Three Years of DLMF: Web, Math and Search*

Bruce R. Miller

Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD
bruce.miller@nist.gov

Abstract. DLMF was released to the public in May 2010 and is now completing its 3rd year online. As a somewhat early adopter of large-scale MathML content online, and exposing a math-aware search engine to the public, the project encountered situations distinct from those with our previous web sites. In the hopes that our experiences may inform developers of current and future Digital Library projects, we describe some of our observations delivering MathML content and trends in both web usage and browser evolution. We will also look at the the ways our readers have used math search, attempting to assess whether they found what they sought, and ways the engine might be improved.

1 Introduction

Three years ago, after a considerable gestation, the Digital Library of Mathematical Functions (DLMF) [5] was released as a free resource to the public. As it is the successor to the *Handbook of Mathematical Functions* by Abramowitz & Stegun [1], we also served the traditional audience with the commercial publication of a companion Handbook [9]. We faced certain challenges [7]: to use what were (when we started) cutting edge technologies like MathML [3] to enhance reuse and accessibility; given the heavy mathematical content, math-aware search seemed essential; we needed to develop tools to assist in authoring XML and MathML content.

In our role as proponents of Mathematical Knowledge Management (MKM), we are enthusiastic about MathML and work to develop enabling technologies such as math-aware search; we'll happily promote these technologies in venues such as the current one. On the other hand, the goal of the DLMF project itself is to provide and make useful the mathematical knowledge it contains. It uses the technologies, even quietly encourages their use, but doesn't loudly force or announce them. The effect of this is that our users are not, for the most part, coming to our site with expectations of MathML or math search; when they submit a search query they are more likely to simply use what comes to mind, than following explicit instructions for 'how to search for math'.

Our DLMF is just one of several different kinds of 'Digital Library' and so not all of our experiences will be relevant to all other library developers. Nonetheless, we anticipate that many of them will be helpful to other developers. After looking at general usage of the DLMF, we will focus on the delivery of MathML and use patterns of math-aware search.

J. Carette et al. (Eds.): CICM 2013, LNAI 7961, pp. 288–295, 2013.

2 Response

The reviews and feedback we have had on the DLMF and Handbook, at least to our faces, has been almost completely positive. Most complaints that we have received question the accuracy of formulae, occasionally with justification. In a few other cases, we have apparently overlooked listing someones favorite software package. Another handful of comments concerned technical problems with MathML, typically missing fonts, or other browser issues. In response, we have made 5 minor updates of the DLMF to include 20 corrections in errata along with various technical and conversion improvements.

An internal study of the citation indices indicates that citations of the Handbook and DLMF are gradually displacing citations of Abramowitz & Stegun. Citations specifically of the online DLMF seem to be a small portion (17%) of the total. Although the continued increase of citations of Abramowitz & Stegun had originally been an important motivation for the DLMF project, the total citations over the last 3 years, ironically, seem to be leveling off. A speculation is that use of the DLMF is indeed displacing both the new *and* old printed handbooks, but that since citation of online materials is unfamiliar to most authors, it is either not being cited as consistently, or its citations are harder to recognize in the citation indices. [We've added a 'How to Cite' page, to try to improve the outlook.] It is difficult to confirm this theory, however.

3 General Log Analysis

Web server logs are notoriously difficult to interpret, or are simply unreliable [4]. A surprising amount of traffic, *half the bandwidth*, seems to be web indexing robots (or worse). We seemingly expend as many resources preparing to find material as we do using it once we've found it! Many of those robots routinely masquerade as familiar web browsers. Normally, various caches shield the server from many page requests and thus skew the statistics. However, since we use content negotiation based on agent identification strings, we mark the pages as uncacheable and thus may be less affected by this inaccuracy. Nevertheless, the server logs are what we have and they will have to serve.

The traffic to our site has been gradually increasing since its unveiling (Figure 1). Initially around 200 visits per day increasing to over 600 — we've arbitrarily defined a 'visit' to end when a visitor has not requested a new page for more than 30 minutes. A visit appears to average around 4–5 pages. We're restricting our attention to what we believe are humans in the following discussions.

Sketchy information encourages speculation. Do the trends of average visit duration, shown in Figure 2, indicate a gradually decreasing attention span? Apparently people spend an average of 5 minutes at the site, but near 10% stay for more than a half-hour.

We spent some effort making snippets of our formula available as TEX or Presentation MathML(pMML). There seems to be a pay-off, as it appears that on average, each visitor (385/day) downloads at least a pMML (202/day) or TEX (196/day) or BibTEX (25/day) snippet.

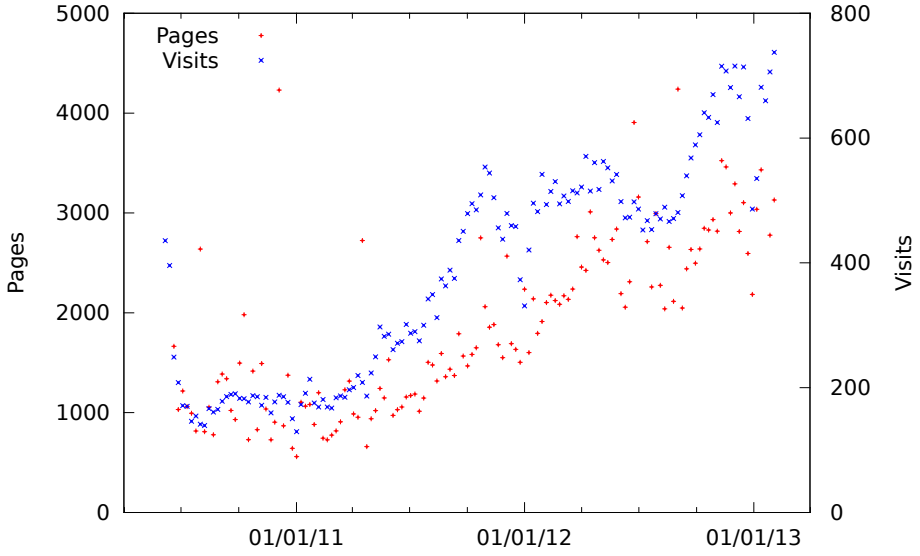

Fig. 1. Daily visits to DLMF

4 MathML Specific Issues

Setting up a website to use MathML may encourage us to be a bit too clever for our own good. We want to serve MathML whenever possible, but otherwise fall back to images for the math. Portal pages, or forcing users to understand and choose the appropriate format, are awkward and intrusive. So we set up our server to determine, by user-agent sniffing, which format the browser supports and send that automatically. It actually works fine, although we have to occasionally update the agent rule base.

One complication is due to breaking the assumptions of stock web analyzers. The analyzer may no longer correctly classify requests as being requests for pages, for example. Moreover, it can yield a wildly misleading picture of browser share which is typically based on the number of successful requests originating from each kind of browser.

The DLMF has a total 1,613 pages, with some 38K math expressions. Thus, to view a typical page, a MathML supporting browser will fetch the single page, with embedded MathML and is ready to display. A non-MathML agent will load the page without MathML, and make an average of 24 extra requests to fetch the images of the math before it can view the page. This severely biases the apparent market share. Interestingly, the total amount of data downloaded in both cases is comparable.

It seems more interesting for our purposes to divide up browsers into 3 categories. DLMF makes fairly heavy demands on the rendering agent, and so only the most complete implementations, whether native or via plugin, are served

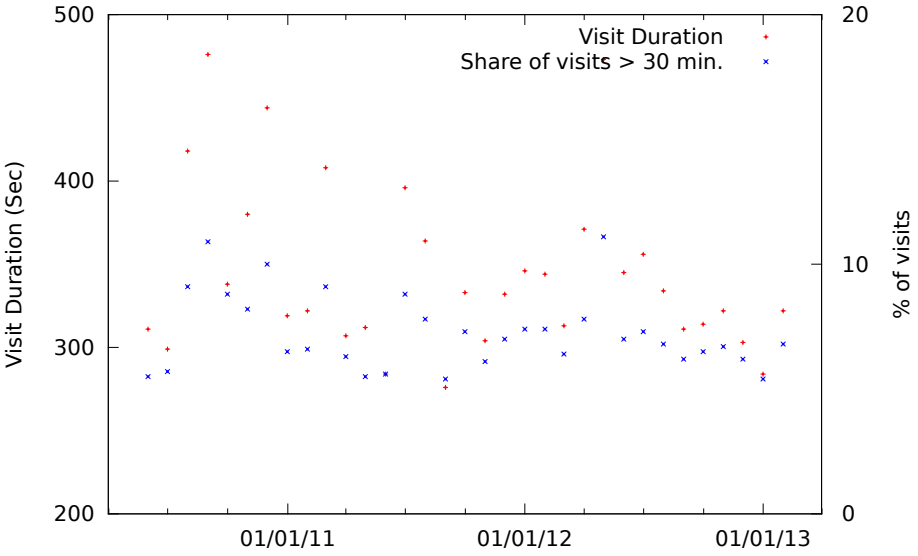

Fig. 2. Lengths of visits to DLMF

MathML by default. A second category has partial (or even sporadic) support for MathML, but not quite good enough to cover DLMF's material; for example, supporting only the CSS profile, or missing crucial elements like prescripts and multiscripts; we'll call these 'almost MathML'. The final category is without MathML support. Figure 3 shows the trend in page views between these three categories of browser. (We'll avoid 'naming names', since support is evolving, and our main interest here is the (positive) outlook.)

One should be careful over-interpreting Figure 3, as these figures seem rather sensitive to the patterns used for robots, and as the robots vary their choice of browser to mimic. Nevertheless, it seems encouraging that MathML supporting browsers, and particularly browsers that could support it, with a bit more effort[1], sum up to such a large and growing share.

In the meantime, there have been two other encouraging developments affecting MathML support. One is the inclusion of MathML in HTML5 [2], along with the support of most browsers. Although many browsers claim to support HTML5, few in fact implement MathML (yet). This lack is partially ameliorated by the second development, the advancement of MathJax [6] which implements MathML rendering using JavaScript and CSS.

One 'take away' lesson is the following. Even if XML and namespaces seems unloved by the HTML5 community, it is only through the use of XML infrastructure that DLMF can almost trivially track this change. We will be adding HTML5 with MathML as a formatting option in the near future; perhaps using

[1] Not that we are offering business advice.

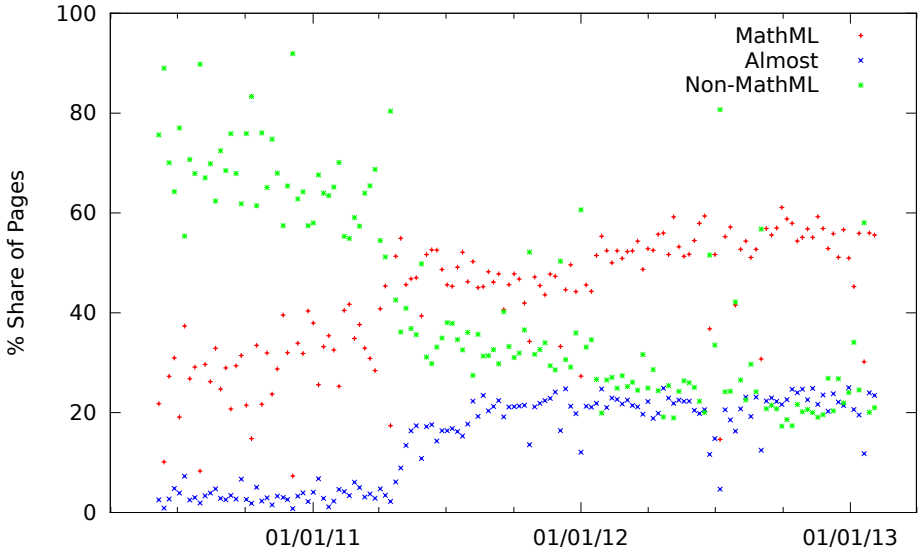

Fig. 3. Browser trends

MathJax as the fallback rendering engine. Additionally, although the agent-sniffing machinery may seem less necessary with the advent of MathJax, it may still be useful for handling other contingencies, such as mobile agents and tablets.

5 Search Issues

DLMF provides a search engine which supports mathematics-specific search, as well as conventional text search. Taking guidance from modern search engines we let the user type whatever query they expect to work, and attempt to make the best of it. We infer, based on the tokens in the query, whether it was intended to find math or text, what kind of notation they're using, and so on. In the following, we'll try to assess whether or not we have succeeded, but it is also interesting to see what queries these untrained users did in fact submit.

Our search engine converts the math to a serialized text equivalent, both in the document during indexing and in the query during search, and then leverages a text search engine to perform the actual search look-up [8]. This type of search engine is oriented towards the more informal usage that we envision. However, it is certainly not the only approach to math-aware search in all contexts.

From our web server analysis, we see that there were an average of 24 searches a day; roughly 1 for every 16 visits. About 16% of them appear to be intended to be math searches, although perhaps a third of these are simple terms, like `exp`, that are easily interpreted as either math or text.

Figure 4 shows the number of 'tokens' (basically sequences of contiguous letters or numbers or individual noise characters) used in math and text queries.

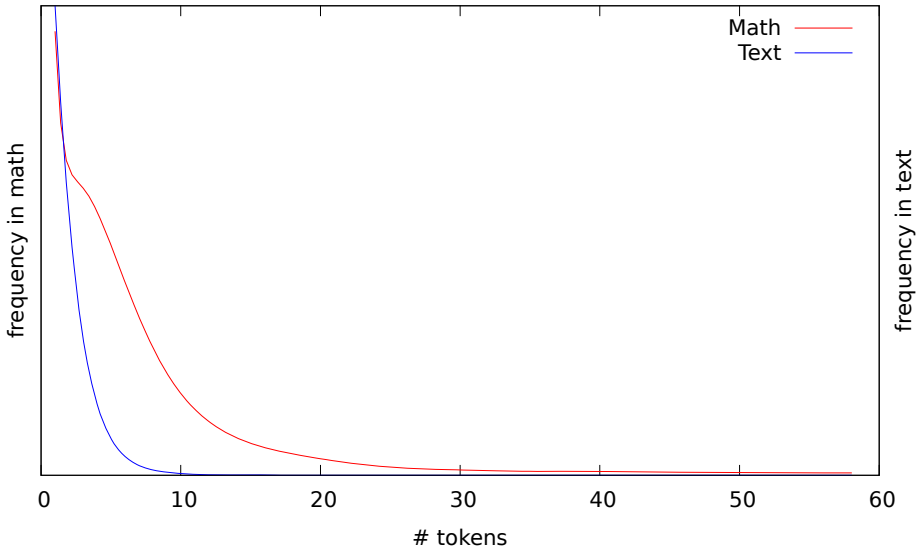

Fig. 4. Distribution of number of search terms per query

The shapes indicates that math queries are most commonly expressions of several terms, but queries up to 58 were seen. Text queries tend to be at most a few words, and a superficial scan of the lists suggest they are almost always phrases.

Out of the math searches, common patterns include

- `$` wildcard (10%: suggesting they *did* read the help file after all!);
- LaTeX markup (5%);
- various identities (e.g. $c^2 = a^2 + b^2$);
- pairs of math symbols presumably expected in the same formula, but not a math expresssion, as such;
- simple formula fragments: `sin 2x`, `sinh cos`.

Some surprising math queries include

- cut & pasted long formula;
- examples from Help page;
- `sinx+cosy` or `sinacosb`;
- `\hbox` and `\vbox`;
- `x_sub{0}`;
- `+` apparently used as query meta-operator (or url encoding?).

But does it work? At a commercial website, a sale is easily recognized as a success, but for a Digital Library, success is your readers discovering the information they desire. Did they leave the site because they found the information, and are now going on to do some productive work, or did they leave out of frustration?

Short of a survey, with its own set of problems, how can we tell if the search engine works for the users?

We therefore turn back to our server logs to attempt to infer success or failure from the users' sequences of actions. What do they do *after* they have executed a search? If they follow the link to one of the search results (we'll call that 'Click Thru' in the following), what do they do after that? One might idealize a a successful outcome as when a user performs a search, inspects one of the results and then, having found the desired item, will visit random other pages within the site. Less successful outcomes would have the user floundering, checking the next page of hit results ('Next 10'), trying other search results or formulating alternative searches ('New Search').

Table 1 collects the tracks derived from our web logs, showing how the behavior of users searching for math differed from those searching for text. It would seem that searchers for text often follow that idealized path suggested above. While it isn't clear that searchers for math are unsuccessful, they certainly appear to need more fishing around to find what they wanted (assuming they did); they were more likely to check the next page of hits or try a different query. Moreover, even after they've clicked on one search result, they were more likely to come back to the search results and try another result or another search. Whether this is somehow due to the different nature of 'searching for a math expression', or is a measure of poor search results is hard to tell.

Table 1. What users did after a search, or after clicking on a search result

		Next page request				
	Total	Click Thru	Next 10	New Search	Other Page	Left DLMF
Searches	23190					
Math	3644 (16%)	37%	12%	30%	14%	7%
Text	19308 (83%)	43%	5%	26%	17%	8%
ClickThru	20888					
Math	2963 (14%)	18%	13%	33%	30%	6%
Text	17751 (85%)	24%	5%	18%	44%	9%

6 Conclusions

The DLMF is online and appears to be appreciated and used after 3 years. Of course, mathematics, let alone, special functions, is a niche, not mainstream, interest; we don't expect web traffic to rival Google or Amazon. Nor do we expect browser implementers to pay as much attention to MathML as to video. Nevertheless, there are reasons for optimism about delivering math on the web; solutions sometimes appear where you least expect them.

We find math search to be used modestly, but this is not surprising given that users don't expect it and we have ruled out being confrontational to promote

it. Our log analysis suggests that math searches require a bit more work to find results than do text searches, but nevertheless appear to serve the users. A more convincing analysis of search behaviors, and indications of search success, would likely require instrumenting the search engine to generate search-specific logs.

Disclaimer: Certain products, commercial or otherwise, are mentioned for informational purposes only, and do not imply recommendation or endorsement by NIST.

References

1. Abramowitz, M., Stegun, I. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, D.C. (1964)
2. Berjon, R., Leithead, T., Navara, E.D., O'Connor, E., Pfeiffer, S.: HTML 5.1, `http://www.w3.org/TR/html51/`
3. Carlisle, D., Ion, P., Miner, R., Poppelier, N.: Mathematical Markup Language (MathML), W3C `http://www.w3.org/TR/MathML/`
4. Kathuria, P.: I, robot? Don't believe your web stats (March 10, 2013), `http://www.limov.com/library/do-not-believe-your-web-stats.lml`
5. NIST Digital Library of Mathematical Functions, `http://dlmf.nist.gov/`, Release 1.0.5 of 2012-10-01. Online companion to [9]
6. MathJax, `http://www.mathjax.org`
7. Miller, B., Youssef, A.: Technical Aspects of the Digital Library of Mathematical Functions. Annals of Mathematics and Artificial Intelligence 38, 121–136 (2003)
8. Miller, B.R., Youssef, A.: Augmenting Presentation MathML for Search. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) AISC/Calculemus/MKM 2008. LNCS (LNAI), vol. 5144, pp. 536–542. Springer, Heidelberg (2008)
9. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W. (eds.): NIST Handbook of Mathematical Functions. Cambridge University Press, New York (2010), Print companion to [5]