Chapter 1

swMATH: Publication-based indexing of software

Hagen Chrapary, Wolfgang Dalitz, Wolfram Sperber, and Moritz Schubotz

The quote "For a scientific institute that does not need experimental facilities, the library is the most valuable asset" (Begehr, *Mathematik in Berlin*[1]) is an exemplary assessment of the importance of mathematical publications in the mid-20th century. Mathematical knowledge was mainly available in the form of publications. With the development of computers and their use for solving mathematical problems, experiments have now also found their way into mathematical research. Computerisation and digitisation have added new forms of mathematical knowledge to the traditional ones. This has led to broad impetus in the application of mathematical knowledge. In particular, mathematical modeling, the development of algorithms, and their implementation, generally summarised under the term Scientific Computing, have become indispensable tools in the industry, the service sector, and administration.

Therefore, the extension of adequate infrastructure for the access and use of this knowledge is required. This knowledge manifests itself in the form of mathematical research data, such as the versions of software and the underlying mathematical models. The management of the different classes of mathematical research data is more complex than that for publications. This is due to several reasons, particularly the dynamic nature of software development, the dependencies on hardware, operating systems, programming languages, data formatting, modeling languages, etc. Nevertheless, mathematical publications continue to form the core of mathematical knowledge and are a source and tool for Scientific Computing. This fact can also be used to develop specific services for research data, shown in the following using the swMATH Open service as an example. With the database swMATH, FIZ Karlsruhe and Zuse Institute Berlin (ZIB) have developed the world's largest catalog for mathematical software, which currently lists almost 40,000 software objects, classifies them, and links them to software archives, such as Software Heritage, for open source software. The idea for the development is based on the evaluation of software references from the mathematical literature. The freely accessible database zbMATH is the world's largest bibliographic database of mathematical publications, with currently about 4,300,000 entries, and forms the basis of swMATH.

Today, software development is often accompanied with a publication describing the software's essential aspects (underlying algorithms, functionalities, hardware, and

[1] Heinrich Begehr, *Mathematik in Berlin – Geschichte und Dokumentation*, Erster Halbband, Shaker Verlag 1998, ISBN 3-8265-4225-8, p. 246

software requirements). These types of publications, referred to as standard publications in swMATH, are particularly highlighted in swMATH. Standard publications are often found in mathematical journals that specialize in mathematical software, such as the journal "ACM Transactions on Mathematical Software" (TOMS). A second relevant class of publications ("user publications") containing information about a software cite this software in connection with the results they have achieved with its use. Many user publications on software can be taken as an indication of the high dissemination and quality of the software.

In the case of standard publications, the name of the software usually appears in the title. In the case of user publications, references to the software can be found primarily in the full text, in the citation lists, and in the keywords. The bibliographic data fields of the database zbMATH, particularly title, keywords, Mathematics Subject Classification (MSC), citations of the database zbMATH, and other sources, such as arXiv and journals and web sources specialized in mathematical software, are evaluated for information about software. The title of the software is used to search for publications in which this software has been used.

The swMATH service provides information about all software versions (under a common name). These objects are uniquely citable via the identifier introduced in swMATH. Every software goes through a life cycle, expressed in the different versions of a software. Unfortunately, there is still no standard for referencing software. In particular, information about the version of the software is often missing, which is, however, indispensable for verifying the results. The biographical information of the zbMATH database is also used to provide information about the software, such as the mathematical areas that were starting points for the development of the software, or application fields in which the software has been used. In turn, the swMATH project led to the recording of references of mathematical software in a further data field in the database zbMATH. On the one hand, this new data field facilitates software identification. On the other hand, it underlines mathematical software's growing importance, especially for users from other scientific disciplines, industry, the service sector, and administration. The increasing importance of software citations is also expressed in the high number of software products in zbMATH; at the end of September 2021, 475,011 software references were found in 244,084 entries in zbMATH Open. Together with zbMATH Open and other publicly accessible sources such as arXiv, almost 250 thousand scientific publications are thus evaluated in swMATH.

Software user groups include software developers, users, providers, and service operators who require different information. Software data models, therefore, comprise a variety of information on the content, tasks, solution approaches, algorithms and limitations of the software, the software code, the requirements for and dependencies on hardware and software, the development status, the business model, and licenses, the developers and providers, test data and application examples, etc. These can differ significantly for individual versions and are sometimes incompatible.

A typical presentation of software entries in swMATH is demonstrated in Figure 1 using the swMATH entry of SCIP, a well-known optimization software:

Figure 1. SCIP (https://swmath.org/software/1091). The URL is persistent and can be used for referencing the software.

First, swMATH provides the information extracted from the zbMATH database. The swMATH entries are based on two pillars:

- the analysis and evaluation of the information from the publications;
- the reference to external sources.

For analysis and evaluation of the publications, the zbMATH Open data provide an excellent basis for analysis and evaluation. From this, the following data are generated via the software for the swMATH entry:

- The description of the software: For this purpose, a summary of the current standard publication of the software is used, if available.
- The keywords from the standard, and user publications: A keyword cloud is generated from all standard and user publications. It contains all keywords and weights them according to the number of frequencies (the font size is chosen according to the frequency).
- The list of standard, and user publications that cite a software: This is of considerable value; a large number of publications indicates high relevance. The authors of the standard publications are usually the software developers.
- An MSC classification of the software: For this purpose, the number of MSC classes assigned to the publications is determined. We assume that the most frequently mentioned MSC classes of the publications also characterize the essential mathematical contents and application areas.
- References to similar software and dependencies between software products: The common mention of software names and a match in MSC classifications in standard and user publications are indicative of a relationship between different software products. However, the nature of this relationship cannot be inferred from mere common naming without a deeper analysis. Frequently, however, reference is made to similar alternative software products.
- Current development status of the software: From the publication dates of the publications, the life cycle of the software is inferred, expressed by the S-curve cycle typical in business administration for describing the life cycle of a product. After publication (date of the first standard publication), the dissemination of the software begins, leading to increasing publication numbers. After the numbers stabilize, the number of publications gradually decreases and ends after development ceases due to new development or the use of alternative products, which is expressed in the decrease of publication numbers. A major reason for the end of software development is the constant development of the framework conditions for a software (hardware and software).

swMATH essentially provides the bibliographical data of a software. If the Web contains more information, e.g., the code or a documentation of a software, then the swMATH entries link to the following information:

- The web pages of the software: The web page of a software usually offers a detailed overview of the current version of a software, contains information about the content and goals of the software development, about the code or the license terms, about the hardware and software requirements, the installation, about the developers and providers, etc.
- Software Heritage: For developers, the software code is of particular interest, for example, further joint development. The Software Heritage service developed by

INRIA (France) is the world's largest archive for software codes, offering information on all versions.

- The code of SCIP is not open source: Instead of the link to the code, the swMATH entry gives the license terms in the swMATH entry of SCIP.
- Internet Archive: The various versions of a software's web pages are archived periodically by the Internet Archive and thus provide an overview of a software's development history. However, only the top levels of the web pages are usually publicly accessible.

The advantage of the swMATH approach is that the information can be mainly generated automatically from the zbMATH data and Internet sources. Of course, further efforts and activities are needed to support the infrastructure for this type of research data. In particular, introducing a standard for software citations, which has been worked on for years, opens new perspectives for improving the swMATH service. The swMATH database is enjoying increasing use, which is also reflected in the visibility of swMATH via the major search engines.

The concept of swMATH can also be used for other classes of research data. Currently, a prototype of a database for mathematical algorithms is being developed. A mathematical software implements a mathematical algorithm, although the distinction between algorithm and software is not handled uniformly, which does not simplify the automatic identification of algorithms.

The publication-based approach of swMATH is the first successful step towards the comprehensive information of all classes of mathematical research data and an essential pillar of the mathematical information of FIZ Karlsruhe.

In the future, the swMATH team will collaborate closely with the German initiative for Mathematical Research Data (MaRDI). By doing so, we bring software and other mathematical research data such as formulae, numerical data, models, statistical data, and interactive notebooks, among many others, closer together. Additionally, we participate in the task force for Infrastructure Quality Research Software of the Europen Open Science Cloud. Therefore, we build a community of infrastructure providers for research software not only within mathematics. Combining forces with institutions with similar intents ensures that swMATH is well connected to similar initiatives from other disciplines.