New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 55-63.

The Digitization Registry at SUB Göttingen A Step towards a DML Registry

Thomas Fischer

SUB Göttingen

In 2002, John Ewing, then executive director of the *American Mathematical Society*, suggested to consider the effort of digitizing the mathematical literature: "The goal was to create a virtual library containing much of the past literature – a library that could eventually grow into a 'World Mathematics Library' "[1] A basic calculation showed that this was an enormous task but essentially feasible, if approached as an international endeavour of mathematicians, librarians and publishers. This is not the place to recall the emergence and development of the Digital Mathematics Library in full detail, but since this is at the heart of the considerations concerning the Digitization Registry, I will try to give enough background information and references to put this endeavour into perspective.

The U.S. National Science Foundation (NSF) funded a two-year (2002–2003) planning project coordinated by Cornell University Library "toward the establishment of a comprehensive, international, distributed collection of digital information and published knowledge in mathematics" with the goal "to make the entirety of past mathematics scholarship available online"[2]. This was extended up to October 31, 2004, to provide means for the transgression of the project from Cornell to the *International Mathematics Union*, where it will be run under the auspices of the *Committee on Electronic Information Communication*, the new website being at `http://www.wdml.org/`.

An international steering committee was formed and directed the process of formulating standards for the planned digital library, which where formulated in several working groups. Planning meetings in 2002 (Washington) and 2003 (Göttingen) laid a groundwork on principles and standards applied to the digitization of legacy text and the presentation and availability of born digital material. The minutes of these and related meetings are available from the Digital Mathematics Library website and provide links to further results, leading to

[1] John Ewing: Twenty Centuries of Mathematics: Digitizing and Disseminating the Past Mathematical Literature. Notices Of The AMS, August 2002 (`http://www.ams.org/notices/200207/fea-ewing.pdf`)

[2] Digital Mathematics Library, `http://www.library.cornell.edu/dmlib/`

the meeting in Stockholm. The final report on the project will be due by the end of October 2004.

A preliminary draft of the final report is available form the Cornell project site[3] where the reports of the working groups are published as well. There is a general consent on the technical basis of the digitization and archiving process, and first considerations on the metadata to be used are agreed upon. The basic problem remains the financial one. Early hopes for funding by the NSF and the European Community did not come to fruition, and no other general funding agencies seemed to be on hand.

Since there was no general world-wide funding available, the practical activities took place on a national level in the participating countries. So the slowly emerging *World Digital Mathematics Library*, renamed last year to stress the world-wide effort (and to avoid a name conflict) consists of diverse projects, which are usually funded on a national or bilateral basis and have focussed either on the national heritage or on specific sections of relevant mathematical literature. Some prominent examples are the following.

NUMDAM

The French project *Numérisation de documents anciens mathématiques* (`http://www.numdam.org/`, not quite appropriately translated as "Digitization of ancient mathematics documents", since the documents are predominantly from the second half of the twentieth century) has digitized six of the most eminent French mathematical journals and made them freely available on the net, using a "moving wall" system to restrict access to the most recent issues: the Annales de l'institut Fourier, the proceedings of Journées Équations aux dérivées partielles, the Publications mathématiques de l'IHÉS, the Bulletin and the Mémoires de la SMF, the Annales Scientifiques de l'École Normale Supérieure.

ERAM

The *Electronic Research Archive for Mathematics* is also known as "The Jahrbuch Project" (`http://www.emis.de/projects/JFM/`). This German project is centred around the "Jahrbuch über die Fortschritte der Mathematik", a predecessor to the "Zentralblatt der Mathematik", which appeared from 1868 until 1942. The *Jahrbuch* itself was keyed in and is now available as a database in the same format as the *Zentralblatt*, but also the most important articles refereed where digitized and are now available through the *Göttinger Digitalisierungs-Zentrum* (GDZ), so creating a "digital archive of the most important mathematical publications of the period 1868–1942".

The development of a distributed digital library of mathematical monographs

This joint project by the University of Michigan, the State and University Library Göttingen and Cornell University provides a virtual collection of the distributed collections available in Ann Arbour, Göttingen and Ithaca. Funded by the US National Science Foundation and the Deutsche Forschungsgemeinschaft, this project does not provide additional digitisations,

[3] See `http://www.library.cornell.edu/dmlib/DMLreportJun29draft.pdf`, where the other reports are linked in as well.

but enhances the accessibility of existing ones. Portals are available at the University of Michigan[4] and at Cornell University[5]. Depending on the preparation of the material, search on the metadata level or full text search is available.

JSTOR

JSTOR – The Scholarly Journal Archive – was established as an independent not-for-profit organization in August 1995 and has the dual mission to create and maintain a trusted archive of important scholarly journals (`http://www.jstor.org/`). The project is based at Princeton University in Princeton, NJ; University of Michigan, Ann Arbor, MI; and at MIMAS in Manchester, UK and has a wide scope beyond mathematics. But the mathematical books and articles can be viewed as a special collection: "JSTOR's Mathematics & Statistics Collection unites 30 titles in the mathematical and statistical sciences from existing JSTOR collections.

EMANI

The *Electronic Mathematics Archiving Network Initiative* (`http://www.emani.org/`) brings together university libraries (Cornell University Library, SUB Göttingen, Cellule MathDoc and Tsinghua University Library) and publishers (Springer and EMIS, the European Mathematical Information Service) to organize reliable archiving of born digital mathematics as well as reliable online access. "EMANI now offers access to 100 mathematical journals from various publishing companies .... As well as digitally produced journals, these also include retro-digitised journals, such as "Mathematische Annalen", which have been published by the scientific publishing company Springer-Verlag since 1869".[6] The EMANI website offers an impressive list of journals provided by the different partners, either already digitised or in the process of digitisation. Obviously, this diversity was hard to keep track of. In the "Notices of the American Mathematical Society", Allyn Jackson published an article "The Digital Mathematical Library"[7] on the Göttingen meeting of the steering committee of the DML project, which summed up the development up to that stage and gave the project a wider publicity. This article also included a list of almost one hundred retro-digitized journals available from different projects and institutions and called for "some kind of centralized access". Since then, Ulf Rehmann from Bielefeld University started a website presenting the different books and journals with an interface allowing alphabetical browsing by title and author[8], and Steven Rockey from Cornell University Library keeps a list of *Mathematics Digitization*[9].

In spite of these efforts, it became increasingly difficult to find out which project was planning to digitize which papers, was in the process of digitization or had already finished it, and if

[4] `http://www.hti.umich.edu/cgi/t/text/text-idx?c=mathcornell;c=mathgoettingen;c=umhistmath;g=mathall`

[5] `http://mathbooks.library.cornell.edu:8085/Dienst/UIMATH/2.0/Search`

[6] `http://www.springeronline.com/sgw/cda/frontpage/0,11855,5-109-2-103447-0,00.html`

[7] Notices of the AMS, (50) No 8, 2003, `http://www.ams.org/notices/200308/200308-toc.html`, also available at `http://www.wdml.org/publications/comm-jackson.pdf`.

[8] `http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_links.html`

[9] `http://www.library.cornell.edu/math/digitalization.php`

the paper was finally available and if so, under which conditions. This is when the idea of a *Registry* was gaining momentum, a central hub that would allow to answer questions of this kind by querying a central database. Actually, different registries with different objectives and focuses are planned or already in operation; I will focus on the *Göttingen Digitization Registry*, which is somewhat modelled after the EROMM database described below.

Background: EROMM

Göttingen has already a history of providing registry services to the library community. In 1994, a database for the *European Register of Microform Masters* (EROMM) was set up in Göttingen for a consortium of four national libraries in Europe. The goal was to provide a registry of microforms (e.g. microfiches, microfilms) produced in Europe and integrate this with information available from around the world: "The copy of a work that has been reformatted in one place according to agreed technical standards must not be reformatted again anywhere else. To prevent duplication of effort two things are necessary: Information on the existence of the reformatted item and availability of a service copy."[10] More partners from different countries joined the project over time and made this the central European reference of reformatted editions. In the last years, the scope was broadened to encompass digitized material as well. Although the access to the database requires identification either individually or through the IP address, the use is free for citizens of the participating countries. The EROMM database allows researchers to effectively find these sources and eventually obtain them from the organization holding them, usually through an integrated order form.

Actors and objects in the digitization process

The basic questions for the set up of the registry are its *scope* and the *data model* used. The question of scope is essentially answered by the goal of this registry, to give reference to all objects digitized or otherwise available under the umbrella of the *World Digital Mathematics Library*, that is, it should contain all documents which are considered as part of this effort by their producers; there is no formal membership to the WDML. Like the EROMM project, the first goal would be to avoid duplicating efforts, so for this it would be sufficient to register the journals or books which are digitized, giving additional information if not the whole volume of a given year or not the whole lifespan of the journal is being digitized – single articles are rarely considered in large digitization project.

But again the second goal is to make the digitized articles readily available, and for this the granularity of the data structure has to be much finer, at least going down to the level of the single article or probably the chapter of monographs. Only this would allow to search in this virtual collection for papers by author and title, or even support the search for keywords or classifications. This would obviously be a very desirable support of the mathematical research process, given the experience that mathematical results have an extremely long "expiration date" compared with science – even papers more than hundred years old are regularly cited in contemporary mathematical articles.

This gives a basic idea of the scope and the necessary information to be stored in the registry. But closer inspection of the actors in the digitization process is useful. The digitization

[10] See `http://www.eromm.org/e-info-e.htm`

occurs usually in the context of some project, funded by some agency and based at one or more institutions. These projects are usually based at one or more institutions (universities, libraries, research institutes), and may develop over time, adding or loosing some partner, changing scope etc. In the spirit of the World Digital Library, this is information which may be useful later on when the objects are to be retrieved, and problems occur or legal issues have to be addressed.

The objects to be digitized are either monographs (probably part of a series like the *Grundlehren*), series of books (e.g. *Encyclopaedie der mathematischen Wissenschaften*) or Journals. All these objects tend to have their particular history, being printed and reprinted, changing editors, publishing companies or names. While again the immediate benefits to the individual researcher are not obvious, the idea to collect the complete mathematical literature will require to provide as much information as possible to fit together the pieces of the puzzle that constitutes the mathematical heritage.

For this, we introduce a data model that allows for separate datasets for the different entities involved: for institutions, projects, authors, agencies etc. The Göttingen Digitization Registry provides containers for the following data formats:[11]

INS: Institution

PER: Person

PRO: Project

SER: Serial Publication (of books or journals)

MON: Monograph (of one or more volumes)

JOU: Journal

ISS: Issue

ART: article

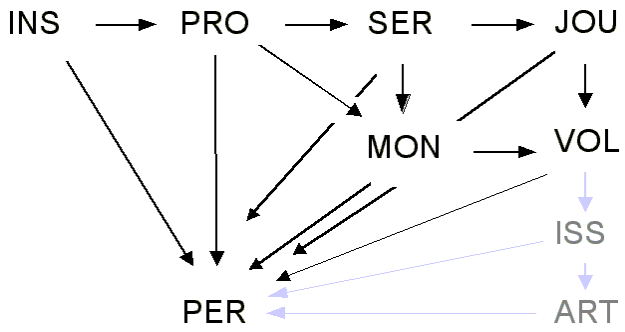
CON: conference proceedings

COL: collection of any material

Although implicitly present, these data formats are not all used yet. The relevant information needs to be collected, and the data have to be connected through a system of unique identifiers. This requires either extensive manual labour (and respective funding) or a sophisticated system for the exchange of information.

A slightly simplified visualization of the relationships among the different data types would look like this:

[11] A more complete (and more technical) description of these datasets is available at the Digitization Registry website, `http://www.sub.uni-goettingen.de/ssgfi/digreg/`.



The graphic shows that the data type *Person* (PER) may appear in the context of projects and institutions (usually as contact person or such) as well as in the context of publications (usually as author or editor). For now there seems to be no need to differentiate the different roles, given that the respective persons are uniquely identified. Additional care is taken in trying to keep track of different dates, e.g. the publisher of a journal may change over time, and rights be held by different corporations. So many fields allow for additional time (range) information. For search and retrieval the relevant information will be joined appropriately to provide as much information as possible without overloading the display. Eventually, the display of the data may be customizable. Note that the entries for Issue (ISS) and Article (ART) are marked in grey, indicating that this information is not present in the Digitization Registry yet. The underlying reason is a problem with the unique identification of the individual articles. While this problem is partially due to the procedure and system used by the GDZ in the ERAM project, the basic question is of general importance: how are the entries to be identified? There are different options available, which open up different possibilities. On the one hand, there are the established identifiers for the printed material, like the ISBN and the ISSN. In the library context, there are additional identifiers, e.g. the Pica Production Number PPN. And in the mathematical community, the are at least the numbers from Zentralblatt and Math Reviews. And to make things more complicated, different editions from the same book will have different identifiers, but will usually be indistinguishable as digitized works. To provide an example, consider Robert Switzer's Algebraic Topology. This was published by Springer in 1975 as volume 212 in the series "Die Grundlagen der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete". But since Springer published this in Germany and the United States, two different ISBNs where assigned to it, as well as codes from Zentralblatt and Math Reviews respectively. In 2002 Springer published a reprint of this book, which was assigned another ISBN and new codes from Zentralblatt and Math Reviews as well. Still, for the purposes of the working mathematician, this is just one single book, and there is no reason why the different versions should be digitized separately. Apart from these identifiers, the are other systems geared towards the digital versions, like URN and PURL as stable identifiers for digital objects on the Web. DOI and OpenURL are used e.g. by CrossRef[12] to provide citation linking, a service of high importance to mathematicians. It is not clear at this point in time, where and how these different identifiers will be administered and connected to provide direction to the appropriate digital object.

[12] `http://www.crossref.org/`

Data exchange: OAI

To connect the different projects contributing to the World Digital Mathematics Library to a unified virtual repository, the scattered data have to be collected, standardized and indexed to provide an effective method for retrieval. Different projects exhibit their data in different ways, and (for now?) the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH)[13] seems to be the preferred method for the exchange of the relevant data. This protocol allows the complete collection of data at the different repositories as well as the incremental collection of the data sets changed recently. The required format for the presentation of the datasets is based on Dublin Core, but an OAI data provider may choose to exhibit the data in several other formats, the protocol offers a mechanism to ask the provider for available formats. The Open Archive Initiative presents a number of tools for the creation and testing of OAI services, among them an experimental browser for OAI repositories.[14] The OAI method is used by a number of universities and mathematics institutes to expose research papers to the interested community. The *Experimental OAI Registry*[15] at Grainger Engineering Library Information Center at University of Illinois at Urbana-Champaign provides an extensive listing with a search interface, and searching for "mathematics" yields way over hundred hits, among them well known preprint services like arXiv, but also some digitization projects. NUMDAM[16] serves as an OAI data provider and exhibits six sets of data, one for each journal available. There are also two different metadata formats available, the standard Dublin Core and another called *minidml*, I will presently return to this issue. In a similar fashion, the Göttinger Digitalisierungszentrum (GDZ) employs the protocol used in the Mathematical Monographs Project to exhibit their collection.[17] In this case, all of mathematics is contained in one huge set, and there are two non-standard metadata formats available: ProPrint and CGM. The set-up at the GDZ is interesting in so far as the same interface provides additional services, like a search which usually would not be covered by the OAI-PMH. The same interface is used by the Cornell University Library for their Historical Math Monographs collection included in the Mathematical Monographs Project, providing the additional format oai_rfc1807. Cornell uses this system also for an OAI interface to Project Euclid[18] with the same metadata formats and more than 30 sets for the mathematical journals published by Euclid. The OAI harvesting options are used already by several service providers, e.g. the *Bielefeld Academic Search Engine* BASE[19]. This project uses advanced modern search engine technology to search in an index of data collected from different sources, including many digitization projects. In France, Cellule MathDoc sat up an inventory of retro-digitized books, Livres Numérisés Mathématiques (LiNuM)[20], that presents not only the French collections from Gallica, but also among others the books available from Cornell, Michigan and Warsaw Universities and Göttingen State and University Library. A similar project for articles is called *Mini-DML* and will be described elsewhere in this volume. This shows that the given

[13] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[14] http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai/
[15] http://gita.grainger.uiuc.edu/registry/
[16] http://www.numdam.org/oai?verb=Identify
[17] http://gdz-srv2.sub.uni-goettingen.de:8091/cgm/servlet/dienstreq?verb=Identify
[18] http://ProjectEuclid.org/Dienst?verb=Identify
[19] http://base.ub.uni-bielefeld.de/help_search_english.html
[20] http://math-sahel.ujf-grenoble.fr/LiNuM/linum_en.html

information is already useful for resource discovery. Unfortunately, for the building of an effective registry it is not sufficient.

Data structure: Towards a DML application profile

The reason for this is that the universally available Dublin Core metadata are not rich and rigid enough, and more comprehensive metadata sets are not generally employed. Not rich enough means that many Dublin Core data sets only include the name of the author, the title of the work and a (probably local) identifier for the object. Since in addition, names and titles may have undergone some transliteration passing from one language to another, the proper identification becomes quite involved. Furthermore, these data usually do not include a link to the webpage that makes the article available; the given identifier and an analysis of the project website may allow to construct a request that displays the desired document, but a standardized method is not available. The lack of rigidity of the Dublin Core format refers to the lack of standards of how given fields are to be filled in. For example, the "last name first" rule is quite standard in the cd.creator field, but adherence to it is not guaranteed, and the rule is definitely not sufficient outside the middle European/Anglo-American context. In addition, one may find the years of birth and death in this field, separated by commas from the name. Since that is decided by the different repositories individually, this is hard to analyze automatically. Another example is given by the dc.publisher field: independent of the original publisher of the work, the Cornell collection gives "publisher: Cornell University Library", which is not helpful for the identification of the work.

The other formats used by these repositories are not necessarily much more useful: rfc1807 used by Cornell for the Historical Math Monographs is a more bibliographic format that gives the number of pages and something like the original publisher in addition to the DC information, CGM used by the GDZ contains only standard DC information, only ProPrint adds page count and separates first name and last name into different fields.

As the name already suggests, the minidml format offered by NUMDAM is something quite different and might well serve as a starting point for the development of an application profile[21] for the data exchange among the participants of the WDML. The minidml format is used by NUMDAM for the description of journal articles. In addition to the standard DC fields, minidml provides

- citation information: one field with the full citation, and additionally the separate parts in individual fields, including the page range, not only the page count,
- the URL of the article on the Web,
- information on the language of the article and the file formats available,
- information on the digitization project,
- the identifiers from Zentralblatt and Math Reviews, if available.

With this information, a registry could provide links to the desired objects, resolving the identifiers from Zentralblatt and Math Reviews and point to the preferred format. Only minor modifications would be necessary for monographs, most notably a field for publisher

[21] For application profiles, information is available from a working group at the European Committee for Standardization CEN. See `http://www.cenorm.be/isss/cwa14855/` and the references cited there

identification. Probably a more elaborate format than the "last name first" could be chosen for the presentation of the author, and it would be desirable to formulate datasets for the individual journals.

Another interesting example are the metadata used by Cornell for the Project Euclid. Although the names are the same, the contents of the metadata in Dublin Core as well as rfc1804 format are quite different from those used for the Historical Math Monographs. In addition to the standard fields, this version of rfc1804 provides keywords, a description or abstract, and the date of publication. The Dublin core version offers the same information, and also the language, the dc.type and the MIME type of the document, but most importantly, three dc.identifiers: the internal identifier of the dataset, the URL of the document on the Web, and a citation for the article. This amounts to essentially the same information as the one provided by the minidml format, albeit in a somewhat less organized form. Although this last example proves that all essential information can be delivered in a Dublin Core package, as general agreement on a more rigid and comprehensive format would be highly desirable. The Euclid Dublin Core and the minidml can serve as excellent starting points and it is hoped that the WDML Metadata Working Group will publish an application profile for digitization projects along these lines.

The Göttingen digitization registry: Present and future

The Göttingen *Digitization Registry* is in operation and available on the Web[22]. The database references about 600 monographs and 100 journals as well as some projects, institutions and serial works. All entries are linked to the appropriate websites so that the digitized objects are readily accessible, searching as well as an index scan is available for the important registers. At the point of writing, the integration of the mathematical holdings of the *Göttinger Digitalisierungszentrum* is under way, including hundreds of monographs and tens of thousands of articles. Every article will be assigned a Pica Production Number which will become the stable reference for this article, used in the Digitization Registry as well as the Jahrbuch data base. This encompasses also the full integration of articles with the chain of relations leading from the article via issue and volumes to the main entry of the journal, including the relationships of inheritance needed. That requires some fine-tuning of the data model employed and will be completed in the first months of 2005. The next step is an update of the present information of all entries from the projects already included, probably already combined with the next goal: the setting up of a OAI service and data provider. The *service provider* will collect the available information on digitized articles, possibly after consultation with the service providers on the available data formats, and the data thus gathered have to be incorporated into the database. The *data provider* will exhibit the data present in the Göttingen Digitization Registry to the other interested parties, offering Dublin Core as the standard data model. For more advanced exchange of data, a common data format for the mathematical community will be required, probably along the lines mentioned above. This will be implemented as soon as it is available.

Received December 30, 2004

[22] Address: `http://DigReg.MathGuide.de/` or `http://digreg.mathguide.de/`