Content analysis and classification in mathematics

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

UDC Seminar 2011
CLASSIFICATION & ontology
Formal approaches and Access to Knowledge
The Hague, 19-20 September 2011

A snippet from the SKOS version of the MSC

….<skos:Concept rdf:about="&msc2010;00Axx"> <skos:prefLabel xml:lang="en">General and miscellaneous specific topics</skos:prefLabel> <skos:prefLabel xml:lang="zh" rdf:parseType="Literal">一般的和各种特殊的课题</skos:prefLabel> <skos:prefLabel xml:lang="it" rdf:parseType="Literal">Argomenti generali e specifici miscellanei</skos:prefLabel> <skos:notation rdf:datatype="&mscsmpl;MSCNotationDatatype">00Axx</skos:notation> <skos:inScheme rdf:resource="&msc2010;" /> <skos:broader rdf:resource="&msc2010;00-XX" /> <skos:narrower rdf:resource="&msc2010;00A05" /> <skos:narrower rdf:resource="&msc2010;00A06" /> <skos:narrower rdf:resource="&msc2010;00A07" /> <skos:narrower rdf:resource="&msc2010;00A08" /> <skos:narrower rdf:resource="&msc2010;00A09" /> <skos:narrower rdf:resource="&msc2010;00A15" /> <skos:narrower rdf:resource="&msc2010;00A17" /> <skos:narrower rdf:resource="&msc2010;00A20" /> <skos:narrower rdf:resource="&msc2010;00A22" /> <skos:narrower rdf:resource="&msc2010;00A30" /> <skos:narrower rdf:resource="&msc2010;00A35" /> <skos:narrower rdf:resource="&msc2010;00A65" /> <skos:narrower rdf:resource="&msc2010;00A66" /> <skos:narrower rdf:resource="&msc2010;00A67" /> <skos:narrower rdf:resource="&msc2010;00A69" /> <skos:narrower rdf:resource="&msc2010;00A71" /> <skos:narrower rdf:resource="&msc2010;00A72" /> <skos:narrower rdf:resource="&msc2010;00A73" /> <skos:narrower rdf:resource="&msc2010;00A79" /> <skos:narrower rdf:resource="&msc2010;00A99" /> <skos:exactMatch rdf:resource="&msc2000;00Axx" /> <skos:exactMatch rdf:resource="&msc1991;00Axx" /> </skos:Concept> <skos:Concept rdf:about="&msc2010;00A05"> <skos:prefLabel xml:lang="en">General mathematics</skos:prefLabel> <skos:prefLabel xml:lang="zh" rdf:parseType="Literal">普通数学</skos:prefLabel> <skos:prefLabel xml:lang="it" rdf:parseType="Literal">Matematica generale</skos:prefLabel> <skos:notation rdf:datatype="&mscsmpl;MSCNotationDatatype">00A05</skos:notation> <skos:inScheme rdf:resource="&msc2010;" /> <skos:broader rdf:resource="&msc2010;00Axx" /> <skos:exactMatch rdf:resource="&msc2000;00A05" /> <skos:exactMatch rdf:resource="&msc1991;00A05" /> </skos:Concept>

……

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC – remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

The Mathematical Subject Classification (MSC) – basics (I)

- MSC - a classification system for the whole mathematics and its applications
- MSC is maintained by Mathematical Reviews and Zentralblatt MATH in a dialog with the mathematical community
- MSC is a broadly accepted standard for the classification of the mathematical literature

MSC - basics (II)

- MSC was developed in the 70-ies
- MSC is updated periodicously (each decade)
- MSC is minimal in the sense, that it covers only
  - the classes
  - the hierarchical structure (tree structure) and some similarity relations (overlapping)
- MSC is a weakly faceted classification scheme: It allows besides the classification of the mathematical subject also the formal type of the classification (in parallel)

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC - remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

MSC – comments and problems (I)

- Definition of the classes:
  An exact definition of the MSC classes is missing.
  The labels of the MSC are not unique (e.g., stability or applications)
- Retrieval:
  the use of classification for searching is decreasing (typically keywords and author names are used for retrieval)
- Number of classes (to large):
  - 63 classes on the top level,
  - more than 500 classes on the second level an
  - more than 5.000 classes on the third level.

MSC – comments and problems (II)

- Granularity:
  The granularity varies from subject to subject.
- Faceted structure:
  The faceted structure of the MSC is poor (a faceted structure has the potential to reduce the number of classes and has the potential for enhanced navigation features)

MSC – comments and problems (III)

- Format:
  The first electronic master version was introduced 2010 MSC2010.
  The master is TeX-encoded (the classes can cover mathematical characters).
  Some other formats are derived (HTML, PDF, Word, a KWIC index, etc.).
- Compatibility:
  The MSC is specific for mathematics. It is not directly compatible with other classification schemes, as UDC, DDC, LCSH, etc. The TeX-master version of the MSC isn‘t machine understandable.

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC – remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

MSC and SKOS (I)

- Making the MSC machine understandable:
  Use of Semantic Web technologies and standards (RDF/XML, OWL, SKOS)
- A pre-comment:
  M. Panzer and M. L. Zeng have presented some difficulties for a SKOS encoding of the DDC. The existing SKOS standard is not rich enough to model all features of the DDC. This is also valid for the MSC.

MSC and SKOS (II)

- Special problems for the conversion of TeX-encoding into SKOS encoding:
  - Encoding of mathematical characters in the labels of the classes (use of Unicode instead TEX)
  - Modelling of the similarity relations
    MSC uses different types: *seeAlso* , *forSee*, *seeMainly)*
    These relations are embedded in the labels. Moreover, the references are of different type: references to single classes but also to sets of MSC-classes (intervals)

MSC and SKOS (III)

- Multilinguality
  There exist some translations of the MSC (Chinese, Italian)
- History
  The current MSC version has to be linked with the previous versions.
- Modelling of facets of the MSC (collections???)

A first implementation (draft version)

MSC and SKOS (IV)

The SKOS encoding provides some additional value:

- The MSC-SKOS encoding allows an automatic processing of the information (main result).
- Concordance relations to other classification schemes can be added.
- The structure of the MSC will become clearer.
- Multilinguality is no problem.
- Preliminary versions of the MSC can be inserted.

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC – remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

MSC and a controlled vocabulary (I)

Semantic enrichment of the MSC

- Controlled vocabulary
  - building a controlled vocabulary of mathematics (thesaurus)
    up to now only (small) subsets of a controlled vocabulary exist
    automatic learning methods
    linguistic and/or statistical methods, e.g., Latent Semantic Indexing, Latent Dirichlet Allocation (term-frequency analysis)
  - definition of the MSC classes by characteristic phrases

MSC and a controlled vocabulary (II)

- Some initiatives to build up a controlled vocabulary for mathematics
  (MKM-community, DeliverMATH-project)
- Encoding the controlled vocabulary as an separate SKOS scheme
- Bundling of MSC and the controlled vocabulary by matching of two SKOS schemes (corresponding to the proposal of Panzer/Zeng)

MSC and a controlled vocabulary (III)

The aims (and the potential):

- use the controlled vocabulary for a description of MSC classes
- use the controlled vocabulary for the keyword extraction of mathematical publications
- use the controllled vocabulary for an automatic classification of mathematical publications and a clustering of publications
- use it for better retrieval in our databases

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC – remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

Some further ideas for the semantic enrichment of the MSC

- Faceted structure of the MSC by typing, reduction of the number of classes, enhanced navigation features)
- Formula index („controlled vocabulary of formulas“)
  - building a formula index for mathematics (basis: mathematics in XML: MathML, OpenMATH, OMDoc, …)
  - use the controlled vocabulary for a characterization of MSC classes by characteristic formulas (project proposal)

Agenda

- The Mathematical Subject Classification (MSC) – basics
- MSC – remarks and problems
- MSC and SKOS
- MSC and a controlled vocabulary for mathematics
- Some further ideas for the semantic enrichment of the MSC
- Conclusions and Outlook

Conclusions and outlook

- MSC can be also an important tool for mathematical information in the future but its role and design will be under a permanent change.
- A SKOS implementation is an important first step in the MSC development.
- We need further semantic enrichments of the MSC
- We are on the way! (broad discussion + practical developments)

Thanks