

Tensions and Questions in JSTOR Data, Math and Otherwise

Nigel Kerr, nigelk@{jstor.org,umich.edu}
Production Programmer, JSTOR

JSTOR (<http://www.jstor.org/>) is a not-for profit with the broad mission of preserving important scholarly journals in a long-term electronic archive, and provide and improve access to that scholarly content. JSTOR's content includes page images of journals, bibliographic information, and OCR'd text for searching.

Important to searching of mathematical content are these two specific guiding principles of JSTOR:

That access to the content should improve over time

The content should be represented faithfully, in both the page images, bibliographic information, and OCR'd text, as far as is possible.

JSTOR includes a great deal of important Mathematics literature (available in some large, multi-disciplinary collections, and a Mathematics and Statistics collection, see <http://www.jstor.org/about/collection.list.html>).

What I Need and Desire:

What I as a representative of JSTOR would most like to hear are things like:

"You could pay attention to this or that forum about these issues."

"You could provide this kind of functionality in searching."

"You need to have data representation of this kind."

"These are the benefits that users get when you have this functionality or that kind of data."

"Here are effort costs and drawbacks for these kinds of features."

With this kind of input colleagues and I at JSTOR can have a good discussion about what we could do better, which we can accomplish in what time frame, and what we should expect to become of those improvements.

What JSTOR Has Now:

Early in JSTOR's development (mid-1990's), we decided that, while the web browsers of the day wouldn't yet be much help to us in displaying marked-up data like TEX directly, we would have better options down the road. We opted to try to encode TEX snippets in the bibliographic fields (title, author, abstract, keywords, titles and authors of reviewed works, captions of figures) anywhere that the character set we were using (ASCII-7 (long story)) could not represent characters or layout by itself. For instance:

x^4

ρ

$\Lambda = \Lambda(c; \Lambda_{0}, R, \Delta G)$

$\int_S |f'(z)|^2 dx dy \leq \frac{1}{2} \int_{\partial S} |f'(z)|^2 \big(\frac{\partial g(z, t)}{\partial n_z}\big)^{-1} |dz|$

We knew that we at least wanted to preserve the information in question here for display to try to be a faithful representation. At that time, we resolved to merely print the TEX snippets out to the user as-is; we did not have a way to display it more interestingly that didn't involve lots of little images to manage, or a browser plug-in (both of which strategies we've tended to steer away from on principle), but we believed that a better option would arise. We would also merely feed the TEX snippets to our search engine as all other text was. This means today, for the examples above, our search engine can see:

```
4
rho
Lambda Lambda c Lambda 0 R Delta G
int int S f z 2 dx dy leqslant frac 1 2 int
partial S f z 2 big frac partial g z t partial n
z big 1 dz
```

So if you're up on your TEX and feeling frisky, you can sort of search mathematical content in JSTOR. Sort of.

Except for the other problem that arises: we were encoding TEX snippets for several years before browsers had advanced far enough for us to do something meaningful with them. We noted that they were there, and that they rendered through a real TEX system to **look** right, to **look** like what was on the page. These TEX snippets in JSTOR are really first about appearance, not about meaning. If the digitization vendor thought they saw a *Whosits* math symbol, but there was really a *Gee-Gaw* math symbol there, and this distinction was lost on the quality control staff of JSTOR, the only way we can correct this error is to essentially re-code the data with a much more sophisticated understanding of the content.

The OCR'd text at JSTOR heretofore is really geared towards plain text, plain ASCII-7 text. If a mathematical formula appears on a page image, only the letters and numbers and symbols recognized will come across, and certainly the identity of the math involved is not captured.

When you come to JSTOR to search (<http://www.jstor.org/search>), you get to use JSTOR's search page to enter terms or strings to search for in the fields or full text. Within the restrictions of the search engine and the quality of data, one can sort of search for math. If the authors of articles always included helpful words in titles or abstracts or the text of your article to grab on to, that's great, but that certainly isn't guaranteed. Once you get some results, then you can go view pages of the articles one by one or as a PDF all together.

This means that just about anything we do with our math will involve a significant investment of some kind. But, on the bright side, it can only be an improvement. An additional consideration will involve the other disciplines in which we used TEX to capture something (largely Biology and Chemistry, but Economics and other areas have their contributions); we doubt there's a panacea here.

So, what could it be? I'd like to hear your ideas and questions at this conference, and indeed afterwards as they arise.

Nigel Kerr
nigelk@{jstor.org,umich.edu}
Production Programmer, JSTOR

