Information Services for Mathematics in the Internet (Math-Net)

Wolfgang Dalitz, Martin Grötschel, Joachim Lügger
Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)
Takustraße 7, 14195 Berlin, Germany
e-mail: dalitz@zib.de groetschel@zib.de luegger@zib.de

Keywords: electronic information and communication, Internet, World Wide Web, Math-Net, metadata, Dublin Core, software libraries, mathematical information, structure and quality of information

ABSTRACT

The present paper gives a brief description of the Math-Net project which is carried out by nine mathematical institutions in Germany, supported by Deutsches Forschungsnetz (DFN) and Deutsche Telekom. The project aims at setting up the technical and organizational infrastructure for efficient, inexpensive and user-driven information services for mathematics. With the aid of active (structured retrieval mechanisms) and passive (profile services) components, electronic mathematical information in Germany will be made available to the scientist at his workplace. The emphasis is put on information about publications, software and data collections, teaching and research activities, but also on organizational and bibliographical information. Decentral organization structures, distributed search systems as well as the use of meta-information (metadata) in accordance with the Dublin Core (hopefully) guarantee a longterm, high-quality repository of data. The well-known mathematical software and data collection *netlib* will be used as an example to illustrate how such a collection can be adapted to Math-Net. An integration of netlib into HyperWave offers additional perspectives and functionalities.

INFORMATION OFFERS FOR MATHEMATICS IN THE INTERNET

The *Deutsche Mathematiker-Vereinigung* (DMV) and the *Konrad-Zuse-Zentrum für Informationstechnik Berlin* (ZIB) have been developing, discussing and publishing ideas and concepts for a distributed system of information and communication for quite some time now [1]. Despite little support by official authorities (a corresponding program by the Federal Ministry of Education, Science, Research and Technology (BMBF) was published in mid-1996 — following years of discussion – and has not taken shape yet) quite a number of initiatives have been brought under way in German mathematics.

All mathematical departments and research institutes in Germany, the DMV itself, and many of its special interest groups have established WWW-servers. This also applies to several special research projects supported by the Deutsche Forschungsgesellschaft (DFG) and BMBF-supported research projects jointly carried out with partners in industry. These servers offer information of (still) varying quality and completeness on topics such as

- **Publications:** preprints and technical reports, PhD and habilitation theses, selected diploma theses in full text;
- **Software and data collections:** mathematical research software, test data collections, data collections from real applications;
- **Teaching:** announcement of lectures, courses, seminars and tutorials, lecture notes, collections of problems, course materials, degree programs, admission requirements, help and info pages;
- **Research:** research areas, lists of special research interests and projects, cooperations, conferences, workshops, talks;
- **Services:** libraries and online catalogues, bibliographical databases, computer and network facilities;
- **Infrastructure:** organization (structure, bodies, addresses, site maps), people (faculty, student and staff directories), personal pages, news.

Some examples of particular activities are the following. The DMV has been operating a fully reviewed electronic journal, the *Documenta Mathematica* [2], covering all areas of mathematics since May 1996. This journal is freely available in the Internet. On the basis of the Harvest [3] information system, the *Arbeitskreis Technik* of the DMV has set up a reference system for mathematical preprints and reports in Germany. ZIB has established a WWW-server for the International Mathematical Union (IMU), which includes a congress server for the International Congress of Mathematicians 1998 (ICM'98) taking place in Berlin in 1998. The *Zentralblatt für Mathematik* has set up a WWW-server for the European Mathematical Society (EMS), which contains an electronic library of currently about 30 mathematical journals in full text. In its electronic library *eLib*, ZIB maintains not only a mirror of *netlib* [4], but also an experimental meta library for netlib-oriented mathematical software libraries, furthermore, a virtual library for mathematical information resources called *Math-Net Links to the Mathematical World* with a multimedia demonstration component, the *Mathematical Museum.*

TOWARDS ORIENTATION IN THE NET

Offering mathematical information in the Internet is a very recent development. For the mathematically oriented reader, it is indeed a problem to find his way through the great variety of information. The Internet does in fact offer a number of subject-oriented virtual libraries that list digital information resources and that even review parts of the information material. There are also efficient search engines that enable access to practically all web servers in the world (on a keyword level). The former require visitors to spend a considerable amount of time navigating through various servers, the latter, however valuable they usually are, often fail to produce satisfactory results by giving too much information if search terms are very general, see e.g., [5].

Information found in the Internet often lacks coherence. One reason for this is the nature of the Web, as it does not contain back references, another reason are authors who do not adequately structure their contributions. Web pages are often linked to other pages to such an extent that the reader gets completely lost *(lost-in-hyperspace syndrome).* Even in hierarchically structured information offers the reader may lose orientation if there are too many levels involved. Frames have been introduced recently in order to support clearer structures. The drawback of these frames is that they hide the inner structure of a web server. Pages screened in this way will not be included in the reader's history, even not in that of an automatic reader. Frames seem to make web pages immune to robots.

Problems of this type are the main focus of the Math-Net Project supported by Deutsche Telekom and Deutsches Forschungsnetz (DFN) [6]. Mathematical departments of the universities Chemnitz-Zwickau, Cologne, Halle, Kaiserslautern, Munich (Technical University), Osnabrück, Paderborn, Rostock and the Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB) led by M. Grötschel are working together in the area of electronic mathematical information in Germany with the following aims:

- Enhancement and completion of local information services, structuring of information and classification by metadata;
- Automatic indexing for retrieval and searching (local, regional and national), with the possibility to limit a search within sub-categories;
- Efficient organization of distributed information offers, which should meet high standards in terms of quality and up-to-dateness;
- Setting up of electronic profile services with e-mail subscription;
- Integration of mathematical information resources from all over the world by interdisciplinary cooperation both on national and international level.

In the context of electronic publishing this involves questions of quality and up-to-dateness, authenticity and identity of documents and, last but not least, costs. The organizational, technical, but also the legal aspects of access and (longterm) archiving result in problems that have to be solved not only within mathematics, but throughout all branches of science (e.g., for PhD theses). Among these are questions of formats, storage media and techniques as well as authors' rights, which include problems of patent law and copyright. A considerable amount of data is of personal nature and therefore involves aspects of data protection, safety and identifiability.

Automatic Indexing and Regional Information Brokers

An important task within this project is the automatic indexing of locally offered information. The Math-Net project is based on the widely spread *Harvest System* [3], which offers approaches to this topic that even enable the treatment of metadata in HTML pages. The system itself consists of the two components *gatherer* and *broker*, which can be combined into a hierarchically structured network. In this hierarchy, information can be directed to higher levels and retrieved accordingly. On the basis of locally offered information repositories, mathematics in Germany (and beyond) can be indexed and made (re-)traceable.

Establishing a network of persons in Germany dedicating (part of) their time and work to the electronic information system is an important organizational task. For this purpose, taking regional criteria into account, information coordinators have been appointed, having the responsibility for local offers of information. They coordinate actions on a regional level and distribute general tasks among themselves. The tasks of information coordinators in their particular region are:

- **local:** Collection of various elementary repositories (WWW-servers, Harvest brokers on a departmental level), caching and mirroring.
- **regional:** Brokering of local indices, alternatively gathering where Harvest is not implemented locally.
- **national:** Brokering of regional indices, search in a central search engine.

Authors Set Up Metadata; Dublin Core

Appointing local information coordinators and establishing local and central search engines, however, only represent a partial solution to the tasks involved. Full text indexing, which is very common today, does not seem satisfactory in view of the sometimes gigantically high number of matches. Adequate document retrieval is only possible if adequate metadata have been recorded together with the data, see e.g., the special section of [5] related to this theme.

It seems canonical that meta information should be provided by the original creators of the documents, the authors. If metadata are set up by others, additional costs will be induced. Authors may be willing to create the neccessary metadata if the extra work is negligible.

In the area of electronic metadata the discussion seems to have advanced to a consensus on the most important categories and formats. The element set assembled under the name Dublin Core currently comprises 15 major metadata (such as author and title) as well as their electronic treatment and representation (e.g., in HTML pages) [7]. The electronic recording forms developed in this project (such as MMM – Mathematics Metadata Markup Editor) can now relieve the author from the syntactical details of the Dublin Core and hide them behind HTML/CGI forms. They reduce the additional work for making a document electronically available to a minimum. Some mathematics departments have started to use these electronic formsheets.

COLLECTIONS OF RESEARCH SOFTWARE, AN EXAMPLE

Sources of research code are stored today in software archives, so-called software libraries, in which single program components (modules) can be identified by means of an index. The best known and most comprehensive of these collections is the netlib created and maintained by J. J. Dongarra and E. Grosse, a library of libraries. It presently consists of more than 150 main libraries with approximately 350 sub-libraries, combining over 1.5 GB of software, test data and relevant documentation under one roof. The netlib is organized hierarchically. At present, its top index and the indices of its sub-libraries are structured according to a certain metadata scheme [4]. This includes a classification according to the GAMS index [8].

The netlib is organized in a central fashion. The sources of the 25,000 programs are stored in a single file system. In its current technical form it is hardly suitable for a distributed (decentral) organization of program libraries (the netlib is completely mirrored at several mathematical sites in the world). This particular form of organization of the netlib, as successful as it is, does in fact produce certain problems discussed below arising from the processes involved in the maintenance and coordination of sources.

In the case of research software, which is frequently modified, it is highly recommendable that users should always have access to the latest code an author provides. This requires not only that modified sources are integrated into the software libraries as soon as possible, but also that users are informed about changes at once. There is, however, research software, where even the central administrators are not informed immediately if a new version is available. Therefore, it might be a good idea to give authors the possibility to distribute their program sources themselves and to use some central "register and location" system. The Internet and the World Wide Web are appropriate means; nonetheless, new problems will arise eventually.

Another drawback of the present netlib index is, for example, that the single short descriptions of the components of a software library cannot be used "isolated", like index cards in a library catalog or a database. Therefore, the netlib administrator puts up an (internal) list of keywords that can be searched (this list, however, is barely integrated into and always lags behind the index texts). The netlib index is a "list of index cards", which are strung together in a single text file. The advantage of this special form of organization is, however, that this structure not only helped to make a retrieval of its single components by e-mail possible, but also made it relatively easy to make the transformation into the hypertext form, in which it is available today.

The line-oriented version of the netlib, e.g., was able to combine program modules as sources according to their (sub)routine call hierarchy. This form of software retrieval is very useful within the context of access by electronic mail, as it drastically reduces the number of mail actions needed to retrieve one full application. In the (new) times of fast networks the transfer of complete (integrated) software libraries through the World Wide Web has become common standard. In this respect, the module-oriented description and documentation of netlib libraries and software modules may be regarded as superfluous. But it is still useful, e.g., for representation in hypertexts or – as will be shown in the next section – for certain types of retrieval that can be combined with hypertext navigation.

Hierarchic Implementation of the netlib in Hyperwave

From the user's point of view it would be best to have a library system which offers both possibilities, (1) the retrieval of software modules by "classical" database means, and (2) a facility to "read" the index files and the software by means of a standard Web browser. It should also be possible to switch between both navigation modes. Thus, it would be helpful to a user, e.g., to locate a certain software module by employing the GAMS index scheme and then to learn more about this special piece of software by browsing its related "software environment", i.e., the index of the library it belongs to and the associated modules or documentation.

The basic idea of implementing the netlib in this way is to use HyperWave [9] to break down its index files into single components and, in this way, to make them searchable in a HyperWave server (and therefore in the WWW, see `http://elib.zib.de/netlib`). In this process, however, the connection between components and indices shall not be lost. One might say that this separation of single components yields index cards. These are (in HyperWave terminology) title lines, keywords and short descriptions of the single software modules and the deeper-lying sub-libraries (a recursive structure), which now can be browsed and searched simultaneously on each level of their hierarchy.

A short description contains, in the present form of its hierarchical implementation, explicit pointers (URLs) both to the superordinate index (i.e., the superordinate software library) and to the source of the corresponding module or the index of a sub-library. To the browsing hypertext reader, such index cards are not visible at first sight. He will primarily use index files which have been transformed into HyperWave structures and which are well known from the standard web implementation of the netlib.

The user may search them, however, according to their keywords and URLs and their components; this can be done by using the title, the textual (short) description or keywords in HyperWave-internal keyword fields. If a search is successful it provides access to the (hitherto) hidden module description, which in itself is the text component separated from the library index. In his search the user may also use parts of the netlib-directory paths (the program sources are available, e.g., in ZIB's public domain ftp-area) or parts of URLs. These have been integrated into the keyword field in a fragmented form. URLs are included explicitly – clickable – and in their complete form into the text, in order to enable not only (problem-free) viewing, but also printing.

The user of HyperWave can start a search at any point in the hierarchy, not only over the entire server, but also – and this is a particular strength of HyperWave – on the partial hierarchy below this point. This makes it possible for the user to structurally limit his search. He can exclude from his search everything that is above his actual position of navigation.

Following a search and the selection of a particular document (an index card) the user has now – thanks to explicit references – the possibility to switch back from the search modus into hyptertext navigation modus and to climb both upwards and downwards in the hierarchy of netlib indices and software sources.

A complete translation of the context of modules, in which the right and left neighborhoods are shown explicitly, would also have been possible; however, the current implementation does without that for economic reasons. The index files used for this purpose also give the required information (besides, the right and left neighborhoods are of minor importance in the case of software modules). More important would have been a call-reference map, which, however, has not been implemented in the current experimental version either.

One of the problems with the hierarchical implementation of the netlib is the permanent updating process. Weekly, sometimes daily, new software is added: modules and sometimes whole libraries are completed and updated or should be deleted. At ZIB a nightly cronjob runs a mirror-software, which obtains and synchronizes the updates from an archive within the original netlib. The mirror program not only builds a mirror of the netlib in the ftp-archive at ZIB, but also gives the required information where to carry out possible changes within the Hyperwave implementation of the netlib libraries. Therefore not only the programs are synchronized daily, but also all related indices, sub-indices and short descriptions (index cards).

Distributed Software Libraries

Because of their explicit references, the digital index cards of the HyperWave implementation of the netlib have in a way become independent of the position of the software they reference. They no longer have to have a fixed position in the hierarchy of netlib structures. As long as the implementation guarantees their up-to-dateness at any given time, they can be copied as often as required and may be used to define further views. This leads us to new potentials inherent to this kind of implementation and that have not been completely exploited by the actual implementation.

From the point of view of a software author, he is now in a position to maintain and offer his software locally (in his own ftp area) and to announce and distribute it centrally (in a HyperWave implementation of netlib situated elsewhere) as long as he adheres to the conventions for the netlib libraries (e.g., by setting up analog index files and descriptions). From the user's point of view it is sufficient if the indices of software collections of this kind are centralized somewhere (the corresponding software may stay remote) and the corresponding digital index cards are available, so that they are readable and searchable. In principle, the centralization can be implemented by an analog mirror mechanism, as is already employed in the mirroring of the original netlib. Software libraries of this type, which are not part of the netlib as such, should, however, be held in a separate meta library in order to avoid name conflicts with "original" netlib libraries.

It is remarkable that the authors of research software in this model write their documentation in the form of ASCII texts, just in the way they would adapt their program library for the traditional netlib. Transformation of this documentation into HTML or the hypertext language of HyperWave is done automatically, as is the splitting up into single digital index cards. Authors of software libraries are free to make their sources available simultaneously and locally, for instance by using the traditional netlib library software. Technically speaking, it would not require too much effort and a storage space to provide also a local mirror of the whole centralized system.

Research software that is made available through the netlib is usually classified according to the GAMS index. Under the conditions described above, this can be used to generate different views of the software made available in this way (either statically, by means of a pregenerated hypertext, or dynamically, according to search queries). Developing the first component is part of the present Math-Net project, the latter is already provided by HyperWave.

Documentation and index structures of software made available through the netlib can probably (and to a large extent automatically) be transformed into a Dublin Core HTML form. In this way, research software, which up until now has been described according to different attributes, could for the first time be searched and identified in the same context as preprints and other research articles (e.g. the documentation of the software itself). This development also belongs to the Math-Net project.

References

[1] W. Dalitz, M. Grötschel, J. Lügger, W. Sperber *New Perspectives of a Distributed Information System for Mathematics.* Newsletter of the European Mathematical Society (EMS), Part I: EMS Newsletter No. 13 (Sep 1994) 6-17, Part II: EMS Newsletter No. 14 (Dec 1994) 6-14

[2] *Documenta Mathematica* E-journal of the Deutsche Mathematiker-Vereinigung

[3] C.M. Bowman: *Harvest: A Scalable Customizable Discovery and Access System.* TR CU-CS 734-94, Dept. of Computer Science. Univ. of Colorado, 1994

[4] S. Browne, J.J. Dongarra, E. Grosse, S. Green, K. Moore, T. Rowan, R. Wade: *Netlib Services and Resources.* Techn. Report UT-CS-94-222, Univ. of Tennessee Comp. Sci. Dept., 1994

[5] *The Internet: Bringing Order from Chaos.* Scientific American, 3 (1997) 42 pp

[6] W. Dalitz, M. Grötschel, G. Heyer, J. Lügger, W. Sperber: *Informationsdienste für die Mathematik im Internet (Math-Net).* ZIB, Report TR 96-13, 1996

[7] S. Weibel, J. Goodby, E. Miller: *OCLC/NSCA Metadata Workshop Report.* Office of Research OCLC, Dublin, Ohio, 1995

[8] R.F. Boisvert, S.E. Howe, D.H. Kahaner: *The Guide to Available Mathematical Software problem classification system.* Comp. Stat. Simul. Comp., 20 (4) 1991, 811-842

[9] H. Maurer: *HyperWave: The Next Generation Web Solution.* Addison-Wesley, 1996