foundations
  - ▶ parsing + type reconstruction + simplification
  - ▶ IDE (web server + JEdit)
  - ▶ change management
- ▶ Continuous development since 2007 (> 30000 lines of Scala code)
- ▶ Close relatives:
  - ▶ LF, Isabelle, Dedukti: but flexible choice of logical framework
  - ▶ Hets: but declarative logic definitions

# Exports from Proof Assistants

- ▶ **General Approach:** Export library as MMT projects, store in MathHub
- ▶ **Library Export Architecture:** (this seems to work sustainably)
  - ▶ System-near export (e.g. to XML or JSON) as part of system code
  - ▶ aggregate, into OMDoc/MMT in MMT API system.
- ▶ **Current state of the collection effort:**
  - ▶ Mizar: set theoretical (initial export done (with Josef Urban))
  - ▶ HOL Family: HOL Light, HOL4, Isabelle, TPS (initial export done (Rabe/Kaliszyk))
  - ▶ Coq or Matita: type theoretical (Work with Sacerdoti Coen ongoing)
  - ▶ IMPS: heterogeneous method (Partial Export Done)
  - ▶ PVS: rich foundational language (Müller/Owre)
  - ▶ TPTP: mostly first-order ATP problems
  - ▶ Computer Algebra Signatures: GAP, Sage (Konovalov/Pfeiffer/Thierry)
  - ▶ Specware, OEIS, MetaMath, ... (experimental)

### ► Example 3.6 (Search in the MMT API/MathHub)

Enter Java regular expressions to filter based on the URI of a declaration

Namespace

Theory

Name

Enter an expression over theory

Use  $\$x,y,z$ :query to enter unification variables.

Search

type of **MOD\_EQ**

$\vdash \forall m:\text{num} . \forall n:\text{num} . \forall p:\text{num} . \forall q:\text{num} . m = n + q * p \implies m \text{ MOD } p = n \text{ MOD } p$

type of **MOD\_MULT\_ADD**

$\vdash \forall m:\text{num} . \forall n:\text{num} . \forall p:\text{num} . (m * n + p) \text{ MOD } n = p \text{ MOD } n$

# Goal: Towards Library Integration

- ▶ Refactor exports to introduce modularity
- ▶ 2 options
  - ▶ systematically during export (e.g., one theory for every HOL type definition)
  - ▶ heuristic or interactive MMT API-based refactoring
- ▶ Collect correspondences between concepts in different libraries (heuristically or interactively)
- ▶ Relate isomorphic theories across languages
- ▶ Use partial morphisms to translate libraries

- ▶ There is a wealth of formal mathematics out there (diversity?)
- ▶ Unfortunately, it is segregated into 20+ silos (need foundational pluralism)
- ▶ System-specific part of the exporter must be part of the exporting system

# Take-Home Message: OAF

- ▶ There is a wealth of formal mathematics out there (diversity?)
- ▶ Unfortunately, it is segregated into 20+ silos (need foundational pluralism)
- ▶ System-specific part of the exporter must be part of the exporting system
- ▶ integration of heterogeneous libraries necessary (DLAI?)

## 4 The OEIS as a Mathematical Resource

## 4.1 The OEIS: Online Encyclopedia of Integer Sequences

# OEIS: Open Encyclopedia of Integer Sequences

- ▶ **Definition 4.1** An **integer sequence** is a function  $s: \mathbb{N} \rightarrow \mathbb{Z}$ .
- ▶ **Applications:** Every parametric phenomenon that can be counted
- ▶ **Example 4.2** **A000944**: Number of polyhedra (or 3-connected simple planar graphs) with  $n$  nodes  $(0, 0, 0, 1, 2, 7, 34, 257, 2606, \dots)$
- ▶ **Example 4.3** **A001222**: Number of prime divisors of  $n$  counted with multiplicity  $(0, 1, 1, 2, 1, 2, 1, 3, 2, 2, 1, 3, 1, 2, 2, \dots)$
- ▶ **Example 4.4** **A031214**: First elements in all OEIS sequences (in order)  $(1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \dots)$
- ▶ **Intuition:** If phenomena grow with the same sequence  $\leadsto$  related?
- ▶ **Idea:** Collect many integer sequences (Neil Sloane 1965  $\leadsto$  OEIS)
  - ▶ started as a book: *A Handbook of Integer Sequences* 1973 (2372 sequences)
  - ▶ online since 1994 (16.000 sequences  $\leadsto$  <http://oeis.org>)
  - ▶ OEIS Foundation: 2009 (Creative Commons License)
  - ▶ today:  $\sim 275.000$  sequences

- ▶ One “record” per sequence with fields including
    - ▶ Identifier: A??????
    - ▶ start values
    - ▶ name (maybe with short explanation)
    - ▶ author
    - ▶ references to papers
    - ▶ program code
    - ▶ **Formulae**
- (DB Key)
- (Mathematica, Pari, ...)

All in ASCII files keyed by one-letter line prefixes.

▶ **Example 4.5 (Fibonacci Numbers)** %I A000045 M0692 N0256

%S A000045 0,1,1,2,3,5,8,13,21,34,55,89,144,233,377,610,987

%N A000045 Fibonacci numbers:  $F(n) = F(n-1) + F(n-2)$  with  $F(0) = 0$  and  $F(1) = 1$ .

%D A000045 V. E. Hoggatt, Jr., Fibonacci and Lucas Numbers. Houghton, Boston, MA, 1969.

%F A000045  $F(n) = ((1+\sqrt{5})^n - (1-\sqrt{5})^n) / (2^n \sqrt{5})$

%F A000045 G.f.:  $\sum_{n \geq 0} x^n * \text{Product}_{k=1..n} (k + x) / (1 + k*x)$ . — Paul D. Harris

%F A000045 This is a divisibility sequence; that is, if  $n$  divides  $m$ , then  $a(n)$  divides  $a(m)$

%A A000045 \_N. J. A. Sloane\_, Apr 30 1991

## 4.2 OEIS Semantification

# Parsing the OEIS format

- ▶ Formulae have no prescribed format (look good to the editors)
- ▶ But they are sufficiently regular (on average) to allow parsing
  - ▶ **infix operators**, e.g. the  $+$  symbol in  $m+n$ .
  - ▶ **suffix operators**, e.g. the  $!$  symbol in  $n!$ .
  - ▶ **prefix operators** (with or without brackets), e.g.  $\sin$  in  $\sin(x)$  or  $\sin x$ .
  - ▶ **infix relation symbols**, e.g. the  $<$  symbol in  $x < 2$ .
  - ▶ **binding operators**, e.g. the  $\forall$  symbol in  $\forall x. x^2 > 0$ .
- and some OEIS idioms like G.f. or g.f. for “generating function”.
- ▶ **Problem**: open-ended set of primitives, e.g.  $\sqrt{\phantom{x}}$ ,  $^{\phantom{x}}$ ,  $\text{sum}/\Sigma$  and  $\text{prod}/\Pi$
- ▶ **Ambiguity**: ASCII formulae have multiple plausible readings, e.g.
  - ▶ **implicit multiplication/application**:  $a(x+y)$  or  $\ln x$
  - ▶ **elided brackets/precedences**:  $\sin x$  or even  $\sin x/y$
- ▶ **Delineating Formulae/Text**:

Note that  $\text{ppzeta}(s) = \sum_{p \text{ prime}} 1/(p^s - 1)$  and  
 $\text{ppzeta}(s) = \sum_{k=1}^{\infty} \text{primezeta}(k*s)$ .

– Franklin T. Adams–Watters, Sep 11 2005.

# The Generated OMDoc

```
<omdoc xmlns:dc="http://purl.org/dc/elements/1.1/">
  <theory id="A000045">
    <metadata>
      <dc:creator>N. J. A. Sloane</dc:creator>
      <dc:title>Fibonacci numbers</dc:title>
    </metadata>
    <symbol name="seq"/>
    <assertion>
      <!-- OpenMath for  $\forall n. \text{seq}(n) = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n \sqrt{5}}$  -->
      <OMBIND>
        <OMS cd="arith" name="forall"/>
        <OMBVAR> <OMV name="n"/> </OMBVAR>
        <OMA>
          <OMS cd="arith" name="equal"/>
          <OMA><OMS name="seq"/><OMV name="n"></OMA>
          :
        </OMA>
      </OMBIND>
    </assertion>
  </theory>
</omdoc>
```

# Implementation

- ▶ Implementation as an extension of the MMT System (2000 LoC)
- ▶ Formula parsing via the Scala PackRat framework (left recursive linear parsing)
- ▶ available at <https://svn.kwarc.info/repos/MMT/src/mmt-oeis/>
- ▶ OEIS corpus:
  - ▶ 223.866 formula lines, The formula parser succeeds on 201384 (or 90%).
  - ▶ Out of that, 196515 (or 97.6%) contain mathematical expressions.
  - ▶ remaining problems: connectives, formula delineation
- ▶ What does the 90% mean?  $\leadsto$  parser accepts formula
- ▶ Manual Evaluation: 40 randomly selected parsed formulae evaluated  $\leadsto$  85% semantically correct
- ▶ Need to scale evaluation  $\leadsto$  involve OEIS editors (see below)

## 4.3 Applications

# Application: Math (Formula) Search

- ▶ We have a Math Search Engine: MathWebSearch
  - ▶ Harvest Formulae  $\leftrightarrow$  convert OpenMath to MathML
  - ▶ index them in MWS (together with full text).
  - ▶ formula converter daemon

(employ it)

(for user input)

**OEIS MathSearch**

The [OEIS MathSearch](#) system is a [text-and-math](#) search engine for the On-line Encyclopedia of Integer Sequences based on [MathWebSearch](#) and [MMT](#). To report an issue click [here](#). You can see existing issues [here](#).

$$\frac{(1+x)^n - (1-x)^n}{y}$$

« 1 2 3 »

MathHub.info : [A001045.omdoc](#)

MathHub.info : [A022103.omdoc](#)

<http://mathhub.info/oeis/oeis/source/A022/A022103.omdoc>  
**Title:** [A022103.omdoc](#)  
**OEIS Link:** <https://oeis.org/A022103>  
  
[...], math4, math5, G.f.: math6 : 
$$a(n) = \frac{(1+sqrt(5))^{n+1} - (1-sqrt(5))^{n+1}}{(2^{n+1} \times sqrt(5))} + \left( \frac{(5+1+sqrt(5)(n-1)) - (1-sqrt(5)(n-1))}{(2^{n+1} - 2) \times sqrt(5)} \right) \cdot \text{Offset math8, math9} \cdot A|$$
  
[Hakanson \(hawkuu\(AT\)gmail.com\[...\]\)](#)

MathHub.info : [A088138.omdoc](#)

- ▶ Demo: <http://oeissearch.mathweb.org/>

# Application: Standardizing Input in OEIS

- **Problem:** 400 OEIS submissions per week (three out of 60 editors really active)  
quality of submissions often low (including syntax)

# Application: Standardizing Input in OEIS

- **Problem:** 400 OEIS submissions per week (three out of 60 editors really active)  
quality of submissions often low (including syntax)
- **Idea:** Parse before submitting (use a normative parser)

[MEP](#) [Home](#) [Report issue](#)

**Editor** ^ >

**Text format:**  
OEIS

**Examples:**  
Number of trees with n unlabeled nodes.

%F A000055 G.f.:  $A(x) = 1 + T(x) - T^2(x)/2 + T(x^2)/2$ , where  $T(x) = x + x^2 + 2x^3 + \dots$  is the g.f. for A000081.

PRESENT ►

**Presenter** ^ >

- A000055  
G.f. ::  $A(x) = \left( (1 + T(x)) - \left( \frac{T(x)^2}{2} \right) + \left( \frac{T(x^2)}{2} \right) \right)$   
where  $T(x) = x + x^2 + (2 \times x^3) + \dots$  is the g.f. for A000081.

**Error log** ^

- **Warning** Unknown symbol "A"
- **Warning** Unknown symbol "T"
- **Warning** Unknown symbol "A000081"
- **Info** No stex:srcf attribute for <theory name="A000055">
- **Info** No stex:srcf attribute for <OMOBJ >

- **Demo:** <http://ash.eecs.jacobs-university.de:9090/>

## 4.4 Finding Relations between Sequences

# Relations between Sequences

- ▶ Understanding relations between sequences is a genuine mathematical concern.
- ▶ **State of the Art:** Matching initial segments of sequences.
- ▶ **Example 4.6** [Ste04] found 117 conjectures proves 100.
- ▶ **Problem:** Sampling limited data gives only conjectures. (need proof)
- ▶ **Example 4.7**  $\lfloor \frac{2n}{\log(2)} \rfloor$  and  $\lceil \frac{2}{2^{1/n}-1} \rceil$  agree for 777451915729367 terms but are not equal [Slo12].
- ▶ **Idea:** use the formulae from the OEIS instead.
  - ▶ they are exact and peer-reviewed (relations found will be theorems)
  - ▶ we have about 50k generating functions (powerful, compact, structured representations)

# Generating Functions for Integer Sequences

- ▶ **Definition 4.8** Let  $s := (a_n)_{n \in \mathbb{N}}$  be an integer sequence, then we call  $g_s(x) := \sum_{i=0}^{\infty} a_i x^i$  the **ordinary generating function** of  $s$ .
- ▶ **Example 4.9** The sequence  $A000012 = 1, 1, 1, 1, 1, \dots$  can be represented as  $1 + x + x^2 + \dots = \frac{1}{1-x}$
- ▶ represent an infinite sequence finitely (cf. Kolmogorov complexity)
- ▶ There are other generating functions: exponential generating functions, Lambert series, Bell series, and Dirichlet series. (use only ordinary ones for now)
- ▶ **Operations on Generating Functions:** induce to operations on the sequences.
  - ▶ constant factor:  $c \cdot g_s = g_{c \cdot s}$ .
  - ▶ shift:  $x^n \cdot g_s(x) = g_s(x + n)$ .
  - ▶ ... partial fraction decomposition, differentiation, integration, ...
- ▶ **Idea:** systematically search for relations on the generating functions in the OEIS induced by such operations

# Relation Finding Experiment

- ▶ **Experiment:** search for relations on  $\sim 50\,000$  OEIS generating functions
  - ▶ Method 1: const, shift, sort (sanity check; expect known relations)
  - ▶ Method 2: ... partial fraction decomposition, differentiation, integration, ...
  - ▶ Method 3: See Enxhell's B.Sc thesis [Luz16]
- Implementation:** import parsed equations into MMT, normalize/transform by Sage, hash, compare.

# Relation Finding Experiment

- ▶ **Experiment:** search for relations on  $\sim 50\,000$  OEIS generating functions
  - ▶ Method 1: const, shift, sort (sanity check; expect known relations)
  - ▶ Method 2: ... partial fraction decomposition, differentiation, integration, ...
  - ▶ Method 3: See Enxhell's B.Sc thesis [Luz16]

**Implementation:** import parsed equations into MMT, normalize/transform by Sage, hash, compare.

- ▶ **Example 4.13 (from Method 1)**  $A001478(n) = -A000027(n)$ . ( $\pm \text{Id}$  on  $\mathbb{N}$ )
- ▶ **Example 4.14 (from Method 2)** accepted in <https://oeis.org/A001787>:

$$A001787(n) = \frac{n}{6} A007283(n)$$

- ▶ **Example 4.15 (from Method 2)** accepted in <https://oeis.org/A037532>:

$$A037532(n) = \frac{5}{57} A049347(n-1) + \frac{3}{57} A049347(n) + \frac{29}{171} A000420(n) - \frac{2}{9}$$

# Relation Finding Experiment

- ▶ **Experiment:** search for relations on  $\sim 50\,000$  OEIS generating functions
  - ▶ Method 1: const, shift, sort (sanity check; expect known relations)
  - ▶ Method 2: ... partial fraction decomposition, differentiation, integration, ...
  - ▶ Method 3: See Enxhell's B.Sc thesis [Luz16]

**Implementation:** import parsed equations into MMT, normalize/transform by Sage, hash, compare.

- ▶ **Example 4.16 (from Method 1)**  $A001478(n) = -A000027(n)$ . ( $\pm \text{Id on } \mathbb{N}$ )
- ▶ **Example 4.17 (from Method 2)** accepted in <https://oeis.org/A001787>:

$$A001787(n) = \frac{n}{6} A007283(n)$$

- ▶ **Example 4.18 (from Method 2)** accepted in <https://oeis.org/A037532>:

$$A037532(n) = \frac{5}{57} A049347(n-1) + \frac{3}{57} A049347(n) + \frac{29}{171} A000420(n) - \frac{2}{9}$$

- ▶ two out of three randomly picked OEIS submissions were accepted by Neil Sloane (third one not deemed to be interesting enough)
- ▶ OEIS acceptance prompted immediate human submission of trivial corollaries

# Relation Finding Experiment: Overall Results

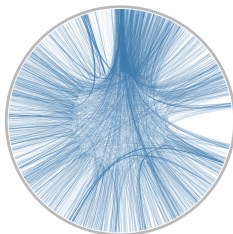
- **Results:** before recent parser enhancements.

Parsed Generating Functions	43 005
SageMath verified Generating Functions	16 065
Parsed Ordinary Generating Functions	35 953
SageMath verified Ordinary Generating Functions	13 400
Method 1 relations	4 859
Sequences in Method 1 relations	853
Method 2 relations	297 284 646
Method 2 relations without normalization	66 427

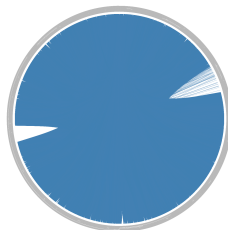
**Caveat:**  $G = A + B + C$  is counted  $4^3$  times (trivial variants)

- **Results:**

current realations



one B.Sc.



## Take-Home Message: OEIS

- ▶ OEIS is an interesting corpus (mostly Data  $\hat{=}$  facts about special individuals)
- ▶ OEIS grows steadily (6000 submitters, 300sub/week, 150 accepted by human editors))
- ▶ It is definitely not formal (but the GF are after parsing)
- ▶ induced GF database allows deriving new theorems

## Take-Home Message: OEIS

- ▶ OEIS is an interesting corpus (mostly Data  $\hat{=}$  facts about special individuals)
- ▶ OEIS grows steadily (6000 submitters, 300sub/week, 150 accepted by human editors))
- ▶ It is definitely not formal (but the GF are after parsing)
- ▶ induced GF database allows deriving new theorems
- ▶ Need a Theorem Appreciator for automated submission (DLAI?)

## Take-Home Message again (If I managed to get here)

- ▶ I only go GOF AI (Good Old-fashioned AI aka. Logic)
- ▶ My Domain of Application is Math (no e.g. protocol verification)
- ▶ no DLAI (applying Deep Learning to everything)
- ▶ BUT we have a lot of interesting Data
  - ▶ OAF: the Open Archive of Formalizations (<http://oaf.mathhub.info>)
  - ▶ arXMLiv preprints and ZBMath Abstracts (licensing problems)
  - ▶ OEIS: "Conjecturing relations between Sequences" ([https://github.com/eluzhnica/\\*](https://github.com/eluzhnica/*))
- ▶ Could use DLAI help (but not in ATP improvements)
- ▶ I am looking for good GOF AI Ph.D. students (maybe even DLAI)



Ingrid Daubechies, Clifford A. Lynch, Kathleen M. Carley, Timothy W. Cole, Judith L. Klavans, Yann LeCun, Michael Lesk, Peter Olver, Jim Pitman, and Zhihong Xia.

*Developing a 21st Century Global Library for Mathematics Research.*  
THE NATIONAL ACADEMIES PRESS, 2014.



Arif Jinha.

Article 50 million: an estimate of the number of scholarly articles in existence.  
*Learned Publishing*, 23(3):258–263, 2010.



Enxhell Luzhnica.

Formula semantification and automated relation finding in the OEIS.  
B.sc. thesis, Jacobs University Bremen, 2016.



Peder Olesen Larsen and Markus von Ins.

The rate of growth in scientific publication and the decline in coverage provided by science citation index.  
*Scientometrics*, 84(3):575–603, 2010.



F. Rabe and C. Schürmann.

A Practical Module System for LF.

In J. Cheney and A. Felty, editors, *Proceedings of the Workshop on Logical Frameworks: Meta-Theory and Practice (LFMTP)*, volume LFMTP'09 of



Neil J. A. Sloane.

The on-line encyclopedia of integer sequences.

<http://neilsloane.com/doc/eger.pdf>, 2012.



Ralf Stephan.

State of 100 conjectures from the oeis.

<http://www.ark.in-berlin.de/conj.txt>, 2004.