swMATH - Challenges, Next Steps, and Outlook

Hagen Chrapary[1],[2], Wolfgang Dalitz[2], and Wolfram Sperber[1]

[1] FIZ Karlsruhe/zbMATH , Franklinstr. 11, 10587 Berlin, Germany
[2] Zuse Institute Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany

Abstract. swMATH [**?**] is currently one of the most comprehensive specialized information services and search engines for Mathematical SoftWare (MSW). It was the intention of the project to support the user community by providing information about MSW and searching relevant MSW. Currently swMATH lists information of more than 13,000 items and 120,000 publications which refer to MSW. Maintaining and updating of the service is mainly done automatically, the number of requests is permanently increasing. This sounds like a perfect solution. But swMATH is only a first step to a powerful information system on mathematical research data. This talk addresses some open problems for the further development of the swMATH service. It is shown that some problems for swMATH lead to central questions for the information management of mathematical research data, especially for MSW. This contains an extended content analysis of MSW, versioning and citation standard of MSW, a typing of the swMATH resources and the presentation of context information and high-quality control within the swMATH service.

1 Introduction

The advent of computers has lead to the development of a new mathematical branch: MSW. MSW has some unique features which are different from the classical form of the presentation of mathematical knowledge, the publications

- MSW is written in a special programming language, not in a natural language as publications. The source code of a MSW is not self-explaining and often inappropriate for human's reading and automatic content analysis.
- MSW depends on hard- and software, programming languages or environments, etc.
- For human users of MSW, special forms of instructions (documentations, manuals, etc.) have been established. This information is often provided on own Web sites of a MSW.
- MSW realizes a mathematical concept or an algorithm for solving special classes of mathematical models ("computational model") which can be presented independently from the MSW.
- Typically, MSW has instances, e.g. versions and develops dynamically. Moreover, MSW has a life cycle. This requires a version management of the MSW (in the past, mathematical publications were static).
- The dynamic character of MSW has led to the development of repositories for maintaining and providing MSW. There are some commercial companies but also a lot of developer communities, e.g. the R community, who run own repositories.

– The increasing use of MSW in mathematics and other scientific fields, industry, services, and administration requires an exact citing of the MSW. Results without exact references of the used MSW instance significantly complicate the reproduction of the results.
– Standardized data models for the description of and a citation standard for MSW are missing. Metadata schemes for publications cannot be transferred one-to-one for MSW.
– Methods for the quality control of MSW are under development. Some specialized journals on MSW were founded. Typically, publications on Open Source MSW consist of a text-based publication describing the algorithm or the MSW supplemented by the software code and testdata. The quality control of these publication contains also a peer-reviewing of the MSW, typically an installation of the MSW and a validation of the results.

The demands on information services for MSW are diverse, especially search, access, version management, and archiving. This has led to the development of different kinds of information services for MSW: directories which list MSW and provide information about it, repositories, which maintain and archive MSW for a special subject and programming language, and individual websites for a MSW or specialized mathematical services.

The swMATH service is a directory for MSW. There are a lot of other directories for MSW. The manual maintenance is expensive. Therefore we have developed a concept for a machine-based concept.

2 swMATH in a nutshell

A central concept of swMATH is the so-called publication-based approach. MSW is often accompanied by publications. By analyzing the mathematical literature, MSW is identified and information about it is extracted. Instead of analyzing the full texts of publications, the information of the database zbMATH [**?**] is used, especially titles, reviews or abstracts and references. Software citations in mathematical publications are often short and restricted to the name of the MSW, e.g. ”SAGE", without precise information about instances, e.g. versions. Often an exact information on the type of the resource is missing. Sometimes also information which is closely related to MSW but not MSW itself is detected, e.g. services, languages and environments, testdata and benchmarks, etc. We have decided to insert also these kinds of information into swMATH. In the following, all entries in swMATH are labeled as MSW.

The heuristic analysis starts with searching for characteristic patterns for MSW. Each publication which has a link to a MSW is classified either as ”standard publication" if a MSW is the main content of the publication or as ”user publication" if the MSW is used as a tool for calculation. As said above, publications and also zbMATH data often don’t contain information about different instances of a MSW. Hence there can exist more than one standard publication (a standard publication can exist for each instance of a MSW). A citation standard for MWS would significantly improve the situation. Some questions on a citation standard will be discussed in detail below.

But the analysis of the publications offers great potential to gain relevant information about MSW. This is done in a second step Currently the main information in swMATH contains the following information (metadata) on MSW:

- the name of the MSW
- a short description of the content of the MSW. This description uses the review or the abstract of a standard publication of the MSW and provides information about the content of the MSW addressing such questions ´Which problems can be solved using MSW?, Which algorithms are realized by the MSW?, What are the special features of the MSW? ˙
- a keyword cloud which aggregates the keywords of standard and user publications
- classification codes of the standard and user publications correspondingly to the Mathematics Subject Classification 2010 (MSC2010) [**?**]
- a list of publications which cites the MSW
- a list of "related software" is calculated by the data of publications
- a profile of the acceptance of the MSW measured by the number of publications per year

Moreover, a simple and an advanced search feature over the swMATH entries is provided. The number of references of MSW is used for a ranking of information. For example, the importance of a keyword in the keyword cloud results from its frequencies, the MSC codes provide relevant information about the mathmatical subjects and the applications areas. The number of references in peer-reviewed publications can be also interpreted as an indicator for the quality of MSW.

Publications contain not all information about a MSW. Typically, the Web sites of a MSW are excellent resources for specific information, e.g. they contain manuals or documentations, provide technical specifications, legal rights, etc. By a websearch the URL of the website of a MSW – if existing – is determined. We are working on more specific methods for the automatic analysis of the websites of MSW. We have started to identify manuals and documentations of MSW on the websites.

Figure 1 shows the swMATH entry for the MSW "Singular" .

Commonly, websites of a MSW contain more or less information about the current instances of a website. But the publications refer not only to a current instance of a MSW but also to outdated instances. Web archives which periodically store the Web could enrich the information about the different stages of a MSW and help to assign the publications to the corresponding instances. Therefore the swMATH directory of swMATH will be used as a seed list for to search information about MSW. It is the plan to combine this information with the existing information in swMATH.

All swMATH entries (families of MSW) get a persistent identifier, the running number of the swMATH input of the MSW. The identifier could be extended by adding data of an instance.

The publication-based approach and the use of the zbMATH database allows a widely automatic maintaining and updating of swMATH. The publication-based approach also ensures the presentation of MSW in the context with the mathematical literature.

swMATH is an open access service for the mathematical community.

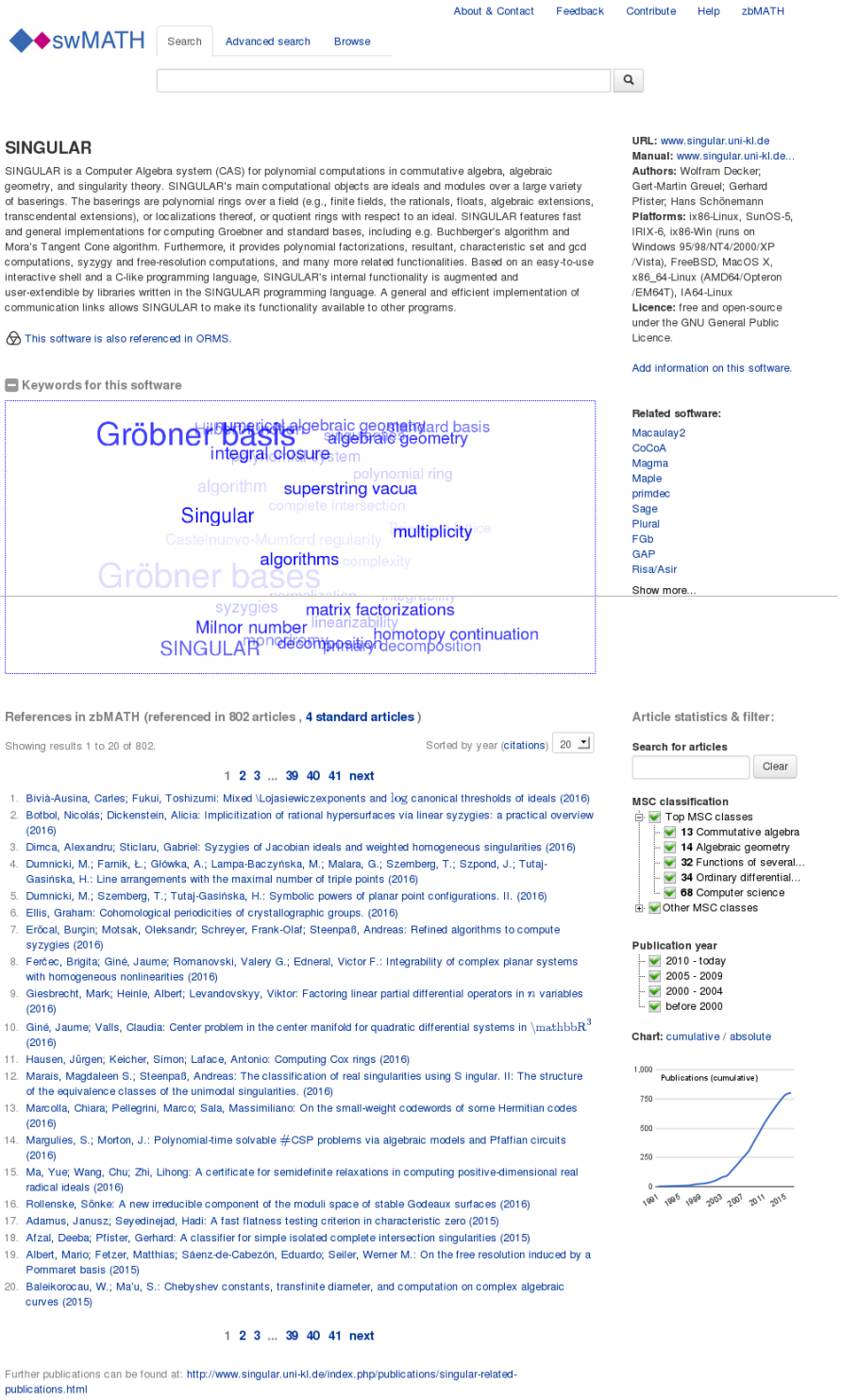

Fig. 1. The detailed information for the software "Singular"

3 Challenges

3.1 MSW instances and citation standard for MSW

The analysis of mathematical publications seems to be an efficient way to search for MSW. But the use of heuristic methods can fail, e.g. if the name is a word from the natural language. Also, the heuristic methods don't allow to determine instances. A standardized referencing of MSW would have important advantages: a precise information about MSW used, better visibility of the MSW, reproducibility of research results, secure identifiability of the MSW, etc.

Software is ubiquitous on computers. A precise citing of the used instance of a software is a precondition for reproducing the results which were achieved by using a software. Different instances of a software can be incompatible, are derived from other algorithms and provide other features. Software has other characteristics as publications, especially its dynamic character. Citing of software is a widely discussed topic in different communities, see the general discussion Mike Jackson's blog [**?**] "How to cite and describe software" . The standardization of citations has different aspects. MSW is distributed in various different ways which has an influence on a citation standard. Some publishers and software providers have developed and implemented own models and recommendations for citing its MSW. A main requirement to a citation standard for MSW is that the citation standard must guarantee the unique identifiability of the software used, typically such a reference must refer the instance of the software used. A MSW can have instances, e.g. versions or revision numbers. Instances can characterize the state of the art of a MSW, dependencies from operating systems, programming languages, or other software, or different legal rights and usability conditions. MSW can be accessible via a repository and explicitly defined versions are missing, here timestamps or hash tags could be helpful information for the verification of the used MSW instance.

The current situation is characterized by proprietary standards and recommendations for the citation of software. A widely used approach for the citation standard is the "Who – When – What – Where" -format, propagated, e.g. by the American Psychological Association. This approach is derived from the citation standard for publications. In principle, this recommendation can be used but the details of this recommendations are not specified.

The R community, a community-driven initiative in mathematics statistics, recommends to cite the corresponding manuals - if existing âĂŞ instead of directly referencing the MSW. Typically, there is an 1:1 relation between a manual and an instance of a MSW. But, the link to the software cannot be generated from the manual. For the MSW based on R this is not important because the CRAN repository allows a fast access to the R packages.

Jackson has published some recommendations which contain necessary information. His general hint: Publications on a software should cited as a complement not as a replacement for citing software.

He distinguishes between the following four types for referencing:

- software purchased off-the-shelf which can be described by the following data: Product name; Version; Release date; Publisher; Location

– software downloaded from the Web which can be described by the following data:
  Product name; Version; Release date; Publisher; Location; DOI or URL; Download date
– software checked out from a public repository which can be described by the following data:
  Product name; Publisher; Location; URL; Repository specific checkout information
– software purchased by a researcher which can be described by the following data:
  Product name; Author; Location; Contact details; Received date

These data describe the used software in a unique way. But the proposal requires to apply different metadata schemes which hampers the acceptance and use. Further information especially operating systems or used packages or libraries could be optional parts of the citation standard.

A further problem of a citation standard of MSW is its implementation. For the mathematical community, TEX is the lingua franca to write mathematical documents. References are done in BibLATEX[**?**] which has widely substituted the outdated BIBTEX software. BibLATEX is more flexible than BIBTEX and allows to define new types and corresponding metadata schemes but doesn't support a type "software" until now. This type must be specified, mandatory and optional fields describing the MSW must be defined and the backend software. A first BibLATEX proposal has been created with "Biber" as backend software.
Remark: The problem of a citation standard for MSW should be discussed more generally for all types of mathematical research data, not only for MSW. We discuss this problem in more detail in the following section.

3.2 Context, resource types, and classification of swMATH entries

The publication-based approach described above links a MSW to the publications which cite it. The publications contain data about the investigated problems and models, their mathematical treatment, data of measurements which are used for the input, parameters, etc. But often these data are incomplete, not explicitly linked to MSW (e.g. which model is solved by the MSW?) which makes it difficult to reproduce and reuse scientific results.

The mathematical subject, concepts, models, and methods as the application of mathematics are very heterogeneous for MSW. A classification of mathematical research data by its subject and type would be useful for retrieval and navigation. It allows to present MSW in its context, especially the relation to the mathematical models which are solved by using MSW, the algorithm, the used data and parameters for calculations.

A comprehensive and structured content analysis starting with the original mathematical models, the parameter and input data, the computational algorithms and the computational models used for MSW as well as the computational results would make the research process more transparent and improve the reproducibility of results.

Subject classification in swMATH

The MSC2010 is the most accepted subject classification scheme for mathematical publications. The scheme covers pure mathematics as well as a broad spectrum of application areas. For publications with the focus on MSW a special classification code XX-04 for the most of the 63 top classes of mathematical subject in the MSC is pÃ§rovided. Last but not least all zbMATH entries are classified by the MSC.

The publication-based approach of swMATH suggested to use the MSC also for the characterization of the mathematical subjects of a MSW. Of course, other classification schemes, e.g. the Guide to Available Mathematical Software [**?**] index could be added.

Type classification in swMATH

Currently, swMATH contains different types of mathematical data with the focus on MSW. Currently, entries of the following types of objects are inside swMATH:

- MSW
  A MSW could be defined as source code, computer programs, or software libraries for analyzing or calculating mathematical problems or models
- Mathematical languages and environments for MSW
  Languages and environments are programming languages for special mathematical subjects, e.g. the language and environment R [**?**]for statistics.
- Mathematical services
  The common property of mathematical services is the presentation of information. The information can result from retrieval in databases or from online computing or a mix of both. The Online Encyclopedia of Integer Sequences or the Digital Library of Mathematical Functions are impressive examples of widely used mathematical services.
- Testdata and benchmarks
  Testdata and benchmarks are important for the verification of results of the use of MSW, to test its performance and to decide which MSW is used for solving a mathematical problem.
- Further mathematical research data, especially models and model parameters, visualizations and simulations This class is very heterogeneous and covers numerous model parameters but also complex mathematical objects and models, e.g. polynomials or ideals for computer algebra.

This first approach for classifying the types is resulting from a pragmatic point of view on the swMATH entries.

But, the scheme reveals also the difficulties of a classification scheme for mathematical research data: The classes are characterized by different properties. Different classes can strongly overlap. The swMATH entries can be elements of different classes. On the other hand, a classification of the resource types for the swMATH entries is useful, e.g. for the presentation of related software. The feature ”Related software" is important for the users of swMATH. It provides an overview about similar swMATH entries. Up to now the similar entries are determined by common references in publications and common MSC codes. But if we don’t want to compare apples to oranges,

e.g. a MSW to a benchmark, we need a typing of the swMATH entries. The resource type scheme allows a structured presentation of the environment of a MSW, not only of similar MSW but also of related benchmarks, models etc.

As usual we propose a hierarchic classification scheme for the types in swMATH. All these classes can be subdivided. Up to now, a first flat type scheme is implemented in the production database of swMATH.

The typing could be presented as an additional facet in the UI of swMATH.

3.3 Applied specific information in swMATH

Here we restrict ourselves to two aspects. At first we discuss the enrichment of the existing information of application aspects in swMATH. Currently, information which is of special interest for users outside of mathematics is included in the MSC classification, especially the MSC codes for the application areas, and in the keyword cloud by application area-specific keywords. But there is no explicit tagging of mathematical and application-specific keywords. Moreover, keywords and a classification codes of the application areas are not mandatory. Typically, non-mathematicians actively participate in the modeling and the evaluation process. They are especially interested to get detailed information about the models, MSW and the results of the calculations. This information would help the users to decide which models and MSW could be suitable for solving their problems. Publications in applied mathematics often contain a description of the original problem and the mathematical models. Heuristic methods could be developed for searching specific information for users in the publications which are listed in the swMATH entries, especially for models.

A more general aspect is the ambiguous borderline between MSW and non-mathematical software. The publication-based approach in swMATH is restricted to publications which are reviewed by zbMATH. swMATH provides an interface to inform about MSW outside of the scope of zbMATH.

3.4 Quality of MSW

Quality measures of MSW are difficult. The quality control of MSW cannot be reduced to a single factor. It is a complex process. Up to now, swMATH doesn't make any statement about the quality of the MSW. References to a MSW within the zbMATH data are a necessary condition for the inclusion of the MSW to the database swMATH. Reference to a MSW in a peer-reviewed publication can be interpreted as an indirect quality measure for the cited MSW. The number of references is also an indicator for the quality. In the last 20 years, special journals with the focus on scientific software were founded, in mathematics especially the journals "ACM Transactions on Mathematical Software" [**?**], "Archive of Numerical Software" [**?**], "Journal of Machine Learning Research" [**?**], "Journal of Software for Algebra and Geometry" [**?**], "Journal of Statistical Software" [**?**], "Mathematical Programming Computation" [**?**], "Numerical Algorithms" [**?**], or "The R Journal" [**?**]. For a complete list of journals with the focus on software see Neil Chue Hong's blog [**?**] "In which journals should I publish my software". A typical publication in these journals consists of an article which describes the algorithm and/or the MSW and the code of the MSW. All these journals have defined

explicit more or less standardized policies for submission and peer reviewing which involve also the software code. The policies include the installation and feasibility of the MSW for the peer-reviewing process and accepted dependencies from other software and programming languages. The verification of the results which were achieved by using the MSW is a further important criteria for the quality control. Therefore, the submissions must include the original testdata. Of course, the evaluation of MSW articles is more expensive than this of a traditional publication. Some journals have technical editors which are responsible for the reviewing of MSW.

Quality-controlled MSW, especially the one which is published in the journals mentioned above could be labeled in swMATH. This would strengthen trust into the MSW, increase the usability of zbMATH, and be of great value for the user community.

3.5 Further problems

Long-term archiving of MSW is an open problem. It requires an archiving of the different software instances but also of the complete environment. The swMATH service plans to check Internet archives to provide information of the lifecycle of MSW and of outdated MSW. Also the linking of swMATH entries in swMATH with the software code is on the agenda. The granularity of MSW is a further problem. There exists very complex and powerful MSW which can be used for solving different classes of problems but also small MSW which is focused on solving a special problem. Until now there are no standardized parameters which characterize the size of MSW. Often complex MSW, e.g. Mathematica, Matlab, or R have a substructure. Packages are MSW entries with the focus on special classes of problems. swMATH has to define a policy for the handling of complex MSW. To improve the actuality of swMATH we plan to extend the publication-based approach of swMATH can be to other resources, especially identifying MSW in the e-prints of ArXiv [**?**].

4 Final remarks

Mathematical research data and a qualified access are useful for the mathematical community and outside users. swMATH is a new information service for mathematical research data especially for MSW. It was shown that heuristic methods can be successfully used for content analysis and the maintenance of swMATH. Moreover the development of the swMATH service has caused discussions about standards and classification. We hope that swMATH can contribute to build up a lean and efficient infrastructure for mathematical information.

References

1. The database swMATH (2012 -),`http://www.swmath.org`
2. The database zbMATH (1868 -),`http://www.zbmath.org`
3. The Mathematics Subject Classification 2010, `http://www.msc2010.org`
4. Netlib (1987 -), `http://www.netlib.org`
5. CRAN, `https://cran.r-project.org/web/packages/`

6. Decision Tree for Optimization Software (Plato), http://plato.asu.edu/guide.html
7. A portal for computer algebra and symbolic mathematics, http://mathres.kevius.com/comal.html
8. Gert-Martin Greuel, Wolfram Sperber, swMATH – an information service for mathematical software, in: Hong, Hoon (ed.) et al., Mathematical software – ICMS 2014. 4th International Congress, Seoul, South Korea, August 5–9, 2014. Proceedings. Springer. Lecture Notes in Computer Science 8592 (2014) , 691-701, http://www.mathematik.uni-kl.de/~greuel/Paper/GeneralType/2014-ICMS_swmath.pdf swMATH, ICMS2014
9. Mike Jackson, How to cite and describe software?, http://www.software.ac.uk/how-cite-and-describe-software?mpw
10. BibLaTeX, https://www.ctan.org/pkg/biblatex
11. Guide to Available Mathematical Software (GAMS), http://gams.nist.gov/
12. The *R* Project for Statistical Computing, https://www.r-project.org/
13. ACM Transactions on Mathematical Software (TOMS) (1975 -), http://toms.acm.org/
14. Archive of Numerical Software (2013 -), https://journals.ub.uni-heidelberg.de/index.php/ans/index
15. Journal of Machine Learning Research (2003 -), http://www.jmlr.org/
16. Journal of Software for Algebra and Geometry (2009 -), http://msp.org/jsag/about/journal/about.html
17. Journal of Statistical Software (1996 -), https://www.jstatsoft.org/index
18. Mathematical Programming Computation (2009 -), http://mpc.zib.de/
19. Numerical Algorithms (1991 -), http://www.springer.com/computer/theoretical+computer+science/journal/11075
20. The R Journal (2009 -), https://journal.r-project.org/
21. Neil Chue Hong, In which journals should I publish my software? http://www.software.ac.uk/resources/guides/which-journals-should-i-publish-my-software
22. arXiv.org e-Print archive, http://arxiv.org/