Editorial – On the Road to MSC 2020

Adam Bannister (FIZ Karlsruhe, Berlin, Germany), Fabian G. Müller (FIZ Karlsruhe, Berlin, Germany), Mark-Christoph Müller (Heidelberg Institute for Theoretical Studies, Heidelberg, Germany) and Olaf Teschke (FIZ Karlsruhe, Berlin, Germany)

Eighteen months ago, the beginning of the revision of the Mathematical Subject Classification was announced.[1] Since then, the mathematical community has already contributed a number of suggestions on the public wiki available at https://msc2020.org/. In this article, we will give a brief overview of the current usage of the MSC, analyse some data related to its effectiveness and precision, relate it to topic clusters generated by data mining techniques and indicate some trends that have become visible in the course of the revision.

Current usage of the Mathematical Subject Classification (MSC)

While the raw scheme had already been introduced in the early volumes of *Mathematical Reviews*, the current shape, as a joint effort of both mathematical reviewing services, evolved in 1980. This came after an initiative of Bernd Wegner to incorporate the system into *Zentralblatt MATH* and to maintain regular collaborative revisions. Since then, MSC has been primarily used by MathSciNet and zbMATH reviewers and editors to classify the mathematical research literature, as well as being adapted by classical and digital libraries and journals. Several recent developments in the zbMATH database, such as author, journal and citation profiles or filter functions, have utilised the subject information beyond its original raison d'être.

How reliable is the MSC?

MSCs are assigned to books and papers by authors, reviewers and editors, with the final classification approved by MathSciNet and zbMATH section editors. Naturally, as a human enterprise, such assignments may be subjective. Hence, it is natural to ask about the degree of subjectivity that comes along with a classification performed by hand and whether it is possible to derive conclusions for the revision from the degree of vagueness. To determine this, a comparison of MSC2010 assignments has been made for 78,063 articles published between 2010 and 2016 in journals indexed cover-to-cover by both MathSciNet and zbMATH. For this corpus, both services coincided at the top level MSC for 62,951 documents, and even for 40,244 at the level of the overall MSC. More precisely, the average F1[2] score for the coincidence of MathSciNet and zbMATH classifications turned out to be 0.83 at the top level, 0.72 at the second level and 0.58 at the third level of the first assigned MSC.[3] The concordance turned out to be significantly larger when permutations were taken into account; indeed, the largest differences by far occurred in the cross-subject MSC sections like 00 (General), 97 (Education), 58 (Global analysis), 19 (K-Theory) and 37 (Dynamical systems). Interestingly, three of them were introduced in the 1991 and 2000 MSC revisions. Hence, while the MSC has overall become less tree-like, with more cross-references introduced in the last revisions, it seems that a large proportion of the articles still fit conveniently into the more classical hierarchical structure of the main subjects. Consequently, it seems justified that there has been no introduction of a new top-level MSC in 2010 and there also seems no need to do so in 2020.

The unreasonable effectiveness of the MSC

The relative reliability of the top-level MSCs can also be derived from the cross-citation Figure 5 in Bannister and Teschke,[4] which shows a strong concentration of references to articles with the same MSC. In this sense, the main subjects can also be seen as most natural clusters of the citation graph. Naturally, due to the interconnected nature of mathematics, this effect is less significant for more granular MSC levels. However, the question remains of whether there are automated ways to organise mathematical literature into subjects. Apart from graph-theory approaches, the last decade has seen tremendous progress in topic modelling by data mining and machine learning techniques. An experiment performed by the Heidelberg Institute of Theoretical Studies (HITS) created several clusters using the ToPMine tool.[5] Human evaluation showed that it performed reasonably well for applied areas (producing, for example, a cluster containing Bayesian inference, posterior distribution and the Gibbs sampler, roughly corresponding to 62F15) but was quite limited for pure mathematics (e.g. it joined the notions of pull back and container loading from category theory and operations research and created the cluster "hyperplane arrangement, traffic jam, speed of light" of hitherto unknown mathematical semantics). Some of the effects may derive from the fact that publication numbers are extremely unevenly distributed in mathematical areas and automated methods tend to underperform for areas with relatively small publication numbers, which are often, however, very important within the mathematical corpus.

[1] E.G. Dunne and K. Hulek, MSC2020 – announcement of the plan to revise the Mathematics Subject Classification. *Eur. Math. Soc. Newsl.* 101, 55 (2016).

[2] Weighted harmonic mean of the fractions of MSC codes in one set that also occur in the respective other.

[3] MSC codes have three levels of increasing granularity, denoted by two digits, a letter and two more digits.

[4] A. Bannister and O. Teschke, An Update on Time Lag in Mathematical References, Preprint Relevance, and Subject Specifics. pp. 41–43. *Eur. Math. Soc. Newsl.* 106, 41–43 (2017).

[5] http://illimine.cs.uiuc.edu/software/topmine/.

Developments toward MSC2020

Taking the mentioned limitations into account, quantitative methods such as those mentioned above can be used to create suggestions for the MSC2020 revision. Phrases that have occurred much more frequently since 2010 have often included developments in the applications of mathematics (which tend to be both more numerous in publications and more fluid in topic denomination), for instance "loop quantum gravity" and "PT symmetry" in quantum theory, "scaling limits" arising both in stochastics and physics, "exponential stability" in control theory, "quantum circuits" and "quantum games", as well as "sparse graphs", "spatial graphs", "circulant graphs" and "phylogenetic trees" connected to the rise of network research, "copulae models" in statistics, "character varieties" in algebraic geometry and topology and the cluster "Khovanov/Heegaard-Floer/HOMFLY homology" from topology, along with transcending techniques like "matrix factorisation". Copulae and character varieties have already been independently proposed in the MSC2020 wiki, as well as many new developments not detected by automated methods, such as "numerical algebraic geometry", "higher categories", "topological data analysis" and "computer-assisted proofs". On the other hand, several recent concepts (like homotopy type theory) are still missing, so please engage in the joint effort and contribute to the MSC2020 wiki at https://msc2020.org/, which will remain open until August 2018!

Adam Bannister [adam.bannister@fiz-karlsruhe.de] has a postgraduate diploma in Geographic Information Systems and currently works on the Scalable Author Disambiguation for Bibliographic Databases at zbMath in cooperation with Schloss Dagstuhl and Heidelberg Institute for Theoretical Studies.

Fabian Müller [fabian.mueller@fiz-karlsruhe.de] studied mathematics and computer science at Humboldt-Universität, Berlin. After finishing his doctoral studies in algebraic geometry in 2013, he started working at zbMATH, where he is responsible for coordinating IT development efforts.

Dr. Mark-Christoph Müller [mark-christoph.mueller@h-its.org] is a research associate in natural language processing at the Heidelberg Institute for Theoretical Studies and currently works on the Scalable Author Disambiguation for Bibliographic Databases project, in cooperation with Schloss Dagstuhl and FIZ Karlsruhe.

Olaf Teschke [olaf.teschke@fiz-karlsruhe.de] was a member of the Editorial Board of the EMS Newsletter from 2010–2017, responsible for the zbMATH Column.