Digitisation, Representation and Formalisation

Digital Libraries of Mathematics

A. A. Adams[*]

School of Systems Engineering, The University of Reading.
A.A.Adams@Rdg.ac.uk

Abstract. One of the main tasks of the mathematical knowledge management community must surely be to enhance access to mathematics on digital systems. In this paper we present a spectrum of approaches to solving the various problems inherent in this task, arguing that a variety of approaches is both necessary and useful. The main ideas presented are about the differences between digitised mathematics, digitally represented mathematics and formalised mathematics. Each has its part to play in managing mathematical information in a connected world. Digitised material is that which is embodied in a computer file, accessible and displayable locally or globally. Represented material is digital material in which there is some structure (usually syntactic in nature) which maps to the mathematics contained in the digitised information. Formalised material is that in which both the syntax and semantics of the represented material, is automatically accessible. Given the range of mathematical information to which access is desired, and the limited resources available for managing that information, we must ensure that these resources are applied to digitise, form representations of or formalise, existing and new mathematical information in such a way as to extract the most benefit from the least expenditure of resources. We also analyse some of the various social and legal issues which surround the practical tasks.

1 Introduction

In this paper we present an overview of the use of information technology in mathematics. Some of the earliest uses of computers was the automation of mathematics. Babbage's Difference and Analytical Engines [Bab23] were designed to engineer tedious, error-prone, mathematical calculations. Some of the first digital computers were designed for the purpose of breaking encryption codes. In the early days of digital computing, much was expected of this mechanical revolution in computation, and yet so little appeared for decades. As with the Strong AI and Ubiquitous Formal Methods communities, early over-selling of the idea of mechanised mathematical assistants led to a dismissal of the whole idea by many in the mathematics community.

This raises a side question as to who is part of this "mathematical community". Athale and Athale presented a modest categorisation of this community in [AA01]. In the sense of developing tools and techniques for mathematical knowledge management(MKM), we must be as inclusive as possible in defining this community. The narrower our definition, the less use our work will be, and the more likely to become a sideline in the progress of information technology. That is not to say that the MKM community should not define specific tasks with narrow, achievable goals. Far from it, but for each of these goals, the needs and milieu of the whole mathematical community should be considered before fixing on the specifics of a project. Some projects will, naturally, only be of interest and benefit to a specific small group within the mathematics community. Even these, however, will benefit from a broad strategic overview being used to inform their development. The general and the specific must

[*] This work is supported by EU Grant MKMNet IST-2001-37057 and UK EPSRC Grant GR/S10919

always be kept in balance when developing technology. Lack of this overarching viewpoint has led to the fragmentation of mathematical tools we see currently. This paper presents a wide-ranging view of who the users of MKM software might be, their needs and what approaches appear fruitful for further exploration.

We will begin by considering the various levels on which information may be represented on computers: Digitisation (for display) in Sect. 2; Representation (purely syntactic detail) in Sect. 3; and finally Formalisation (both syntax and semantics included) in Sect. 4. Once these categories of mathematical information have been described in detail, we will turn our attention to the uses of this information: in Sect. 5, to the dissemination of mathematical information to the various types of user; in Sect. 5.1, to the methods of dissemination historically and to the present; to some ideas on future directions in Sect. 5.2; lastly to legal and social issues effecting MKM in Sect. 5.3. In Sect. 6 we survey prior efforts in aspects of MKM, many of which form the background to the work of current researchers in the community, and which will form background or components for our developing technologies. We finish with some concluding remarks in Sect. 7.

2 Digitisation

New mathematical material is being produced by digital means. In many subjects this would lead to knowledge management specialists for that field discounting the necessity of spending much effort on bringing the older non-digital material into the digital world. In mathematics, however, the effort would not be mis-placed. Even given the ever-increasing rate of publication of new mathematics (estimated to be rising exponentially at present: the total amount of mathematics that has been published appears to be doubling every ten years). This compares with the science in general where it is estimated that only the number of new works published every ten years doubles (as opposed to the total published) [Kea00]. Mathematics is a very hierarchical subject. It is rare that new work completely supersedes older work. Unlike physics, for example, where new theories expand upon and quite often replace older work, leaving the older to languish in the annals of the history of science, older mathematics is rarely completely replaced. Newer proofs may be shorter or more elegant, even appearing to approach some divine perfection [AZ99]. However, when a new area of study opens up it is often the case that older proofs, in their less refined and honed state, may give more insight into approaches, than the newer subject-specific ones. So, mathematics, and mathematical knowledge, is rarely, if ever, superseded. See [Mic01] for an interesting look at a specific digitisation effort.

Given this idea, that older mathematics may be as useful as the latest up-to-date material, it is imperative that efforts at MKM do not begin from 1962 (LISP 1.5 [McC62]); 1979 (TeX [Knu79]); 1988 (Mathematica [Wol])... and work forward with existing represented and formalised mathematics. (In Sect. 3, we will discuss the work on automatic translation of digitised mathematics texts into a representational format.) Seminal works may still go out of print, or be available from small publishers and almost unobtainable, try to buy a copy of [ML84] for instance. University libraries periodically cull their stock, and printed matter is always susceptible to accidental destruction. Material is still subject to copyright and, as we shall discuss in Sect. 5.3, the copyright holders may not even be the original authors. For instance, I am reliably informed that photocopies of the difficult-to-obtain (supposedly still in print) [ML84] are available. While this may be true for widely-acknowledged seminal works, it is less likely to be true for a wide range of material, languishing in the hands of uninterested publishers refusing to relinquish rights gained from highly biased contracts. A project such as Project Gutenberg [Har71] should be created as part of a large-scale investigation into MKM, not to transcribe works of mathematics, but simply to capture their graphic essence, to guard against loss and degradation of the physical originals.

That is not to say that digital materials are themselves completely proof against loss and degradation. [JB01] includes many details on the general issues, following research by and for

The British Library on digital collections. Preservation of material in digital form requires a different approach to the preservation of physical material. Some issues are easier to solve for digital copies than physical ones: digital copies, when made carefully, are identical to the original and further copies of the copy also do not degrade. Other issues are more tricky in the digital world: language changes over time may make the "format" of a physical copy difficult to process, but this usually takes centuries to become a serious problem (except in the case of rapidly disappearing languages), whereas the formats of binary files may change very quickly and backwards compatibility between formats may not be preserved even for years, let alone decades or centuries. The physical form of the digital data must also be considered in terms of potential damage or outdated storage mechanisms. The University of Edinburgh, for instance, donated their last punched card reading machine to a computer museum in the 1990s. A year or so later the University had to ask for it to be loaned back because a number of members of the Physics and Chemistry department had data that would be expensive to reproduce stored on punched cards in filing cabinets. CD Roms have only a limited shelf life and although backwards compatibility with CD Roms is currently being maintained this may not be the case for decades longer. We will discuss other issues related to this in the section (3) on represented mathematics.

Current work on the digital display of mathematics focusses on how to efficiently transmit mathematical data which is already digitally represented, and how to coherently display that mathematics to capture the intent of the author while fitting with the restrictions of the reader's display method. While an issue for MKM, this is much less important (and recent work has been very successful at addressing these issues) than the reverse engineering one of taking displayed mathematics, commonly a mixture of consistent and inconsistent printing formats, and creating a digital archive, which may then be amenable to further processing, which we now consider.

3 Representation

The problem of cross-communication between differing systems of formalisation of mathematics has received a moderate amount of interest over the last decade or so. We will address this issue in detail in Sect. 4. In this section, we concentrate on the gap between formalised mathematics and digitally displayed mathematics, which is the simple representation of the syntax without necessarily linking this to an underlying semantics. Davenport's overview [Dav01] provides useful background to both this and the following section.

Typeset mathematics comprises two primary types of text: the mathematical text and the explanatory text. The richness of the combination is one of the causes of the massive expansion factor involved in fully formalising a published mathematics text. Modern mathematics may be typeset in a variety of systems. Some of these systems include some level of formalisation supporting the mathematical text (e.g. in a Mathematica Workbook). Others, such as LaTeX [Knu79, Lam94] or Display-MathML [Fro02], only contain typesetting information which may be partially converted to a formal semantic scheme. (We will consider representations which include fully or partially formalised semantics in Sect. 4.) There are two primary sources of represented mathematics: papers produced by mathematics users/producers for which the original encoding is available (as opposed to merely the graphical output); and scanned mathematics papers. We will consider these two types of information separately as there are some very different issues to consider. After this, we will return to the connection between the mathematical text and the explanatory text.

3.1 Representation from Digital Encoding

Some interesting work has been done on recovering some of the author's *mathematical intent* from the printed or digital graphic format of published mathematics. A very interesting paper is that of Fateman et al [FTBM96]. The main issues in this topic are the recognition of symbols in the rich language of mathematics, and the parsing of the two dimensional presentation used in typesetting. Take the simple instance of integration of real-valued functions. An integral presented in typeset form might look like this:

$$\int\limits_{a}^{b} f(x)\mathrm{dx}$$

or this: $\int_p^q f(y)\mathrm{dy}$. These may have identical semantics, yet are displayed somewhat differently. Mechanical parsing of these two-dimensional graphical languages poses problems that have received limited attention from the optical recognition community (see comments in [FTBM96] for details. Since recovery of representational material from existing graphical representations (whether scanned or original digital documents) is important to the field of MKM, it is up to us in the MKM community to seek collaboration with those working in the area of parsing of graphic images to develop this field.

Without a suitable underlying representation language in which to embed the translations, however, the utility of parsing techniques for graphical images is entirely pointless for a large scale effort at enabling access to a wider corpus. We will return to this point in Sect. 4.1, where we will consider current systems which may be good candidates for such representation.

3.2 Representation in the Primary Source

Representations which are available with the source code (e.g. a LaTeX file) are already in a form suitable for automatic parsing to add in some level of formalisation. Even a veneer of formalisation, providing it is accurately and automatically produced, can be highly useful in the task of MKM. However, challenges still remain. Authors use their own style files and redefine commands. Some intelligence must be applied to the parsing of even a well-written LaTeX file, since the aim of the author is to present the text clearly in the graphical output to a human reader, rather than to produce machine-parseable code. This same issue attaches to other methods of typesetting mathematics to a greater or lesser degree: word processor files using an equation editor; papers written using a Computer Algebra System (CAS); etc. This issue is also related to that of the explanatory text, and its relationship to the mathematical text. While inline formulae are obviously connected to the sentence that surrounds them, English is full of hanging anaphora, even in the most clearly written technical paper. Cross-references to equations and lemmas buried in other parts of the paper, in other papers, and inside floating tables and figures, further cloud the issue. Even with the abstraction layer of `\label` and `\ref` in a LaTeX file, such internal links may be obscure and are almost certainly incomplete.

Some very good initial work on these kinds of problems can be found in [BB01], which looks at the problem of developing new documents from a variety of originals, through the use of text-slicing. Some automation is present, but most of the ‘”meta-information” (information about the document) is hand-crafted. Beyond this initial input of meta-information, however, there is an automated layer which ties a variety of documents together in a coherent form. The purely practical issues of converging a variety of sources has also been considered in this project.

3.3 Representation and Explanation

The relationship between the natural language parts of a mathematics text and the symbolic content is further complicated by a variety of nomenclatures and symbol sets. The older the mathematical text we wish to represent in a manipulable form, the wider the deviation is likely to be from current standard usages. Given that this is already a problem requiring the application of intelligence to understand the context of the pure mathematical text, it is unlikely that current automatic tools will be able to solve this problem. Automatic cross-referencing within a paper can help but we must accept the limitations of current technology and not expect at this stage too fine grained an automatic analysis of maths texts.

One area which should be relatively easy within this area, however, is to cross-link documents both forwards and backwards via citation indices. Each mathematical document in a digital mathematics library should have an associated set of links pointing to papers which reference it and which are also available in the library. Likewise links to the papers referenced from a work should be present, both where they appear in the text and from the bibliography section at the end. See Sect. 5.4 for further discussion of cross-referencing to external items.

3.4 Maintenance of Archives of Represented Material

As previously mentioned, the mere fact that data is digital is not a guarantee of its incorruptibility. Some basic precautions, such as adequate backup facilities, are obvious, and the process of mirroring data worldwide is another step in the right direction. However, the subject of maintaining digital archives is a current topic of much discussion amongst library scientists, and the MKM community should take note of such developments. Beyond simple data hygiene and archival storage of snapshots of data, there is a definite and growing problem in formalised mathematics that the amount of effort needed to keep formal developments up to date with the growing, and therefore changing, capabilities of the formal systems is increasing. If the required effort increases beyond the funding available for maintenance on such projects then there is a substantial risk of the prior work degrading, possibly degrading beyond the point where it is quicker and cheaper to recreate the lost data when it is needed than to try and bring it up to the state of the art.

There are some interesting developments in this area coming from the NuPrl group at Cornell University. Growing from a perceived need to maintain large formal developments of abstract and applied mathematics with their own systems, the NuPrl [CA+86] group developed a “librarian” architecture which maintains not only the current proof status of objects with reference to the “version” of the system, but also attempts to patch proofs that fail with new versions using automated deduction techniques. The system also maintains records and copies of the older versions of the system in which earlier developments are valid, and documents that changes between versions. This provides a framework for keeping developments valid with as up to date a version of the system as is feasible given limited resources. Following on from this development, the Office of Naval Research has funded a project at Cornell to produce a “library of digital mathematics” to support “the production of mathematically proved programs”. This has consisted of integrating various other theorem proving and some other specialist mathematical systems within the same “librarian” framework. In addition, a fair amount of work has been done on reconciling the differences between the underlying logics of these various systems (HOL, NuPrl, MetaPRL, PVS, Coq [BB+96], Isabelle [Pau88]. . .) where possible, and on identifying the difference where not. Thus, a theorem proved in one system may be identified as compatible or incompatible with the basic logic of another (such as the incompatibility derived from one system being classical and another constructive in nature).

The development of mathematical software has reached the point of maturity where there are many users and constant small upgrades to the capabilities of the software. It has also reached the point where it is usable not only to experts in the field but to users from other fields wishing to benefit from the advances. This is a crucial juncture in the development of a field like this and efforts in preserving a legacy now should prevent much wailing and gnashing of teeth in the future over incompatible and out of date archives having to be reinvented instead of merely updated.

4 Formalisation

Formalised mathematical developments might be thought to be the easiest to control and make use of, given their highly structured nature, and the embedded semantics necessary for their development. This semantics, however, is purely an operational semantics. To be useful in supporting human mathematical endeavour, some form of denotational semantics or Kripke semantics is still needed to connect the type `Real` in PVS [SOR] with the type `Real` in HOL [GM93], and to distinguish it from the type `Real` in Mathematica [Wol] which is actually a floating point representation.

A note of caution about management of knowledge in formalised mathematics. As the body of work in the various systems (PVS, HOL, Mathematica etc.) progress, variations on a theme become more likely. As the body of work in these areas grows, the chances of re-inventing the wheel become ever more apparent. Large formalisations run to thousands of lemmas and definitions, and despite efforts at consistent naming, the sheer number of variants required for a formalisation leads to a lack of clarity. Textual searching through the formal development, while more easily achieved than textual search through less structured representations of mathematics, can still lead to an overload of possible results. So, we must not ignore the needs of the user communities of these systems for good search mechanisms based on higher order matching and some form of semantic matching.

4.1 Representation Languages and Formats

There are a number of possible frameworks for use in storing fully formal representations of mathematics. Some of these are briefly discussed in Sect. 6 on Prior Art which informs and provides a foundation for MKM. Here we will compare some of the existing systems in their scope and utility. A full exploration of these varied systems would require at least one, probably many, papers so by necessity this is only a brief overview.

OpenMath [Dew00b, Dew00a] and OMDoc [Koh01] form an interesting starting point for consideration of formats for representing mathematics. As pointed out in [APSCS01, APSC$^+$01], however, the source of OpenMath as primarily a transport layer for CAS, later expanded to cover theorem provers, leaves some aspects of it unsuitable as a generic formal representation system for mathematical information.

There are a number of individual systems with a good underlying format for the representation of parts of mathematics, such as Mizar [Try80]. However, these systems tend to be tied very strongly to a particular system and a particular viewpoint of mathematics. So, while these systems are worthwhile and form a highly useful and usable source of mathematical knowledge, it is unlikely that they will be useful as a format for generic representations.

The same can be said of the various theorem provers and computer algebra systems. Their languages, and even more cross-system developments such as the CASL [Mos97] are still more focussed on a single kind of mathematical knowledge.

In bringing MKM into sharper focus there must be a review of the various languages in terms of their capabilities, their current usage and their suitability for the needs of MKM. This must be done in parallel with identifying requirements and useful goals in the type of information needed to be accessed, and how that information can be described in a formal way as well as stored in a formal way.

5 Information Dissemination

So, we have seen that the various types of mathematics held on and accessible by computer have their problems. The job of MKM, like that of a librarian, is not just to catalogue, file and cross-reference information. It is to provide it to those who desire it, in as usable a form as possible. Hence in this section we consider the problem of MKM in terms of the people involved, their needs and the possible ways MKM might meet their needs.

5.1 Different Users and Their Needs

When considering types of user for MKM, we are not considering individuals and splitting them into discrete sets. Rather we are (to use a term from the Human-Computer Interaction (HCI) community) considering the "modes" of working displayed by users at different times. These modes can be categorised as:

Producer creating new mathematical information.
Cataloguer filing and sorting a variety of mathematical information into new collections.
Consumer using mathematical information in an application, or for educational purposes.

Note that in the process of producing mathematical information, most will also be consuming other pieces: it is a rare mathematician who does not depend on the body of mathematics built up over two thousand years, at least by direct reference to prior art. A cataloguer is a special form of producer, but one with such an obviously key role in MKM that we feel that mode should be singled out for more detailed attention (see Sect. 5.4 in particular). Finally, even in the "simple" act of applying prior art, a consumer may still produce new material worth being disseminated to others.

The other useful distinction between classes of user is in their type of employment. Mathematicians have always included highly skilled "amateurs" in their ranks as well as those professionally involved in the production of mathematics. These, and those primarily involved in pure academic positions have specific goals, generally involving as wide as possible a dissemination of their work.

Non-mathematician academics in a wide variety of fields use mathematics and produce their own brand of mathematical developments. Computer scientists, engineer, economists, archaeologists and many others contribute to and benefit from a wide range of sources of mathematical information, and need good MKM for the development of their field.

Industrial mathematicians may also be interested in wide distribution of their own work (for instance see the notes on Intel in 5.3 below), or may be interested in allowing paid access to parts of it. They also desire access to the work of others to improve their own developments.

These different types of user also have different modalities of operation. All, at one time or another, can be regarded in the mode of "learner" or "student". Of course, this includes the user for whom this is the primary (sometimes sole) mode, the student of mathematics. Following on from the student mode of learning a piece of mathematics, there is the "application" mode, where the user is getting to grips with a piece of mathematics and using it to solve a problem. Lastly there is the "development mode" where the user understands the mathematics well and may come up with novel twists or uses for it, or create entirely new concepts based on the original.

5.2 From Caxton to CERN

From the development of movable type through to recent developments in MathML and online journals, the dissemination of mathematics has gone through many changes. Originally one could access such material only by talking to the mathematician who originated a new concept, or via a limited number of libraries containing scrolls and then books laboriously copied by hand, an error-prone procedure, especially when carried out by those with no mathematical training themselves. The days of movable type improved this process, but the expense of professional typesetting of mathematics, together with personal typewriters and early word-processors (with type limited to a few symbols beyond alphanumerics) led to a large number of monographs, theses and the like being produced with typed words but hand-drawn symbology. Gradually this gave way to more powerful digital solutions and word processors with equation editors, TEX and LATEX, and CAS with various output options.

The business of publishing has gone through many similar changes. In the beginning natural sciences were supported by the church and universities (many of them supported by the church as well, though some more state-oriented institutions existed quite early on). Movable type in particular led to the development of a business model whereby a publisher put up the costs of the printing (either directly by ownership or by paying a separate printing business) and took a share of the profits gained from distributing the product. Books and journals on mathematics have proliferated over the years, and the advent of cheap self-publication on the Web has led to changes in both the business models and the academic models. However, these models are slow to change and technology has overtaken the speed with which slow evolution can happen. Pressures are building in both the academic and business worlds which MKM must take account of in order to help rather than hinder the goal of MKM, which is to better disseminate mathematical information to those who need it, while allowing due benefit to those who produced it.

5.3 Copyright, Corporate Publishing and Quality Concerns

In recent years the academic publishers, once concerned with the quality of their output at least as much with its profitability, have mutated out of recognition. The “independent” publishers have mostly been bought up by large multi-media conglomerates and despite the efforts of those who care about the content, the model of the academic journal, edited and refereed by unpaid academics for the prestige they gained, is becoming less and less viable with respect to these large conglomerates, who have priced themselves out of the market in many cases, by increasing the costs to libraries well beyond the increase in University funding in most places. Increasingly, academics are beginning to wonder what these costs are doing other than expanding the bottom line of the publishing houses. Even the University publishing houses are under pressure from their parent institutions to produce more income, as other sources decrease. (This is not true everywhere, but even where the institutions are not under financial pressure, the profits made by the non-University houses leads to similar increases at the University houses in many cases.)

The unpaid labour provided by academics and scholarly societies is increasingly begrudged. The ever more restrictive copyright terms demanded by publishers is another cause for concern, leading the International Mathematics Union to recommend a counter-proposal for copyright agreements [Bor01]. One leading mathematics academic has even announced that he had abandoned reading journals in favour of the Alamos archive of “live” papers. As academics face pressure to produce more, the time and energy they have for the unpaid work of quality control on the publications of their peers continues to diminish. MKM surely has a role to play in improving the efficiency of this work, allowing faster, more insightful refereeing within new models of publication. This needs to be true not just for formal mathematics, which is supposed to automate some of the refereeing process by using a proof checker to test the validity of claims, allowing the referee to study the definitions and final resulting theorems rather than needing to work through the intermediate result. The role of a referee can be seen as part of the cataloguing work that needs to be done for the efficient management of mathematical information. The role of referees in determining the importance and overall quality of mathematics that is a candidate for publication would be very difficult to automate, but the book-keeping involved in this process can be made much easier with the correct attribution of meta-data, and the easy availability of referenced prior art. The recent scandal of falsified results in physics at Bell Labs [Bel02], with papers published in Science and Nature, reinforces the necessity of providing good support for the peer review system in mathematics and related subjects.

The attitude of the free software movement, that better code comes from free distribution and the right to produce derivative works, has many similarities with the attitudes of mathematicians (particularly academics). Even companies who business model is built primarily on the value of certain classes of intellectual property (such as Intel Corporation) recognise that other intellectual developments (such as the formalised mathematics of Harrison [Har00]) are more valuable in free dissemination than in closed-source exploitation. Formalised mathematics still mostly follows this ideal and MKM tool developers need to be aware of the desires of the mathematics communities in similar respects.

The idea of the mathematical toolbox, like the IT toolbox being developed as Open Source Software, can be shown to benefit a variety of users. Mathematics, like IT, is more often a tool in wealth generation, than the product generating the wealth directly. it is in the interests of society and mathematicians that the tools in these toolkits are as available as possible, and flexible in their applicability.

So far in this section we have focussed on the issue of new mathematics being published. As we have said before, however, MKM is at least as much about enabling access to the existing corpus as it is about new material. Here we have a difficult problem, posed by existing agreements between publishers and the original authors. While there are many possible copyright agreements between author and publisher, the ones which concern us are those which either assign copyright (or all rights of exploitation) to the publisher, or those which restrict electronic rights to the author themselves. For material which needs digitisation (from journals, books or conference proceedings) some of this is probably now out of copyright (though less and less new material is being added to the intellectual commons in this way due to ever-increasing lengths of copyright protection). However, the vast majority of material published after 1930 will be under copyright still, and care would have to be taken with any large-scale digitisation effort to comply with the appropriate law (especially with international treaties since such an effort would almost certainly be undertaken in various global locations). Any estimates of the effort involved in such work should include costings on contacting, negotiating with, and possibly paying, copyright holders for the right to transfer their material to a new form (i.e. producing a derivative work). It should be noted that even where an author has not signed over copyright in the work itself to the publisher, the publisher still retains copyright over the layout of the material and must be included in discussions.

5.4 Meta-Information

The Web in general is beginning to suffer from information overload. Legal arguments about deep-linking, the copyright status of "collections of information" and digital copyright laws are contributing to the chaos in general information management at present. While it can be expected that these arguments will go on for some time, it can be hoped that relatively stable situations can be reached with respect to parts of the mathematics community quite quickly.

In respect of this, it is in the development of good infrastructure for supporting meta-level information (information *about* a piece of digital mathematics) that MKM may have the quickest and most prevalent effect. Meta-Information can be produced automatically or "manually", and in the case of information only digitised but not represented, manual production is quite probably the only feasible way forward. Some of this meta-information already exists, in fact, such as the straight citations of papers. As we shall see below, however, even these are not without problems are present. As with much of MKM development, the limited resources available must be put to good use, and frequently, the manual or machine-assisted production of good quality meta-information will be of far more use than the translation of digitised material into symbolic representations. Only once automatic translation and automatic extraction of meta-information become possible will this change. All these areas should be explored and developed but the near-term realisable goals should be pursued with greater vigour in the short term while laying the groundwork for improved automation in the longer term.

While MathSciNet [AMS], the Alamos archive, citeseer [Cit] and other systems are useful, the lack of a good underlying representation of the mathematical content of pages reduces that utility exponentially in the number of papers stored. Note, for instance, that most journals in mathematics and the mathematical sciences, such as computing and physics, generally advise authors to avoid use of mathematical formulae in the titles of papers, so that text-only searching can be applied at that level. Good symbolic representations and formalisations are obviously the way in which meta-information should be stored an manipulated, and it is in this area that the experience of those working in theorem proving (automated and machine-assisted) and CAS should prove highly useful. HCI experts are also important to such undertakings. An excellent database of meta-information about digitally stored mathematics remains useless if the interface is counter-intuitive. Not only must the underlying representation of meta-information be suitable for automatic processing and searching (a hard problem in itself) it must also be possible to enter and display the queries and results in a usable fashion. The experience of HCI experts in theorem proving and CAS should again be invaluable in this arena.

The various technologies mentioned above (OpenMath, OMDoc, HELM, "book-slicing") are a vital starting point for identifying suitable technologies for storing and accessing mathematically useful meta-information. Once good methods for storing and entering meta-information for mathematics have been developed, there needs to be a concerted effort to "sell" it to as many producers and cataloguers of mathematics as possible. This does not mean ask them to part with money for the right to be involved in the web of knowledge, but to persuade them that being involved is in their interests and that making their meta-information freely available to others in return for reciprocal access, will benefit them in the long run. The intellectual property issues, in both a legal and a social sense, make this area a sensitive one that must be handled with care.

6 Prior Art: Previous and Current Projects of Relevance to MKM

It would be impossible to cover all the projects that have led to the formation of the current MKM community. However, there are various projects, either recently completed or currently ongoing, to which members of this community should pay close attention, particularly to the projects they themselves were not involved with, the assumption being that they are well aware of the potential contributions of their own previous work. So, in this section we will give a quick description of a number of project, primarily EU funded, which form the background to current developments of a concerted effort in Europe, not ignoring the contributions of those elsewhere, particularly in the US. Some of these projects have already been discussed in more detail but it is useful to have them mentioned here again.

- **Monet: http://monet.nag.co.uk/**
  Mathematics on the Net. A recently funded EU project starting in early 2002. This project is aimed at the general theme of "mathematical web services", generic mathematical services offered over the internet.
- **OpenMath: http://monet.nag.co.uk/openmath**
  A Common Mathematical representation language for mathematical software, particularly CAS and theorem proving systems. The OpenMath Society continues to support and extend the standard. A new EU project started in 2001, following on from the original project which ran from 1997 to 2000.

- **Calculemus: http://www.calculemus.net/**
  A project to combine the capabilities of CAS (excellent for calculation) and Theorem Provers (good for proving logical statements) into better mathematics software. In addition to an EU funded network of sites with junior researchers, there are annual workshops/conferences and a recent autumn school.
- **MoWGLI: http://www.mowgli.cs.unibo.it/**
  Mathematics on the Web, Get it by Logic and Interfaces
  The MoWGLI abstract from their website:
  The World Wide Web is already the largest resource of mathematical knowledge, and its importance will be exponentiated by emerging display technologies like MathML. However, almost all mathematical documents available on the Web are marked up only for presentation, severely crippling the potentialities for automation, interoperability, sophisticated searching mechanisms, intelligent applications, transformation and processing. The goal of the project is to overcome these limitations, passing from a machine-readable to a machine-understandable representation of the information, and developing the technological infrastructure for its exploitation. MoWGLI builds on previous standards for the management and publishing of mathematical documents (MathML, OpenMath, OMDoc), integrating them with different XML technologies (XSLT, RDF, etc).
- **Euler**
  Euler is a relatively recently funded EU project to develop a web portal for mathematics available on the web. This project is aimed at the user interface level of searching for, and using, mathematical information and services via the internet.
- **Mizar**
  One of the oldest Formalised Mathematics projects. Since the early 1970s the Mizar project has been a central point for a large development of fully formalised mathematics and the development of the Mizar system and its components. Various aspects of the system have influenced the development of related projects, such as the ability to produce "human-readable" versions of proofs.
- **MKMNet: http://monet.nag.co.uk/mkm/**
  This EU 5th Framework project (IST-2001-37057) has been funded to support the development of a 6th Framework Integrated Programme for Mathematical Knowledge Management. It is run by a consortium of (primarily) academics most of whom were present at the MKM '01 workshop.
- **NIST Digital Library of Mathematical Functions**
  Based on the *Handbook of Mathematical Functions* [AS72], this NIST project aims to digitise and extend the utility of this seminal mathematical reference text. Two papers were presented about this project at MKM '01: [Loz01, MY01].
- **Digital Library of Mathematics: Cornell University**
  This project has already been described in some detail in Sect. 3.4. it aims to produce a library of formalised mathematics, from various theorem proving and related systems, for use in program verification and synthesis, for the US ONR.

7 Conclusions

We hope we have covered a wide range of the issues surrounding MKM in an interesting and enlightening way. The differing viewpoints of technical issues (the differentiation and the links between digitised, symbolically represented and formalised mathematics) and user issues (copyright, HCI, mate-information) have been covered and some of the links between them explored. This is something of a position paper, and some of the ideas are drawn from the discussions since the first MKM workshop in Linz [BC01].

Attention must be paid not only to the technical issues but the social and legal ones as well. Without the engagement of as many of the stakeholders as possible, the development of MKM will be held back.

The role of mathematics in the world continues to grow. While computers seem ubiquitous, it is really mathematics that is ubiquitous, from e-science initiatives, to cryptographic protocols, they're all based on mathematics as well as computer hardware. Without good MKM, the development of knowledge economies and a net-enabled world is delayed and made poorer.

7.1 Acknowledgements

Many of the ideas in this paper are refined from discussions with members of the MKMNet consortium and attendees at MKM '01. I would also like to thank the anonymous referees for a number of good suggestions including the list of related EU project which have contributed to the development of the MKMNet project.

References

[AA01] M. Athale and R. Athale. Exchange of mathematical information on the web: Present and Future. In Buchberger and Caprotti [BC01].

[AMS] AMS. Mathscinet. www.ams.org/mathscinet/.

[APSC$^+$01] A. Asperti, L. Padovani, C. Sacerdoti Coen, G. Ferruccio, , and I. Schena. Mathematical Knowledge Management in HELM. In Buchberger and Caprotti [BC01].

[APSCS01] A. Asperti, L. Padovani, C. Sacerdoti Coen, and I. Schena. HELM and the Semantic Math-Web. In Boulton and Jackson [BJ01].

[AS72] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs and mathematical tables.* Dover, 1972.

[AZ99] M. Aigner and G. M. Ziegler. *Proofs from THE BOOK.* Springer, 1999.

[Bab23] C. Babbage. On the Theoretical Principles of the Machinery for Calculating Tables. *Edin Phtl Jrl*, 8:122–128, 1823.

[BB$^+$96] B. Barras, S. Boutin, et al. The *Coq* Proof Assistant Reference Manual (Version 6.1). Technical report, INRIA, 1996. Available on-line with *Coq* distribution from ftp.inria.fr.

[BB01] P. Baumgartner and A. Blohm. Automated Deduction Techniques for the Management of Personal Documents (Extended Abstract). In Buchberger and Caprotti [BC01].

[BC01] B. Buchberger and O. Caprotti, editors. *MKM 2001 (First International Workshop on Mathematical Knowledge Management).* www.risc.uni-linz.ac.at/conferences/MKM2001/Proceedings, 2001.

[Bel02] Research Review, 2002. www.lucent.com/news_events/researchreview.html.

[BJ01] R. J. Boulton and P. B. Jackson, editors. *Theorem Proving in Higher Order Logics: 14th International Conference.* Springer-Verlag LNCS 2152, 2001.

[Bor01] J. M. Borwein. The International Math Union's Electronic Initiatives (Extended Abstract). In Buchberger and Caprotti [BC01].

[CA$^+$86] R. L. Constable, S. F. Allen, et al. *Implementing Mathematics with the NuPrl Proof Development System.* Prentice-Hall, 1986.

[Cit] Citeseer. www.citeseer.org.

[CRL01] J. A. Campbell and E. Roanes-Lozano, editors. *Artificial Intelligence and Symbolic Computation, International Conference AISC 2000. Revised Papers.* Springer LNCS 1930, 2001.

[Dav01] J. Davenport. Mathematical Knowledge Representation (Extended Abstract). In Buchberger and Caprotti [BC01].

[Dew00a] M. Dewar. OpenMath: An Overview. *ACM SIGSAM Bulletin*, 34(2):2–5, June 2000.

[Dew00b] M. Dewar. Special Issue on OPENMATH. *ACM SIGSAM Bulletin*, 34(2), June 2000.

[DM97] M. Dauchet and Bidoit M., editors. *Proc. Intl. Symp. on Theory and Practice of Software Development.* Springer LNCS 1214, 1997.

[Fro02] M. Froumentin. Mathematics on the Web with MathML. www.w3.org/People/maxf/papers/iamc.ps, 2002.

[FTBM96] R. Fateman, T. Tokuyasu, B. P. Berman, and N. Mitchell. Optical Character Recognition and Parsing of Typeset Mathematics. *Journal of Visual Communication and Image Representation*, 7(1):2–15, March 1996.

[GM93] M. J. C. Gordon and T. F. Melham, editors. *Introduction to HOL.* CUP, 1993.

[Har71] M. Hart. Project gutenberg, 1971. www.gutenberg.org.

[Har00] J. Harrison. Formal verification of floating point trigonometric functions. In Hunt and Johnson [HJ00].

[HJ00] W. A. Hunt and S. D. Johnson, editors. *Formal Methods in Computer-Aided Design: Third International Conference FMCAD 2000.* Springer-Verlag LNCS 1954, 2000.

[JB01] M. Jones and N. Beagrie. *Preservation Management of Digital Materials (A Handbook).* The British Library, 2001.

[Kea00] T. Kealey. More is less. *Nature*, 405(279), May 2000.

[Knu79] D. Knuth. *TEX and METAFONT: New directions in Typesetting.* AMS and Digital Press, 1979.

[Koh01] M. Kohlhase. OMDoc: Towards an Internet Standard for the Administration, Distribution and Teaching of Mathematical Knowledge. In Campbell and Roanes-Lozano [CRL01], pages 32–52.

[Lam94] L. Lamport. *LATEX: A Document Preparation System, 2/E.* Addison Wesley, second edition, 1994.

[Loz01] D. Lozier. The NIST Digital Library of Mathematical Functions Project. In Buchberger and Caprotti [BC01].

[McC62] J. *et al.* McCarthy. *LISP 1.5 Programmer's Manual.* MIT Press, 1962.

[Mic01] G. O. Michler. How to Build a Prototype for a Distributed Mathematics Archive Library. In Buchberger and Caprotti [BC01].

[ML84] P. Martin-Löf. *Intuitionistic Type Theory.* Bibliopolis, 1984.

[Mos97] P. D. Mosses. CoFI: The Common Framework Initiative for Algebraic Specification and Development. In Dauchet and M. [DM97], pages 115–137.

[MY01] B. R. Miller and A. Youssef. Technical Aspects of the Digital Library of Mathematical Functions *Dreams and Realities*. In Buchberger and Caprotti [BC01].

[Pau88] L. C. Paulson. The Foundation of a Generic Theorem Prover. *J. Automated Reasoning*, 5:363–396, 1988.

[SOR] N. Shankar, S. Owre, and J. M. Rushby. *The PVS Proof Checker: A Reference Manual.* Computer Science Lab, SRI International.

[Try80] A. Trybulec. *The Mizar Logic Information Language*, volume 1 of *Studies in Logic, Grammar and Rhetoric.* Bialystok, 1980.

[Wol] www.wolfram.com.