# Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2015)
# European Conference on Data Analysis

University of Essex, 2015-09-03

Michael Kohlhase (Jacobs-University Bremen)
Wolfram Sperber (FIZ Karlsruhe)

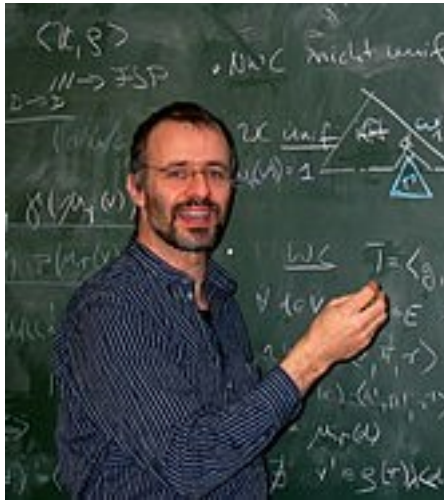
FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

# Agenda

- The background
- Glossaries: Terminology
- The Data Model of SMGloM

  The modules
  - Module signature
  - Language Bindings
- The SMGloM prototype
- The future of SMGloM

# The (personal) background



Michael's background:
Knowledge management in mathematics
(flexible) formalization of mathematics
(Module System for Mathematical Theories language, shortly MMT language)

My background:
zbMATH database, previously Zentralblatt für Mathematik und ihre Grenzgebiete
Machine-based content analysis of mathematical publications:
more in detail: reviews/abstracts – key phrases – classification (MSC)

# The background: The DeLiVerMATH project

Two of our projects in the last years:
**DeLiVerMATH** project:
Aims:

- keyword extraction
- controlled vocabulary (list of relevant key phrases, weighting by frequencies)
- classification (MSC)

Methods

Machine-based methods for key phrase extraction and classification (using reviews and abstracts):

- use of NLP methods (POS tagging - noun phrase extraction)
- SVM methods for classification

Open question:

Relevance of noun phrases

.

# The background: The MathSearch project

**MathSearch** project

- Development of methods for mathematical formula search
- Common mathematical language: Duality of text and formulae
- Formulae – a product of historical development
  'Mathematical alphabet': symbols for objects, operations, and relations but the comparison is not really possible, mathematical has a lot of characters, symbols are not only characters but also complex words
- Advantage of formulae:  highly condensed presentation of complex mathematical concepts, objects, and statements
- Moreover, the presentation is 2D, a lot of different symbols
  one of the first Markup languages:
  - TeX, also today the smart  input standard for publishing mathematics
    disadvantage of TeX: too little semantics

# The background: mathematical languages

- solution: other XML languages
  most prominent:
  MathML (XML language for mathematics, a W3C recommendation):
  - Presentation MathML (TeX can be converted automatically to Presentation MathML)
  - Content MathML (semantic enriched MathML, requires semantically enriched TeX: STeX)

Other semantic mathematical languages:

- OpenMATH, (origin CAS, compatible with Conten MathML)
- OMDoc (metastructure of mathematical documents)

# The background: Formula search

Concept for the formula search:

- Symbols and formulae encoded in TeX are converted to Content MathML (with the LaTeXML converter)
  but this is - in general - not unique
  Why? There are the same problems for symbols and formlae as with words and phrases:
  - symbols and formulae can have different meanings

  and
  - different symbols and formulae are used for concepts, objects and statements

Open problem:

disambiguation of the mathematical symbols and formulae

# The background:
# The concept for formula search (I)

Concept for the formula search:

- Symbols and formulae encoded in TeX are converted to MathML (with the LaTeXML converter, conversion to Presentation plus Content MathML)

  but the conversion to MathML is - in general - not unique
  Why? There are the same problems for symbols and formulae as with words and phrases:
  - symbols and formulae can have different meanings

  and
  - different symbols and formulae are used for concepts, objects and statements

# The background:
# The concept for formula search (II)

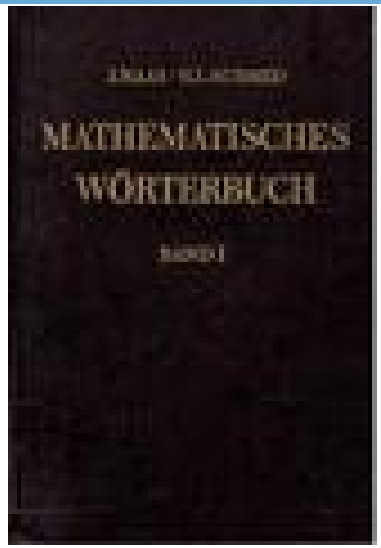
- Hence, the Content MathML encoding is not unique
- The symbols and formulae must be disambigued,
- This can be done by context analysis

For both, the extraction  of key phrases (and also classification) and formula search, a semantic glossary as a knowledge base  are useful. Such a semantic glossary must involve also symbols and formulae

# The background: Some general remarks about the mathematical language

- The mathematical language is a natural language.
- The 'common' mathematical language is a mix of common language and (mathematical) formulae.
- Mathematics is term-driven: Mathematical terms which describe the mathematical objects or concepts play a essential role.
- Mathematical terms are given by definitions.
- The set of terms are well-structured.
- Mathematics needs formulae: Formulae allow arbitrary complex presentations of mathematical objects and statements (it is very difficult, to express a complex formula in common language).

# Mathematical glossaries




- Mathematical glossaries are an important source for the usage of mathematics mathematical and mathematical knowledge management.
- Mathematical encylopedias (and dictionaries) have a long tradition (iMathematische Wörterbuch, Encyclopedia of Mathematics)
- In the Internet:
  Wikipedia (more than 20,000 terms),
  Encyclopedia of Mathematics,
  Planet Math, ...
  but they have two deficits:
  - semantic relations are not presented in a systematic way
  - symbols and formulaes are not semantified (symbols and formulae are encoded in TeX, not in STeX

→ development of a new data model for SMGloM

# Glossaries: Our terminology

- **Glossaries (~ encyclopedias)** (in the traditional meaning):
  lists of terms with short definitions which are ordered alphabetically
  (not only lists of key phrases)
- **Term**
  word or phrases that have a specific meaning describing objects or concepts
- New quality: **Semantic Glossaries**
  adding sytematically structural (terminological and domain) relations to the terms
  (semantic glossaries are ontologies)

# Glossaries: Our terminology

More in detail

- **Terminological relations**
  ***S**emantic relations* between terms, e.g., the Wordnet relations
  - synonymy
  - hypernymy and hyponymy
  - meronymy and holonymy
  - antinomy

  Remark: These relations cover only some selected relations.
  The relations must be extended for the needs of mathematics.

- **Domain relations**
  *Domain relations* describing for example domain-specific methods (e.g., number theory) → classification

# SMGloM: The data model

The SMGloM data model consists of **modules**:
A module is corresponding to a definition of a concept.

Remark: Of course, a mathematical term can be defined on different ways. But, in SMGloM a module for each definition is introduced. The equivalence between two modules is modelled by views. Views can be defined in a flexible way.

# SMGloM: The data model

A module consists of the
- **module signature**
  and
- **language bindings**.

The **module signature** contains the *language independent* part:

- the identifier of the module,
- the relations to other SMGloM modules which are necessary for the definition of the module
  (e.g. integers are used for the definition of prime numbers)
- a semantified input of the symbols and formulae
  by assigning a name to a symbols
  (e.g., the 'plus' operation in $\mathbb{R}^n$)
- alternative symbols

# Module Signature: An example

[the module]:

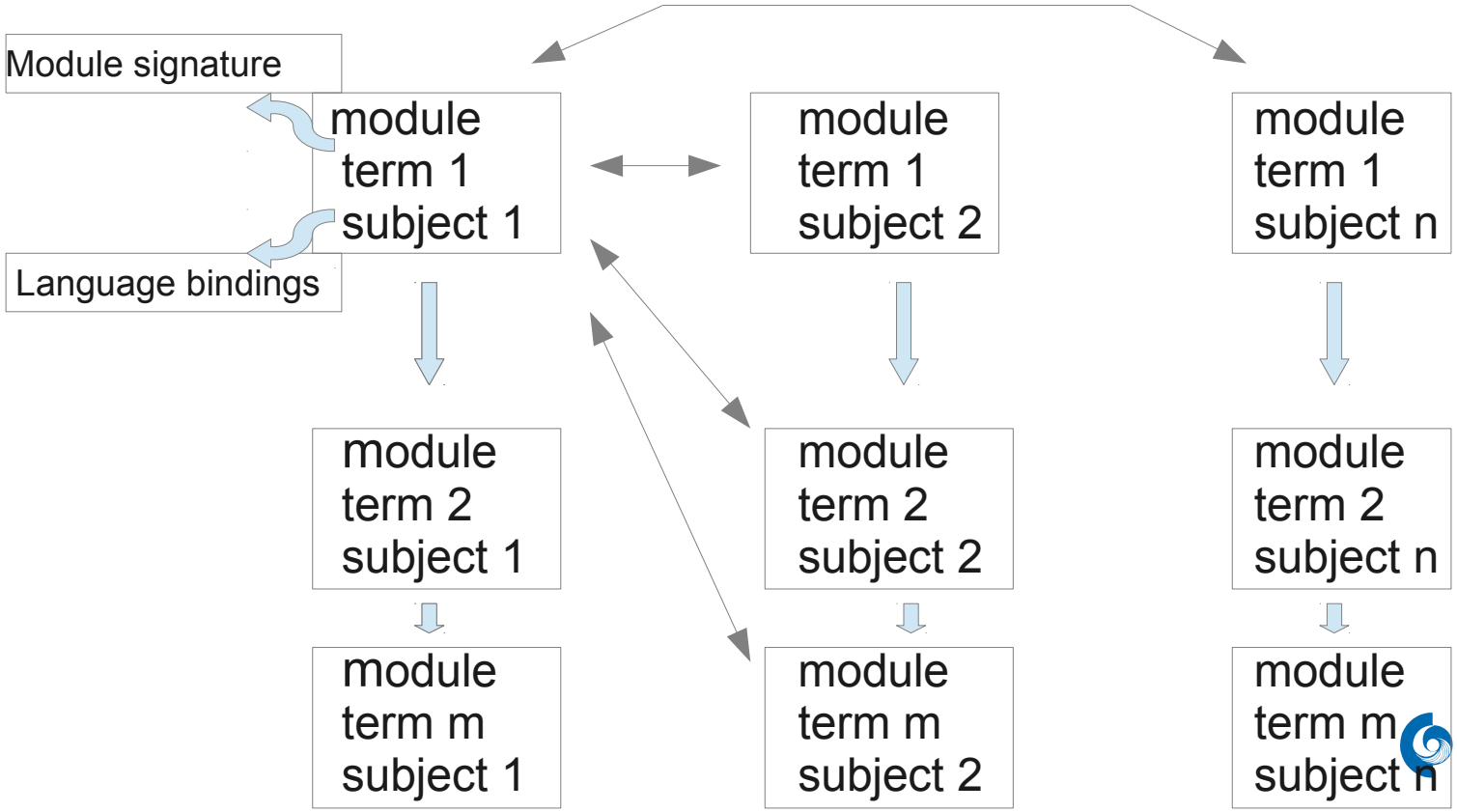
```
\begin{modsig}[creators=jusche]{primenumber}
\gimport[smglom/numberfields]{arithmetics}
\gimport[smglom/numberfields]{divisor}

\symdef[name=prime-number]{PrimeNumber}{p}
\symdef[name=prime-number]{nPrimeNumber}[1]{p_{#1}}
\symtest{nPrimeNumber}{\nPrimeNumber{n}}

\symdef[name=NumberPrimeNumber]{NumberPrimeNumberOp}{\pi}
\symdef{NumberPrimeNumber}[1]{\prefix\NumberPrimeNumberOp{#1}}
\symtest{NumberPrimeNumber}{\NumberPrimeNumber{20}=8}
\end{modsig}
```

# The data model of SMGloM

The **language bindings** (one language bindung for a common language) contain the definition and the name(s) of the term.




FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

# Language Bindings (English): An example

[the language binding for the English language]:

```
\begin{modnl}[creators=jusche]{primenumber}{en}
\begin{definition}
A \defii{prime}{number} is a \trefii[naturalnumbers]{natural}{number}
greater than $1$ that has no positive
\trefi[divisor]{divisor}s other than $1$ and itself.
\end{definition}

\begin{definition}
The number of \trefii{prime}{number}s not greater than
$n$ is written as $\NumberPrimeNumber{n}$.
\end{definition}
\end{modnl}
```

# Language binding (German): An example

[the language binding for the German language]:

```
\begin{modnl}[creators=jusche]{primenumber}{de}
\begin{definition}
Eine \defi[prime-number]{Primzahl} ist eine
\mtrefii[naturalnumbers?natural-number]nat"urlicheZahl,
die genau zwei nat"urliche Zahlen als
[divisor?divisor]Teiler hat. Die Anzahl der
Primzahlen kleiner oder gleich $n$ bezeichnet man
mit $\NumberPrimeNumber{n}$.
\end{definition}
\end{modnl}
```

# SMGloM: Proof of concept - the prototype

Up to now, SMGloM covers nearly 500 modules and 1,500 language bindings (English, German, Romanian, Chinese, ...)
with the focus 'number theory' (one of the top classes of the MSC) beside elementary mathematics.

The modules were created manually.
Our experience:
It is time-expensive to create modules  for some reasons:

- definition (therefore thrustworthy resources are necessary)
- relations and embedding in the graph structure
- encoding and syntactic correctness

First tools were developed to check the structure and the syntax of the entries.  More tools to support the input and control the formal correctness are under development.

# SMGloM: An example

- Prime factorization Definition, Concept Graph de
- prime gap Definition, Notations, Concept Graph de
- prime number Definition, Notations, Concept Graph de

| Languages | Arguments | Rendering |
|---|---|---|
| mathml | 1 | $p_a$ |
| mathml | 0 | $p$ |

A **prime number** is a natural number greater than 1 that has no positive divisor s other than 1 and itself.
The number of prime number s not greater than $n$ is written as $\pi\ (n)$.

https://mathhub.info/mh/glossary

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

# Graph representation of semantic realations in SMGloM (SVG)

MathHub Home Contribute Libraries Sources Search Glossary Math Dictionary

## Navigation

Forums

Help

Report issue

## User login

**Username** *

**Password** *

- Create new account
- Request new password

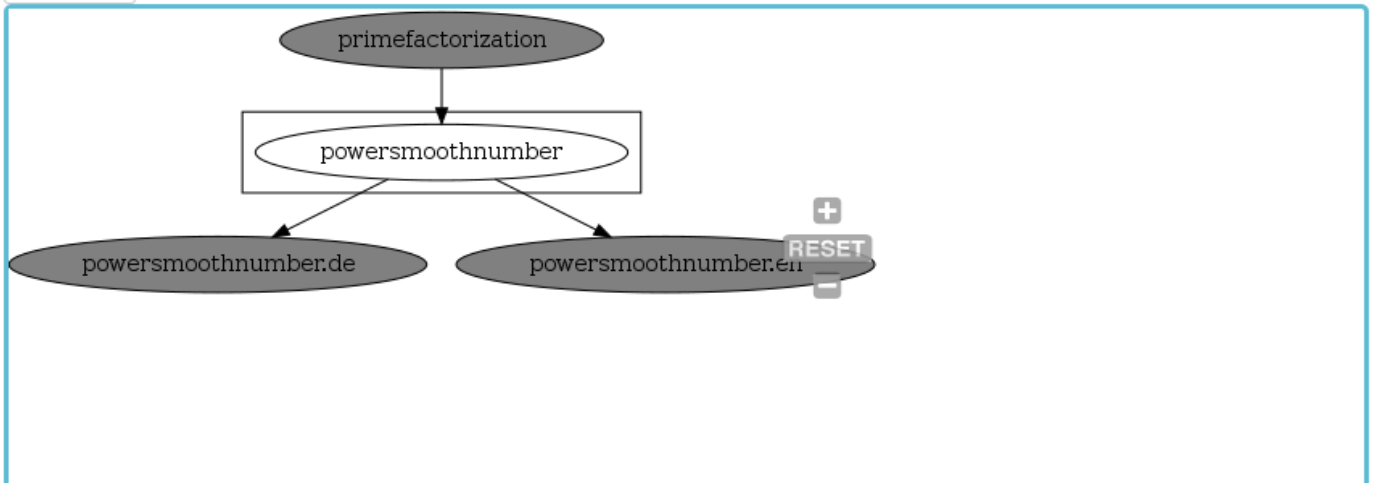
Log in

Libraries / smglom / numbers / powersmoothnumber / SVG

## powersmoothnumber

View OMDoc Source SVG

Full screen




FIZ Karlsruhe

Leibniz-Institut für Informationsinfrastruktur

# SMGloM: Proof of concept – more remarks

SMGloM is compatible with the more general MMT concept.

The data model of SMGloM is under development, especially the relations between the modules, the 'views' must be defined in a more precise way. Views must be able to define relations between modules in a flexible way (e.g., synonyms, examples, generalizations etc.)

# SMGloM and applications

SMGloM can be used for a wide set of applications

- terminological base of mathematics and fast and structured access to the mathematical terminology for human users
- dictionary for human and machine-based translation
- graph presentations of ontologies of mathematical subjects
- retrieval especially formula search
- semantic enrichment of mathematical publications
- content analysis, e.g. classification or clustering, in digital mathematical libraries

# SMGloM: The prototype

The aim is a high-quality glossary of the whole mathematics which provides useful information for both
- humans and
- machine processing of information

Today, no mathematician has the expertise to create such a glossary for all mathematical subjects.
It is clear that such a glossary cannot be produced by a single expert. It must be a cooperative activity. Hence, SMGloM is planned as a community-based initiative.

→ an open license for SMGloM

# SMGloM: Further remarks

How it could work?

- quality control by lenses (subject and topic-specific evaluation groups)
- syntax control by automatic means
- automatic tools to create candidate lists for SMGloM, to extract definitions etc.
  (therefore a new project proposal at the moment is under work)

Thanks!