Some facets of knowledge management in mathematics

Wolfram Sperber (Zentralblatt Math)
Patrick Ion (Math Reviews)

Facets of Knowledge Organization
A tribute to Professor Brian Vickery
ISKO UK biennal conference
4th-5th July 2011
London

Agenda

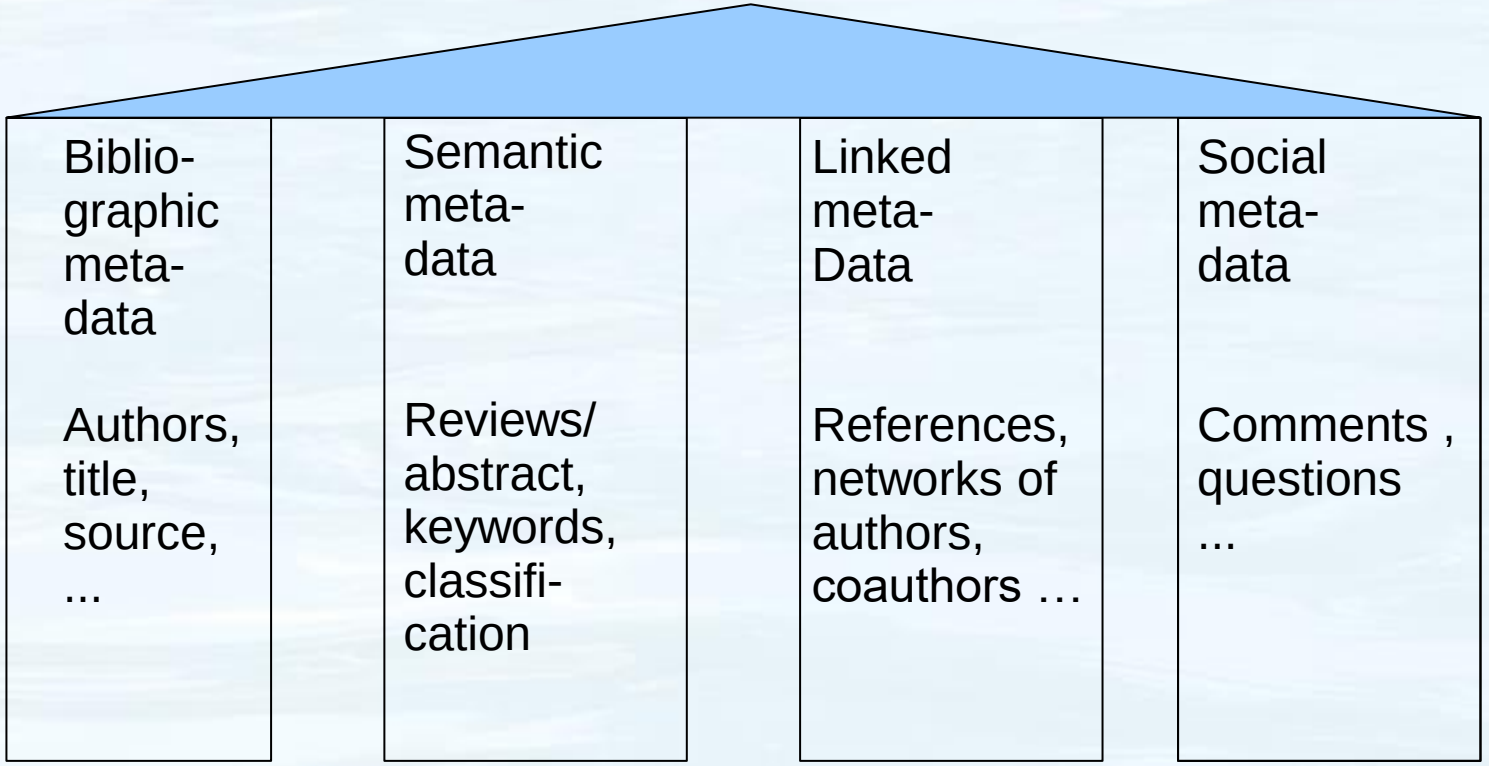
- A state-of-the-art analysis
- Enrichment of the MSC - new approaches: SKOS and a controlled vocabulary for mathematics
- Conclusions and Outlook

State of the art

Zentralblatt Math and Math Reviews:

- the leading reviewing journals in mathematics and its applications
- coverage: more than 3,000,000 bibliographic entries of mathematical publications (journal articles, monographs, textbooks from 1820 up to now
- systematic analysis of the whole literature of mathematics

Facets of content analysis



(in Zentralblatt Math and Math Reviews)

Different levels of semantic metadata

- Reviews (individual)
- Keywords (semi-formal, but no controlled vocabulary exists for mathematics up to now)
- Classification (formal, no degrees of freedom)

Classification in mathematics: Mathematics Subject Scheme (MSC)

Features of the MSC:

- a topic-specific classification scheme
- nodes: more than 6,000 nodes (63 on the top level,
  more than 500 on the second level, more than 5,000 on the third level)
- relations: hierarchical relations are the most important relations within the MSC,
  but there are further two types of similarity relations:

See also and *For … see …*

Printed and electronic versions of the MSC

- up to 2010 the master of the MSC was the printed version (advantages: nearly linear reading), but is only of limited use for the retrieval in the database
  reasons: too many groups, too complex, not intuitive for the most users, it is much simpler to search for keywords and names
- there is an electronic master of MSC 2010 (TeX-encoded), the TeX-encoded MSC is not machine-understandable

Restrictions and deficits of the MSC (I)

- the TeX-encoded version doesn't use standards for a semantic analysis of the structure of the MSC, so it is not interoperable with other classification schemes;
  the classes are defined only by their labels and their location within the MSC
- the labels of the classes are not unique
- the MSC is heterogeneous: the classes have different types, especially: modeling,mathematical objects, theories and methods, etc.

Agenda

- A state-of-the-art analysis
- Enrichment of the MSC - new approaches: SKOS and a controlled vocabulary for mathematics
- Conclusions and Outlook

Enrichment of the MSC: Transformation to SKOS

Create a SKOS-encoded form of the MSC
(SKOS – Simple Knowledge Organization Scheme)

Why SKOS?

- SKOS provides a standardized vocabulary for classification schemes, thesauri, etc.
- SKOS is based on XML and RDF,
  this means SKOS can be extended to individual requirements, e.g., formula analysis in mathematics

The first step:
Encoding of the MSC in SKOS:

- a 1:1 translation form the TeX-master to a SKOS master (so we can model the MSC given by its classes and hierarchical relations)

the result: we have the same content as in the current version, but

- the content is encoded in a machine-understandable way (so it can be used by other schemes and applications)
- the scheme is extensible: we can add further information

```
msc:MSC2000 rdf:type skos:ConceptScheme;
 dct:title "MSC 2000";
 dct:creator msc:MRandZbl ;
 dct:created "2002-07-08T00:00:00-04:00" ;
 dct:modified "2002-07-08T00:00:00-04:00" ;
 skos:hasTopConcept msc:msc00-XX ;
 ...
 skos:hasTopConcept msc:msc39-XX ;
 ...
 skos:hasTopConcept msc:msc97-XX .

....
msc:msc39-XX skos:prefLabel "Difference and functional equations"@en ;
 skos:notation "39-XX"^^mscd:MSCNotationDatatype
 skos:inScheme msc:MSC2000 ;
 skos:inScheme msc:MSC2010 .

msc:msc39Axx skos:prefLabel "Difference equations {For dynamical systems, see 37-XX; for dynamic equations on
time scales, see 34N05}"@en ;
 skos:notation "39Axx"^^mscd:MSCNotationDatatype ;
 skos:broader  msc:msc39-XX ;
 skos:related  msc:msc37-XX ;
 skos:related  msc:msc34N05
 skos:inScheme msc:MSC2000 ;
 skos:inScheme msc:MSC2010 .

...

msc:msc39A10 skos:prefLabel "Difference equations, additive"@en ;
 skos:notation "39A10"^^mscd:MSCNotationDatatype
 dct:created "2002-07-08T00:00:00-04:00" ;
 dct:modified "2002-07-08T00:00:00-04:00" ;
 skos:broader  msc:msc39Axx ;
 skos:inScheme msc:MSC2000 ;
```

Enhancement of the MSC model

- up to now: the MSC model is just a classical graph model
- overlapping of classes couldn't be modeled in the printed form, but now we can do it!

- the idea is very simple:
  we use the terminology used in mathematical publications and add this information to the scheme

Description of MSC classes by terms

The idea: each class will be characterized by a (weighted) vector of terms

In more detail:
Using machine-learning tools
Step 1: Vocabulary in MSC and other sources provide a start vocabulary
Step 2: Analysis of existing keywords in the databases
Step 3: Keyword extraction of the (classified) information of the databases ZBMATH and MathSciNet

The usual problems

- relevant terms are typically phrases, not single words
- synonyms and homonyms
- different grammatical forms of phrases
- abbreviations which are often used

⟹ Controlled (semi-automatic) processing

The (fictional) result for MSC class
Ordinary Differential Equations (MSC 34-XX)

Terms	Occurences
Linear ODE	371
Nonlinear ODE	1072
Fractional ODE	96
Stability	781
Periodic Solutions	37
...	...

Benefits

- a precise and dynamic characterization of the MSC classes
- a controlled vocabulary for mathematics
- a tool which can be used for
  - clustering of documents (similarity analysis of documents,(semi-)automatic keyword extraction and classification,
  - keyword generation by authors,
  - sophisticated retrieval features (MSC as a hidden method for retrieval)

Further steps

- a rigid facet structure for the MSC (reducing the size of the MSC)
- a typing of the MSC classes
  - mathematical modeling,
  - mathematical objects (e.g., ordinary differential equations),
  - theories and methods (e.g., K-theory),
  - qualitative aspects (e.g., stability)
  - applications (within mathematics and in other fields)
- formula analysis and formula search (MathML)

Agenda

- A state-of-the-art analysis
- Enrichment of the MSC - new approaches: SKOS and a controlled vocabulary for mathematics
- Conclusions and Outlook

Conclusions and Outlook

- knowledge management in mathematics is at a turning point, we need new machine-based methods for content analysis and a new quality of service
- using standards formats for the MSC (e.g., methods of Semantic Web allowing a machine-processing of semantic information)
- enhancement of the MSC by combining different (classical) methods of semantic analysis (e.g., classification, controlled vocabularies, etc.)

- We are on the way!

Thanks!