The Future World Heritage Digital Mathematics Library: Plans and Prospects

Symposium organized by the Committee on Electronic Information and Communication (CEIC) of the International Mathematical Union (IMU), National Academy of Sciences, Washington DC, June 1-3, 2012

On Access Infrastructures to Digital Libraries

Gert-Martin Greuel

University of Kaiserslautern, Mathematisches Forschungsinstitut Oberwolfach, Zentralblatt MATH (ZBMATH)

On Access Infrastructures to Digital Libraries

Overview

1. What do we need for a WDML?

2. The WDML from the perspective of a local visitor center

3. Reference databases as a potential access infrastructure to the WDML

1. What do we need for a WDML?

In 2002 **John Ewing** (Notices of the AMS) noted **three goals**:

1. digitize a preponderance of scholarly mathematical literature that is not already in digital form

2. set technical standards for making digital mathematical literature accessible online

3. negotiate a protocol for making future digital mathematical literature available in the future

He mentioned **four major problems**:

1. **Content** (deciding what has to be included and what not)
2. **Copyright** (clearing complicated legal issues in international copyright)
3. **Initial Format** (technical format for presentation)
4. **Archiving** (technical format for archiving and a model for financial maintenance)

In 2006 the **International Mathematical Union (IMU)** and the **Committee on Electronic Information and Communication (CEIC)** of the IMU formulated

Digital Mathematics Library: A Vision for the Future

They required that each article (or item) in a digitization project should include **four components:**

1. Accurate **metadata** consistent with agreed upon **standards**

2. A separate list of references (when available) with **links to the indexing databases Mathematical Reviews** and **Zentralblatt Math**

3. A **high-quality scanned image** of each page

4. The text derived from **optical character recognition** (which is normally hidden from the reader, but keyed to the image for searching)

While points 3 and 4 are nowadays more or less standard and provided by public or commercial digitizing enterprises, work on points 1 and 2 has not even started.

Today we see several changes

- A big amount of mathematical literature **has been digitized** by a variety of different commercial and non-commercial providers
- As a consequence **there exist different formats** for data and metadata of a large number of digital ibraries organized by different providers under different access conditions

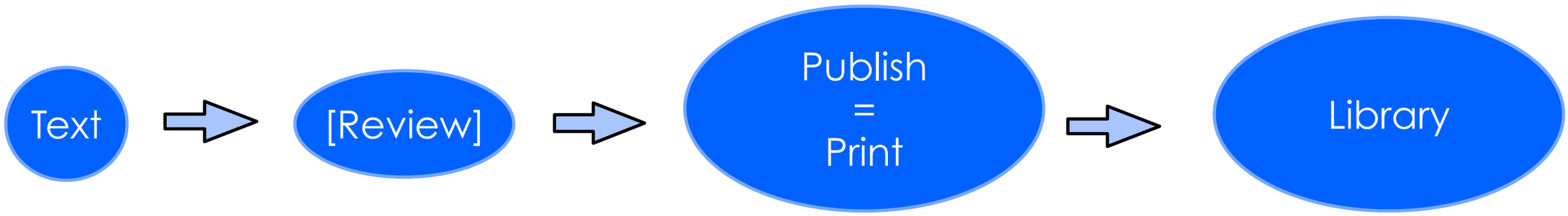
Problems concerning Formats and Content Analysis

- The existing digital mathematics libraries use different formats for
  - publishing and storing documents
  - content analysis and metadata
- Formats and metadata are under permanent development
- We need better content analysis to realize a search to knowledge and not only a search for publications
- We need (semi-)automatic tools for the content analysis

A central digitizing and archiving organization seems unrealistic. The very much differing formats and level of accessibility are, however, a major problem.

Also Publishing has changed dramatically:

classical:



current technologies:

search, semantics, archiving,
only default rendering

Business competition for
rendering the library stuff

- **Public funds should only be given to the digital library** which has to provide full open access
- Rendering, creating portals etc., could be an **opportunity for business**, combining it for different purposes and different categories of users

The model I propose assumes that the library itself isopen access, offering APIs but only default rendering. This offers e.g. commercial publishers opportunities to provide extra functionality, link it with their own services, etc..

We have not yet a comprehensive digital mathematics library

Consequences

- It seems that a decentralized form is the natural form of the WDML
- We have to develop new models
  - for the organization of the WDML; the different digital mathematics libraries must cooperate
  - for the access to the WDML; the content of the WDML as a whole must be searchable for the user from a unique point of access

A good access infrastructure for the WDML has become a major challenge

By access infrastructure for the WDML I mean keywords, classification, searching, semantic content analysis, ... at a unique point of access

The EuDML could serve as a prototype, which we could use but from which we certainly should learn.

How to achieve a World Digital Mathematics Library?

- We need an **open technical framework**, at least providing **a scheme for metadata and content analysis** in linked local DMLs

- We need to include digital content **from open access literature and from commercial publishers** to the DML under well defined conditions

- A WDML will need **public funding,** business models based on donations or on advertisements will not work

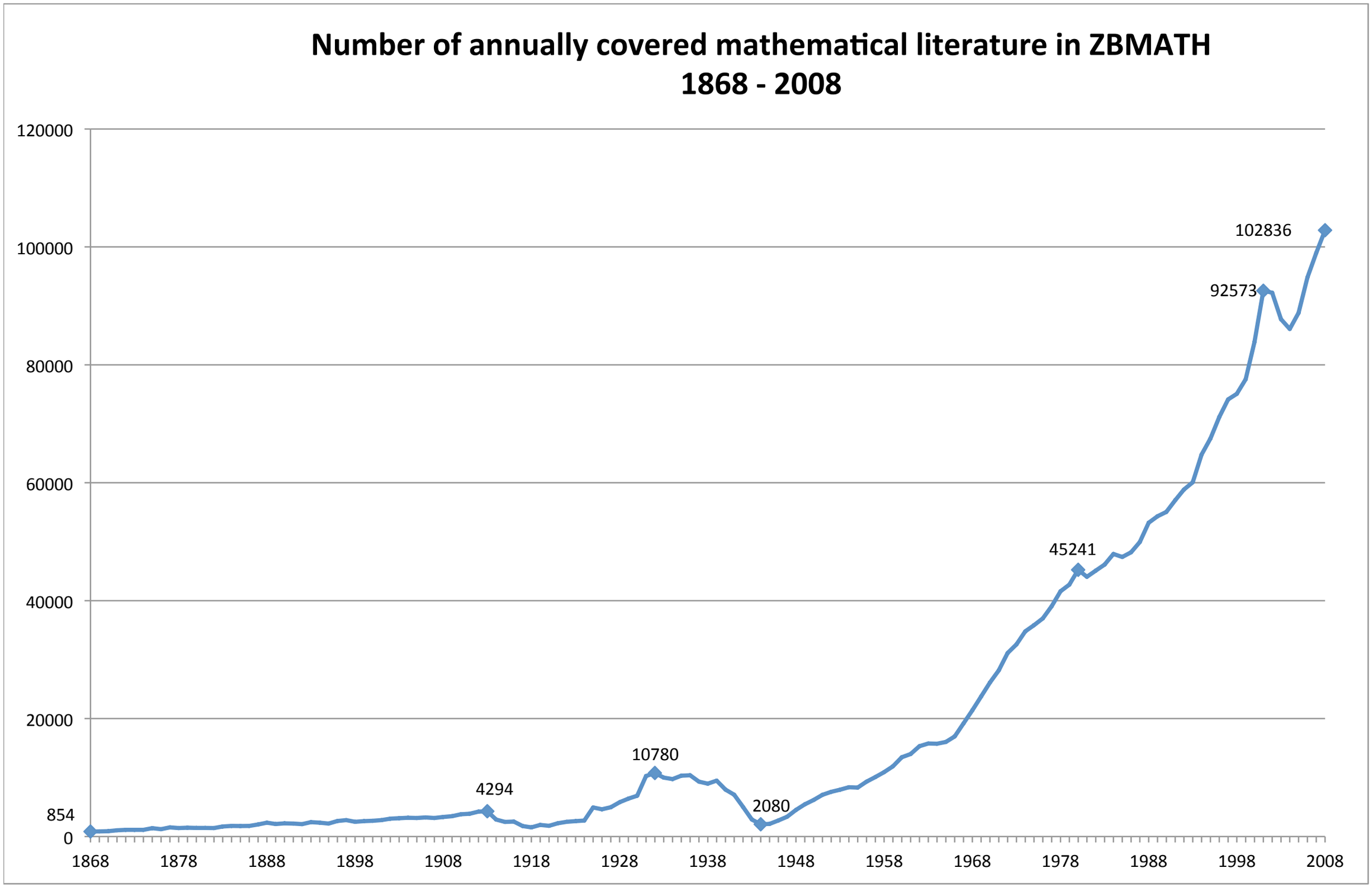
- In connection with public funding we should promote the idea of **clean open access**

- We need a **non-commercial WDML consortium,** controlled by the mathematical community (IMU), and **committees for different tasks** (technical standards, access structures and meta-data, integration of different resources, funding, etc.)

- We should agree on **milestones** and a **time schedule**

To a large extent I agree with the position statements of Thierry Bouche and Jiri Rakosnik.

Growth of the mathematical literature



The enormous growth of the mathematical literature is of course also a challenge for a WDML.

Even very old articles are still cited

	2000-2009	1990-1999	1980-1989	1970-1979	1960-1969	1950-1959
2000-2009	514740	518498	213640	118891	61508	26300
1990-1999	19	223020	349027	156424	74091	27865
1980-1989	0	66	166912	208979	87496	30224
1970-1979	0	0	47	114793	125224	37017
1960-1969	0	0	0	18	43386	34057
1950-1959	0	0	0	0	20	13818

e.g. 665 references in articles published in 2000-2009 refer to articles published in 1850-1859

	1900-1909	1890-1899	1880-1889	1870-1879	1860-1869	1850-1859
2000-2009	2264	1602	1203	855	628	665
1990-1999	1712	1246	963	676	406	362
1980-1989	1364	1004	791	556	323	291
1970-1979	1212	902	608	362	186	157
1960-1969	950	475	286	159	103	62
1950-1959	667	335	190	98	43	39
1940-1949	270	130	76	44	18	18
1930-1939	781	407	227	113	48	14
1920-1929	733	365	229	138	48	24

All data are taken from ZBMATH.

2. The WDML from the perspective of a local visitor center

- The **Mathematisches Forschungsinstitut Oberwolfach (MFO)** is one of the leading international research centers
- The Institute concentrates on cooperative research activities such as one-week workshops or longer stays of small research groups
- Leading representatives of particularly relevant research areas from all over the world are invited to Oberwolfach (about 70% coming from abroad)
- In all activities, participation of promising young scientists plays an important role

Oberwolfach as a potential user of the WDML

- Oberwolfach develops a library portal for the local visitors at the MFO
- The WDML would significantly enhance the portal and improve the research conditions at the MFO
- Oberwolfach is strongly interested in the WDML

Besides individuals, also research institutes are potential users of the WDML. They may wish to create specialized services by using the WDML.

Oberwolfach as a potential provider of content for the WDML

Content created by the MFO

- Oberwolfach Reports (OWR)
- Oberwolfach Preprints (OWP)
- Oberwolfach Digital Archive (ODA)
- Oberwolfach Photo Data Base (OPDB)
- Oberwolfach References on Mathematical Software (ORMS)

All content is open access

MFO is prepared to provide its publications to the WDML

Like Oberwolfach, other research institutes may wish to become a provider of content for the WDML.

3. Reference databases as potential access infrastructure to the WDML

Why should reference databases like ZBMATH and MathSciNet be used as an access infrastructure?

Recall the IMU/CEIC requirement:
Each article should include a separate list of references with links to the indexing databases Mathematical Reviews and Zentralblatt Math

Reference databases have several advantages:

- Provide identifiers for the indexed mathematical literature
- completeness of mathematical literature
- high quality metadata
- well-structured metadata
- qualified search options (e.g., field search)
- exclusive to mathematical literature, little noise
- semantic content analysis (MSC, keywords, abstract, reviews)
- additional feedback from the community (reviews)
- reference lists
- linking of information (e.g., with full texts, if available)
- author disambiguation, author profiles

Reference databases are engaged in the development of tools for the WDML

- development of metadata schemes for mathematical publications (adding of new metadata, e.g. references)

- maintenance of the Mathematical Subject Classification (MSC). A permanent task, soon: the transformation of the MSC to semantic web technologies (SKOS)

- pilot partner for the use of the methods for publishing and presenting mathematical knowledge (e.g., use of MathML as presentation format)

- development of new methods of content analysis (semantic tools)

- predestinated for the linking of different data, especially linking metadata and full texts (in EuDML project, 80.000 links have been created between EuDML items by using ZBMATH)

Hence reference data bases can provide **core services** for the WDML

The Reviews and Zentralblatt work together in developing MSC and SKOS. It is desirable that they also work together in providing core services for the WDML.

Remarks for Content Analysis in reviewing journals

The two big reviewing journals in mathematics:

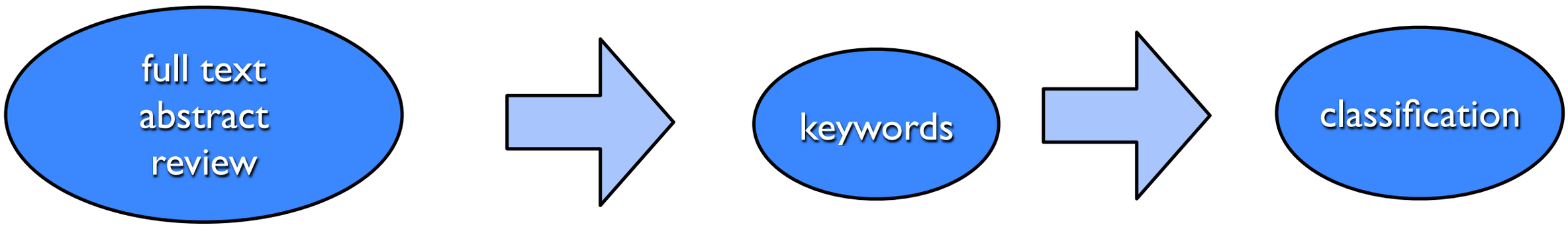
- Mathematical Reviews (complete from1940)
- Zentralblatt MATH (complete from1868)

have a great experience with content analysis of mathematical publications

Current Elements of Content Analysis

Abstract
Review
Keywords
Classification (MSC)
Further helpful information: author identification, citation analysis

There are close semantic relationships between the elements



All these elements are of different nature and have its own value. E.g., abstracs and reviews are short summaries of the content, keywords and classification are important for embeding a publication in the scientific canon.

Keywords - Controlled vocabulary

- Today, keyword search is a normal way to find relevant information (cf. Google)
- A qualified keyword analysis would be helpful for retrieval and also for (automatic) classification and clustering
- A controlled vocabulary could be a useful tool for keyword analysis.
- A typical method for keyword analysis is keyword extraction based on full texts, abstracts and reviews
- This can be done manually or automatically

Keyphrases

Two main methods to create a controlled vocabulary:

- linguistic methods
- statistical methods (the methods can be combined)

A first linguistic attempt in ZBMATH

- building up of special dictionaries (e.g. MSC labels, names of mathematicians, synonyms, acronyms, ...)
- definition of typical patterns
- looking for these patterns within the ZBMATH data (frequency – keyphrase)

More about content analysis and semantics tools can be said by Michael Kohlhase.

Example: The most frequent key phrases (of the length 4) for the MSC classes 13 and 14, based on ZBMATH data 2005 - 2011)
Typically, the number of keyphrases for each MSC class is huge (>> 10.000)

MSC 13 (4 word groups)

==========

332 principal polarized abelian variety
187 smooth complex projective variety
99 complete discrete valuation ring
58 connected reductive algebraic group
49 smooth complex projective surface
47 smooth complex projective curve
41 finite dimensional vector space
35 connected linear algebraic group
34 principal polarized abelian surface
33 algebraic closed residue field
33 simple normal crossing divisor
32 complete discrete valuation field
32 irreducible holomorphic symplectic manifold
32 nonsingular real algebraic variety
30 reductive complex algebraic group
....

24 finite generated abelian group
24 large complex structure limit
23 ha only rational singularity
23 isolated complete intersection singularity
21 completely integrable hamiltonian system
21 henselian discrete valuation ring
20 absolute simple abelian variety
20 differential graded lie algebra
19 algebraic closed ground field
19 minimal graded free resolution
19 smooth connected projective curve
19 smooth projective algebraic curve
19 special lagrangian torus fibration
18 only rational double point
17 affine real algebraic variety
17 ha only canonical singularity
17 irreducible smooth projective curve
....

We see that the extracted keyphrases must be checked manually. The checked keyphrases define a first controlled vocabulary.

Outlook

- Up to now, the content analysis is targeted to the content analysis of a publication as a whole
- But for the future, we need new filters to search for the relevant information, allowing a detailed search in the publications
- We need better content analysis methods for details. The normal full text search is not enough
- A better content analysis is the way for an enhanced WDML ("Semantic WDML")

As a WDML does not yet exist, the **EuDML may be considerd as protoptype**:

- A consortium of 12 content and service-providers
- Currently a total of 230,000 items (articles, books)
- During this project, 80,000 digital links have been created between the EuDML items, relying on ZBMATH services and data

More about the EuDML project will be said by Thierry Bouche.

Summary

General

- We should agree on standards for metadata and content analysis
- We need digital content from open access literature and from commercial publishers
- The business model should be based on public funding
- We should promote the idea of "clean" open access
- We need a WDML consortium and committees for different tasks

Access stucture

- As required by the IMU/CEIC, all articles and references in the WDML should be linked to MathSciNet and ZBMATH
- An easy access to the WDML is essential for the acceptance in the community
- Therefore, new methods for an efficient (automatic) content analysis must be used
- ZBMATH and MathSciNet work on such automatic methods; we should agree on standards which allow easy integration of both services
- The reviewing services could play an essential role as provider of core services for the WDML

Clean OA is a green OA + high-quality/peer-reviewed + guaranteed access at the library and "clean" of profit concerns.