PREPRINT
Published in ACS Publications' LiveWire e-Newsletter, April 2006.

The Discreet Charm of the PDF

David Flaxbart
Chemistry Librarian, University of Texas at Austin

There is a curious paradox that most science librarians are probably well aware of.

Many researchers have long since abandoned printed journals for their online counterparts. They consume the scientific literature more ravenously than ever. They have eagerly adopted all-electronic modes of information-seeking behavior: searching indexes, browsing tables of contents, setting up alerts, retrieving articles, etc. And yet the end product of all the searching, linking, and downloading is most often this: the PDF article. Yes, the Portable Document Format, a fabulously successful proprietary e-document format that so effectively mimics the old-fashioned printed page. Users have consistently shunned the more "web-native" article formats, despite the latter's greater potential for dynamic multimedia content and hyperlinking.

This fact stands out starkly in journal usage reports. Table 1 shows the percentage of total article downloads that were in PDF format for each title in the ACS package (excluding the Archive).

Table 1. Percentage of PDF Article Downloads, 2005

ACS TITLE	Pct. PDF
Organic Letters	96.9%
Journal of Organic Chemistry	95.4
Organic Process Research & Development	94.7
Organometallics	93.2
Macromolecules	91.9
Journal of the American Chemical Society	91.6
Journal of Chemical Theory and Computation	90.2
Nano Letters	89.7
Journal of Medicinal Chemistry	89.5
Accounts of Chemical Research	89.4
Journal of Physical Chemistry A	89.2
Crystal Growth & Design	88.6
Journal of Chemical Information and Modeling	88.6
Chemical Research in Toxicology	88.5
Chemistry of Materials	88.3
Industrial & Engineering Chemistry Research	88.2
Journal of Physical Chemistry B	88.0
Environmental Science & Technology	87.8

Chemical Reviews	87.5
Langmuir	87.4
Journal of Combinatorial Chemistry	86.8
Journal of Proteome Research	86.5
Inorganic Chemistry	85.5
Journal of Natural Products	85.4
Energy & Fuels	84.3
Bioconjugate Chemistry	83.9
Analytical Chemistry	81.9
Journal of Chemical & Engineering Data	78.7
Molecular Pharmaceutics	78.3
Biotechnology Progress	77.1
Journal of Agricultural and Food Chemistry	76.6
Biochemistry	76.0
Biomacromolecules	75.1
Total for All Journals excluding Archive	**89.9%**

Source: ACS COUNTER Usage Reports, University of Texas at Austin, 2005.

Many questions emerge from the HTML/PDF paradox. Why are users who are so thoroughly wired still so wedded to a printed-page format for articles that has changed little in the last 50 years? Do people really print most of their articles for reading offline, or do they read them on their screens? What don't they like about HTML-type article formats? What advantages do PDFs have? Are there differences across various disciplines in their preference, and if so, why?

To get some insight on these questions, I conducted a dangerously unscientific telephone survey of two senior faculty members at UT-Austin (N=2). The first was a prominent chemical engineering professor whom I knew to be a voracious and enthusiastic consumer of e-articles. I stated the paradox and asked him for his thoughts on it. As if he'd been preparing for days, he immediately launched into a long and well thought out list of reasons why he only uses the PDFs, never the HTML. These are some of the points he made:

- PDFs are far more readable than HTML on modern flat computer screens. There's less scrolling, images are actual size rather than tiny thumbnails, and the resolution at 125% zoom makes for easy reading.

- He never prints his articles. He reads them onscreen and stores them on his hard drive. (His students tend to print a lot more, he says.)

- He "archives like crazy" and stores as many as 1,000 or more articles a year on his computer drives. This is far more than he ever could have printed and filed in paper days. He says he's essentially building a customized, focused digital library of his own.

- PDFs are easily searchable using the internal search features of Adobe Acrobat, when he needs to find a particular fact or statement. His operating system's search options can similarly search text across his stored PDFs.

- He uses Adobe Acrobat to add comments and notes to the papers.

- He shares PDFs with his students at his desk. He can e-mail important new articles to his research group with just a few keystrokes.

- He browses far more than he reads; he actually reads maybe one in ten articles that he downloads. Yet he thinks he reads about five times as much now as he used to in the print world. As a result he feels much more up to date on relevant research than before, and less reliant on serendipity and others to make him aware of things.

- He essentially dismissed HTML articles as unpleasant to read and navigate, and not printable. However, he had not thought much about the hyperlinking features publishers (including ACS) have incorporated into HTML articles. He admitted that allowing users to link directly to articles in a paper's bibliography was a useful idea.

In short, he provided most of the likely reasons why PDF is a nine-to-one favorite over HTML for ACS journals. Yet his end result is not a print-out, but a stored file.

The second survey subject was a biochemistry professor, and he offered a contrasting picture. He stated that recently he has moved more to scanning and reading an article in HTML first, and choosing PDF only if he wants to print it out for filing. He said that, while the graphics and figures in HTML versions are very good in general, they are also very good in PDF versions, at least on the computer screen. But if you don't print out an article on a good color printer on high-quality paper, the printed PDFs aren't as useful.

It's clear from the table that the PDF choice varies a great deal by title, and thus by subject. *Organic Letters* usage shows a striking 35:1 PDF ratio, whereas *Biomacromolecules* is only 3:1. Looking at the whole list, one can discern that the PDF preference is not nearly as strong in biology-related fields as in organic chemistry or engineering. It's understandable that a large color model of a protein is much easier to see and manipulate in an HTML environment; but organic chemists seem to still prefer the standard page layout.

Looking at usage data from other publishers makes for interesting comparisons. The top-used chemistry journals from Wiley Interscience showed an equal or greater preponderance of PDF usage.

Table 2: Percent PDF from Top Wiley Chemistry Journals

	PCT. PDF
Eur. J. Org. Chem.	95.9
Chem. Eur. J.	95.4
J. Polym. Sci. B	93.4
Eur. J. Inorg. Chem.	93.2
J. Polym. Sci. A	93.0
Angewandte Chemie	93.0
J. Appl. Polym. Sci.	91.9
Rapid Comm. Mass Spect.	84.7

Source: Wiley Interscience COUNTER Reports, University of Texas at Austin, 2005
Note: Wiley titles that do not provide an HTML full text option are excluded from this list.

Yet curiously, usage of the two most prestigious general science journals, *Science* and *Nature*, did not reflect this preference at all:

Table 3: Percent PDF from Science and Nature

	PCT. PDF
Science (all articles)	48.7%
-- Science chemistry articles	60.0
-- Science biochemistry articles	48.2
Nature	46.1

Source: Science Online and Nature Publishing Group usage reports, University of Texas at Austin, 2005

Since ACS, Wiley, Science, and NPG are all COUNTER-compliant and should count downloads in the same way, one can only speculate that the differences may be based somehow on the different types of content offered in these titles, and the different ways the content is presented to users and to external linking systems. (1)

These data raise many questions. Why are there such big differences across multiple publishers? How do chemistry and biochemistry compare to other scientific disciplines? Do HTML layout and style sheets make some journals more "readable" than others? Does the amount of editorial and news content (front matter) in a journal influence the ratio? Does the way a publisher presents options for HTML full text, PDF full text, and abstracts in TOC displays influence a reader's choice of format? How large a role do link resolvers, such as CrossRef and ChemPort, play in determining what format is obtained? Does choice of discovery tools, such as SciFinder and PubMed, in turn have an influence on the "best copy" linking mechanisms used?

Some of these questions are addressed in a forthcoming article by Philip Davis and Jason Price (2). They apply some rigorous statistical analysis to a large set of usage data from several publishers, including ACS, across a number of institutions. This study confirms the PDF preference at ratios similar to those shown above, and postulates that publisher interface design has much to do with the level of PDF preference. However, the paper does not address the subject differences within a publisher's set of journals, nor does it delve into the complex behavioral patterns of users across different scientific disciplines. But the authors point out that it is very difficult to control for all but one variable when analyzing such disparate data across publishers, titles, and user populations. Of course, if you are only looking at a single publisher's journals, as we are here with ACS, the interface variable is removed, indicating that a journal's subject matter must also affect the level of PDF preference.

A 2003 article by Davis and Leah Solla did look at some of these aspects specifically for ACS journals, and concluded that users were essentially "using the system primarily as a networked photocopier, for the purposes of creating print-on-demand copies of articles rather than for browsing and knowledge discovery" (3). This makes perfect sense in light of aggressive efforts on the part of libraries, index providers, and publishers in recent years to facilitate "one-click" movement from separate discovery tools to full text article content using link resolvers. As these mechanisms become more sophisticated, it's possible that users may rarely see or interact with a publisher's web interface, as they move from an index record or article reference directly to a new PDF file with no intermediate stops.

So what does HTML have going for it? Web search engines can index PDF documents as easily as web pages, and hyperlinks can be embedded in PDFs as well. Larger screen size and improved resolution may be overcoming users' longstanding reluctance to read entire articles on their computers. Since formatting articles in markup languages is not a trivial task, some publishers might begin to question whether there's much point in continuing to develop features and programming for web-native content, or even whether to offer articles in HTML at all in journals where its usage is very low. (4)

No doubt there are tech mavens who could enumerate many current and potential advantages of using markup languages over static page images and proprietary formats. But so far these bells and whistles have not made much of an impact on users. After a decade of the e-journal "revolution," the web-native formats still have a lot of catching up to do. In the eyes of the beholder, the typeset article format in use for centuries still dominates the field.

Notes

1. The COUNTER Code of Practice (2005) requires publishers to ignore double-clicks on one and the same article links that occur within a set time limit (10 seconds for HTML, 30 seconds for PDF). However, a user who first views HTML and then immediately clicks on the PDF link for the same article registers two downloads of that article, regardless of the time differential.
2. Davis, Philip and Price, Jason. 2006. "eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER." *Journal of the American Society of Information Science and Technology*, 57(9), in press.
3. Davis, Philip and Solla, Leah. 2003. "An IP-level analysis of usage statistics for electronic journals in chemistry: making inferences about user behavior." *Journal of the American Society of Information Science and Technology*, 54(11) 1062-68.
4. Some Wiley Interscience titles, such as *Advanced Materials*, have never offered an HTML option.

DISCLAIMER:
Download percentages shown above were calculated by the author based on publisher-provided usage reports. The content of this article reflects the views of the author and not of the American Chemical Society or the University of Texas.