Is Content Still King?

Peter B. Boyce

Is content still king? An interesting question. Electronic delivery of information is changing rapidly now. And in so doing, is it changing how we work and what we, as scholars, think is important? Now, with five years of experience in exploring a few of the options provided by using the electronic medium, we can look back and assess the impact of electronic publishing upon the scholarly community.

In summary, we are seeing the print and electronic versions of articles diverge – with the electronic version becoming accepted as the authoritative, and more complete version. We can now recognize the tremendous added value provided by electronic interlinking among articles. And, it is now clear that continued future access to the electronic material will be vital to future generations of scholars, and that the only way to maintain access is to plan from the beginning for the ongoing management and updating needed to preserve access to the electronic materials in the future.

Good Journals, Bad Journals

To get a feel for the current state of electronic publishing, we must come to understand the differences among electronic journals. They are not all equally effective. I like to characterize them as the Good, the Bad, and the Ugly.

But, before casting judgement upon the various sorts of electronic journals, it is important to remind ourselves that the prime function of the journal is to transfer information from the author to the reader. Therefore, a good electronic journal must be a **linked**, **permanent** information resource for **transferring** reliable and accurate information from the producer of the information to the user of that information. The three important terms for a GOOD journal are “linked”, “permanent,” and “transferring.”

Consider the BAD journals, which still make up an all too high proportion of the electronic offerings. A BAD journal is not well linked, either within itself or to other information resources. Such journals don’t transfer information very effectively. They don’t make use of the advantages of the electronic environment.

Then there are the UGLY journals, those which are only published in page image format. They almost always have the added problem of having no links. And to make things worse, such page image journals can not claim to be really permanent. Three strikes against them, and, to borrow from American baseball, three strikes and you’re out.

Why do I consider the page image journals UGLY? They are often produced in PDF format, and I find them very difficult to read on the screen. From the usage statistics of the *Astrophysical Journal Letters*, our readers agree, using the HTML version (presumably for browsing) five times as often as the PDF (which is used for printing out the articles the reader wants to have a paper copy of). Either the font is too small, or you can't see enough of the page on the screen. So, the reader is left with concentrating so much on the mechanism of reading that the content of the article almost becomes secondary. Such journals are basically nothing more than rapid delivery of the same old paper pages, unchanged in design for a century or more. Trying to shoehorn the "portrait" format of the usual printed page (higher than it is wide) into the "landscape" orientation of the screen (wider than high) makes for difficult reading and cumbersome navigation of static, unlinked pages. While it certainly is a great advantage to have pages available right on your own desktop computer, the undynamic, unlinked page images of a PDF format will not satisfy the scholar who is used to the live links, easy navigation, and search capabilities of a well designed, HTML/SGML presentation.

Another problem of the page image journals is the use of a proprietary, albeit popular, format. While the specifications of both PDF and PostScript are openly available, there is no guarantee that they will continue to be readable several decades from now. Do you think formats do not go out of style? Just look at what happened to Wordstar, the primary word processing software of the early 1980s. I know from the personal experience of trying to save one of my friends from his Wordstar cul-de-sac that you can not retrieve and use a Wordstar document today. Will this be the case with PDF documents a half century from now? I say probably so.

All this is not to say that PDF format is worthless. PDF is a fine format when offered as an option, with the goal of allowing users to print out copies of their journal articles. It is great for that. But, as the only delivery option, the page image formats are very poor.

What Have We Learned?

With five years of experience behind them, the astronomical community provides an example of how a set of well linked electronic information resources can work to the benefit of the users. We will look at each of the three characteristics of a good electronic journal, effectiveness of transferring information, linking, and permanence to see how the astronomy journals differ from the paper-based approaches.

The information transfer effectiveness depends upon two things, the efficiency with which a reader can browse and read the journal, and the ease with which the reader can locate the appropriate article to read. The journals of the American Astronomical Society have been designed for doing things which the reader needs in her daily work. The journal is designed to load fast, includes an article table of contents for ease of navigation within the article, and has abundant internal links for rapid skipping from section to section. See Boyce (1996a, b) for a further discussion of the features of the AAS journals.

As an example of how readers use the internal links to improve their browsing capabilities, astronomers, when looking at an article in their field, often jump first to the page of references. This page, in which the authors appear in alphabetical order, has links back into the text of the article where the reference was mentioned. Astronomers first look for their own name in the list of referenced authors, then they jump directly into the middle of the article to see what the article has to say about their work.

This is an example of the valuable electronic capabilities which can be added, which match the way readers use the journal, and which thus aid the researcher. Yet, only a small percentage of electronic journals actually include this simple capability. Why? When I ask, the answers illustrate the stranglehold which the paper format has upon our thinking; "Why would you want to do that?" and " I never thought about that." are typical. But, the worst is, "We couldn't do that. Inline references are not our style." If the journal publishers, arguing from tradition, don't include all the helpful features the electronic environment now makes possible, they will lose their readers to the innovative publishers and the ad-hoc groups of users who understand the changing conditions.

The Wider Set of Information Resources

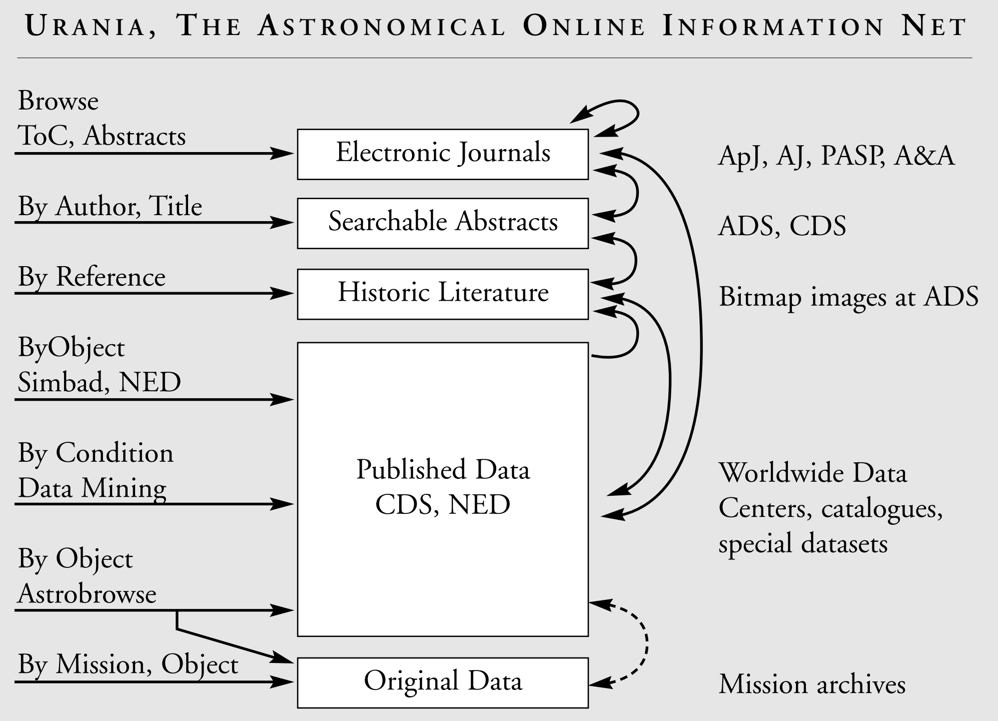
Returning to the subject of effectiveness and finding information, the electronic journals in astronomy have become part of the overall set of electronic resources. Finding relevant articles depends upon the linking capability inherent to the Web. Various astronomical information providers have cooperated in building the infrastructure which supports interlinking and seamless interoperation among the various astronomical resources. Owing to the lack of available standards, we have developed our own article and object identifier system – based upon a journal, volume, page, year combination which works well for the journal literature. We also run a name resolver at each site which allows links to operate through the identifier. URLs, which may change, are not used.

The user can enter the astronomical data system at a number of points, and end up finding the relevant information, be it articles, abstracts, or data. Central to the system is the searchable database of abstracts maintained at the Astrophysics Data System (ADS). This is a valuable resource for the community, developed and maintained with funding from NASA which contains the abstracts of about 500,000 articles. Coverage of the core literature in astronomy is complete from 1975 on (Kurz, et.al, 2000). Searches can be made by author, journal, year, and title, as well as by keyword and full text of the abstract. The search results return the abstract and also link to the full text of the article, either in the electronic version of the journal at the publishers' site, or, if that does not exist, to the scanned page images of the legacy literature. The ADS has completed scanning the full text of all the major astronomical journals starting with volume 1, number 1. For instance the Astronomical Journal is complete back to 1849.

The full-text journal articles, in turn, have their references linked to the abstracts at the ADS, and, in the case of the journals published by the University of Chicago Press, there are links directly to the electronic full text articles, eliminating the need for the extra link through the ADS.

However, the ADS provides a number of other links which are of great use to the reader. They link to future citations which refer to the article, to tabular data in the international astronomical databases, to image libraries, and to other useful resources.

From the user's standpoint, the links work seamlessly, and it is often not clear when the user leaves the journal Web pages and enters the abstract database, and vice versa. Figure 1 illustrates the various astronomical information resources and their interconnections. The user can choose where to enter the information network depending upon their needs. A user can enter by giving the name of an astronomical object to one of the database, which will return information on the object and link to the articles where that information was published.



The distributed electronic information system which serves the international astronomical community can be entered at a number of points, but is all linked together.

Or the user can browse the latest journal table of contents, and from there read the articles, link to the references, check the future citations, and even search for similar abstracts. Thanks to the collaboration and cooperation of nearly forty information providers, the astronomical researcher now has a convenient and powerful electronic resource.

One example from astronomy is indicative of the power of this web of resources. Say someone is reading an article about a newly discovered type of object, one which looks like a star, but may have more infrared or heat radiation than expected, and more ultraviolet light than expected, indicating a newly formed star. It may be an interesting object because planets have been discovered around it. The reader may want to get a list of other similar objects to look for planets. In the pre-electronic days, this was a long and tedious task, involving hand searching many different catalogs of data. Now it is simple. By entering one of the databases of over 2500 catalogs of astronomical objects, the reader can search for objects with excess infrared radiation, and lots of ultraviolet light. Since the data formats have been standardized within the databases, it becomes a simple task to mine the databases and come up with a comprehensive list of objects which might be expected to also have planets around them. The whole process takes five minutes. It is easy to see, even from this one example, how this electronic distributed resource is dramatically improving the capability to do astronomical research, freeing time to think about the information rather then focusing upon the manual effort of finding, retrieving and assembling it. Other examples of the astronomy information system can be found on the Web (Boyce, 2000).

What About the Long Term?

The electronic scholarly record must endure, must remain accessible to future generations of researchers. At the pace with which the formats and the browsers are changing, this is not an easy task. As described above, even a document written with the Wordstar software only fifteen years ago is now virtually unreadable. The disks are still readable, but the software to translate the information into a form we can use, or even read, no longer is available to the general user. For all intents and purposes, the information is as lost as if it were inscribed on clay bricks in ancient cuneiform. Those people who have been using the older material, and who translated it into a modern word processor at the time when the conversion software was commonly available are fine. Those who let their old Wordstar files languish unconverted now find themselves out of luck.

There are two lessons in this story. One, it takes careful management of the electronic material to keep it available and accessible. Two, the lifetime of the medium is now longer than the lifetime of either the storage format or the conversion software. Managing the electronic resources, and migrating them to new formats, transforming them to new standards, and making them work effectively with new Web browsing software can either be simple or a difficult, expensive chore. The key is in how the material is prepared in the first place.

If the material is prepared in a robust standardized format – say in a good version of SGML – which preserves all the information needed to reproduce the article, then it becomes relatively easy to develop translation software which will automatically prepare the versions which the public sees on the screen. It also is trivial to change that translation software to make new versions which are compatible with new versions of the browsers. However, if the primary version of an article is in PDF, HTML or some other presentation-oriented format which does not include all the information required to reproduce the article, than the translation will eventually fail.

Maintaining the access to the knowledge contained in electronic journals takes more than preserving the storage medium and migrating the information to new formats as they evolve. Good electronic journals need to be updated, to refer to future citations as they accumulate, to point to corrections and addenda (as appropriate), and to take advantage of new electronic tools as they become available. As an example of the latter, we seem to be edging closer to effective machine translation. When that day arrives, it will be much more likely that we will be able to translate an article and its associated metadata if it is coded in SGML than if it is just a PDF page image.

A New Electronic-centric Publishing Process

For the electronic journal, the future is clear. The production process must be totally redesigned to provide the electronic SGML version as early in the process as possible. From this archival-quality electronic version, the other manifestations of the article can be derived by automatic translation – the paper version, the screen version, and the PDF version for local printout by the reader. The electronic version cannot easily be created at the end, after using the old fashioned process to produce the paper pages. As some publishers have found out, waiting until the end to make the electronic version complicates the electronic production and adds increased costs. The experience at The University of Chicago Press illustrates that significant savings can be achieved by re-engineering the production process to translate into SGML at the beginning, incorporating automated tools to streamline the production process, insert links, check validity, and do a myriad of copyediting chores.

Once we have an archivally robust master version of every article in SGML, and have developed the translation software needed to produce the browser versions, it becomes trivial for the publisher to make changes in those programs and to remake – "republish," if you will – the entire set of screen versions to take advantage of advancing browser capabilities. This has been done several times for the AAS journals at very little cost, much less than one percent of the yearly operating budget.

The Enormous Complexity of an Electronic Journal

Another factor that is not well understood is the complexity of an electronic journal. For several hundred years, a journal was pages of paper, an independent and self-contained object. However, an electronic journal is composed of an electronic master archival version, plus several public versions, a whole library of special characters, graphics files, tabular files, a name resolver system to ensure that links are permanent, and a whole bevy of scripts and program fragments to make it all work together. The 25,000 paper pages of a year of the Astrophysical Journal actually require about 250,000 different files to hold all the pieces. An electronic journal is, in fact, a very complex system to deliver electronic information.

What does this mean for archiving the electronic material? First, it means that very few libraries could actually undertake to maintain electronic journals from one publisher, let alone the wide variety of different formats that would be encountered from different publishers. The library is simply not able, in the electronic world, to serve as the archive of the material. That leaves it up to the publisher. Only the publisher will have the expertise to be able to maintain and update the journal at a minimum cost. And, if they have not built the maintenance of the journal into their electronic design, they won't be able to do it either.

The more difficult problem is getting the publisher to commit to maintaining a journal. Then there is the question of whether the publisher can be trusted to live up to the commitment. The American Astronomical Society and the American Physical Society have made the commitment, and since each are 101 years old this year, they are very likely to honor the commitment. On the other end of the scale, the early announcements from Springer Science that they will only guarantee to keep the electronic material up for two years does not inspire confidence. And, as should be abundantly clear by now, the paper version of a good electronic journal is not complete enough to serve as the archive. While it is not clear who will be responsible for preserving the electronic material, it is clear that the traditional roles will change. Libraries probably can not do it alone any more, and publishers, who have been used to abandoning any responsibility once they ship the issue or the volume, will have to play a continuing role in the maintenance of the material.

If It's Not on the Web, It Doesn't Exist

What is even worse, the scholarly community must have the material in electronic form. Within astronomy, with most of its material so conveniently available on the Web, the electronic material is all that is ever consulted by the majority of the young astronomers. In a recent paper, librarians Stevens-Rayburn and Bouton (1998) noted that to most astronomers now, "If it's not on the Web, it doesn't exist at all." It is a mark of the coming times, and it shows how deeply we are committed to the electronic future. We have entered, almost without realizing it, into a new paradigm for the exchange of scholarly information. In this new regime, the scholarly journals are but one component of the suite of available information resources. And, with abundant links and easy interoperability, the boundary between the journals, the databases, and other information providers has grown completely indistinct. Is it not apparent any more where the journals stop and the databases begin. And, indeed, it doesn't matter.

Hints of the Future

With five years of experience in the new interlinked information landscape, we have developed a few insights into what the future will hold for scholarly information exchange.

First, we are beginning to see screen versions of articles formatted upon demand. The standards of the two major browsers are diverging. It is impossible to take advantage of the latest capabilities of Netscape and Microsoft Internet Explorer unless the material is formatted specifically for each browser. Some Web sites are doing that. The major publishers will eventually have to follow suit. Additionally, the users will want to structure the material they receive to suit their own needs, omitting material they have no need of. It will not be many years before the "journal" as seen by one reader will look entirely different that the same "journal" seen by another reader with different interests.

Second, there will be more and more "best current value" databases. We already have the human genome and protein molecule databases where the latest information is placed as soon as it becomes available. They take advantage of the ability of the Web to distribute information rapidly on a global scale. Contributors to these databases currently earn a high reputation among their peers, the same as they used to get for publishing in a peer reviewed journal. The databases are always up to date with the best current measurements. And the latest results are as available to the researcher in a small third world institution as to the most advanced, large research center in the United States. This sort of activity will increase. Databases of the "latest info" will range from sites that give the position of the moon and planets in tonight's sky (prepared for the public), to comprehensive collections of tree ring age dating (prepared for the specialist).

Third, the flood of information will continue to increase. The manual scanning which now serves many scholars – especially those which use the preprint servers – will soon not suffice as a reasonable mechanism, either to keep up with their field, or to find relevant information. Effective metadata, and intelligent search and retrieval tools will have to be developed. Users of the Internet will want relevant information immediately, and with the irrelevant information stripped out. The thousands of pages retrieved by today's general search engines will soon no longer suffice. Even though work on the use of neural network methods and intelligent classification engines is still in its infancy, enough progress has been made that the first working tools are now available. They are still cumbersome, but herald exciting possibilities for more effective information search and retrieval in the future.

In summary, it is a very different information world than we have been used to. Many publishers and many users have been slow to recognize just how different it is and to react accordingly. Even many users have been slow to leave their comfortable working styles. But, within a field such as astronomy, once the transition takes hold, it goes with amazing speed. Once the train starts moving without you, it is hard to climb back on – and there is no way to hold it back.

References

P. BOYCE (1996a) "A Successful Electronic Scholarly Journal From a Small Society" Presented at ICSU Press – UNESCO Expert Conference on Electronic Publishing in Science, Paris, France, 19–23 February, 1996. http://www.aas.org/~pboyce/epubs/icsu_art.html

P. BOYCE (1996b) "Building a Peer Reviewed Scientific Journal on the Internet," *Computers in Physics*, v. 10, p. 216.

P. BOYCE (2000) "What Does the Future Hold? Ask an Astronomer," Talk given at NC Serials Conference, Chapel Hill, March 16, 2000. http://www.aas.org/~pboyce/epubs/NCSerials2000/NC2000.html

M.J. KURZ, et. al. (2000) "The NASA Astrophysics Data System: Overview", *Astron & Astrophys Suppl.*, 143. 41.

S. STEVENS-RAYBURN and E. BOUTON (1998) "'If it's not on the Web, it doesn't exist at all': Electronic Information Resources – Myth and Reality" in *Library and Information Services in Astronomy III*, ASP Conference Series, Vol. 153, 1998 Editors: U. Grothkopf, H. Andernach, S. Stevens Rayburn, and M. Gomez. http://www.stsci.edu/stsci/meetings/lisa3/stevens_rayburns.html