Beyond the PDF

The online versions of published research articles can challenge the prevalence of the offline PDF but will require added effort by authors and journals.

With a few exceptions, the print circulation numbers of journals have declined massively, and the reading of articles in print is being supplanted by digital alternatives. So far, scientists have shown a strong preference for the portable document format (PDF) version of individual articles. PDFs can be easily annotated with personal comments, stored and read offline, and emailed to colleagues. But ultimately PDFs are little more than portable facsimiles of traditional print articles. In spite of the greater functionality that an online hypertext markup language (HTML) version of an article can offer, this format often seems to serve as little more than a location holder for the PDF.

The present reality is quite different from what was predicted in the early 1990s: that extensive tagging and indexing of HTML-based articles would usher in a utopia of targeted reading. Scientists would be able to quickly search, filter, link and compare related content from a variety of sources. Needless to say, reality hasn't matched expectations, but publishers are now devoting considerable effort to developing their HTML-based articles. The new open-source journal *eLife* just debuted with a design and functionality built to improve the online reading experience. Nature Publishing Group (NPG) has been steadily enhancing its online platform, and other publishers have made their own efforts.

This month *Nature Methods* and the other *Nature* journals started using a newly organized and expanded list of subject terms, or ontology, to improve the search and delivery of research articles from journal and publisher pages. The chosen subject categories are not, however, intended to include all the terms that could be assigned to the work. For additional fine-grained annotation of papers, technical editors at NPG have separately been annotating all chemical, genetic and protein entities in a growing number of NPG research journals, beginning with *Nature Chemical Biology*. These annotations can be highlighted and linked to details about the entity or used to deliver related content (such as protein-specific antibodies).

Although the broad subject tagging benefits both the HTML and PDF versions of articles, the combination of these two forms of tagging can provide the structure necessary to deliver research results in new forms through HTML.

Development of the HTML-based display of articles can also be used to improve data presentation. For several years now, the *Journal of Cell Biology* has been providing access to raw image data through the DataViewer. Last year *Nature Methods* began integrating supplementary videos into the main text of the manuscript by using embedded links that open a pop-up window for viewing the video without disrupting the reading process.

The publication of the group of Encyclopedia of DNA Elements (ENCODE) papers in *Nature* in 2012 was accompanied by code development for several types of interactive figures. These displays help expose the underlying data in two-dimensional visualizations and can improve data exploration. Elsevier has also developed enhancements in data display, including interactive figures.

The development of online articles faces two challenges. The first challenge is resources: these features require additional development by publishers but also additional effort on the part of authors and journals. Even the seemingly simple assignment of subject terms to manuscripts requires oversight because of the considerable subjectivity in their assignment. NPG manuscript editors are responsible for checking the subject terms assigned by authors and changing them as necessary to indicate the major topics of the manuscript in a manner consistent with related manuscripts. Otherwise, the value of the tagging can be severely compromised. More acutely, displays of both raw images and interactive figures require authors to expend considerable effort to supply additional data, and they necessitate journal editors or technical personnel to oversee this process.

The second challenge is Internet connectivity. Rich HTML-based articles require a browser with an active Internet connection and are therefore less portable than PDFs. This limits convenience. Apps on tablet devices can help bridge some of this portability divide by, for example, supporting interactive figures. Yet targeted reading that makes full use of interactive functions requires an online connection.

A primary value of PDFs has been the ease with which they can be stored and transported between devices and read without Internet access. Browsing through a personal library of research articles along with any notes that accompany them can happen completely offline. The PDF will remain a popular format for research article archiving unless a suitable alternative can be created. But for reading, the HTML versions of articles will increasingly become a more powerful way of accessing the information in published manuscripts and for identifying related research, given that authors and journals provide the necessary support.