From Photostats to Home Pages on the World Wide Web: A Tutorial on How to Create Your Electronic Archive

By Eugene Garfield

The Scientist 13[4]:14. February 15. 1999

It was Derek De Solla Price who in 1963 said that "80 to 90 percent of all the scientists that have ever lived are alive now."[1] While De Solla died just 20 years later, most of his contemporaries have survived. A majority of these researchers were, at one time or another, readers of Current Contents (CC).

The success of CC was due to three critical factors--timeliness, its multi-disciplinary nature, and access to author addresses. Its address directory was estimated at one time to generate as many as 15 million reprint requests per year.[2] The large-scale exchange of reprints has become an almost bygone culture. Widely available document delivery mechanisms and increasingly, the World Wide Web, have displaced routine requests for reprints. Nevertheless, many of us still enjoy the convenience of the reprint, and when possible, the pleasure of a note and autograph by the author.[3]

The reprint culture supported thousands of worldwide "invisible colleges."[4] It would be fun and nostalgic to dwell on the many facets of the reprint culture. During the Cold War, when contacts between scientists worldwide were largely maintained by reprint exchanges, the children of scientists enjoyed collecting all those foreign stamps from reprint request cards. It is an interesting commentary on the present state of affairs that among the most common requests for reprints even today are those requests from Cuba and the former Soviet bloc nations.[5]

As indicated above, the reprint request by snail mail is rapidly being displaced by the convenience of the Web. Technology has its rewards, but we also pay a price. The Web, however, has increased the opportunity for more direct, personal, and almost instant contact. Recognizing the ultimate role that the Internet would play in the exchange of information and "electronic reprints," The Scientist became the first full-text scientific journal continuously published simultaneously in print and on the Internet.[6] Even before there was a Web, The Scientist was available free of charge on the Internet via the AT&T Gopher site supported by the National Science Foundation. File Transfer Protocol (FTP) was an expression often heard then. And in that primitive era we were limited to the use of American Standard Code for Information Exchange (ASCII) text. When the Web was introduced, we simultaneously produced the ASCII and HyperText Markup Language (HTML) versions of The Scientist. The latter version resided on a Web server at the University of Pennsylvania Library, where it has remained (www.the-scientist.library.upenn.edu).

We recently completed converting most of the old ASCII files to HTML. Thus electronic "reprints" of our articles covering the last 10 years are instantly available. It would be impossible to

determine the number of "reprints" that have been generated from this archive.

Over the many years that *The Scientist* has been available electronically, I was frustrated that I could not provide similar access to the many articles and *CC* essays I had published over a 40-year period. Indeed, to this day, I continue to receive reprint requests even though my last *CC* essay was published in 1995. It's amusing to meet contacts who say they still read them.

A few years ago my assistant, Meher Mistry, began the arduous process of creating my personal Web site. It is located on the same University of Pennsylvania server as *The Scientist* archive at: http://165.123.33.33/eugene_garfield. (This abbreviated version of the URL is based on the IP address. The more descriptive URL is: www.the-scientist.library.upenn.edu/eugene_garfield.)

As is not uncommon today, my Web site C.V. includes not only biographical data but also a complete bibliography of everything I have ever published. The main listing is the year-by-year cumulative contents page for 15 volumes of *Essays of an Information Scientist*. The Essays contain reprints of every *CC* essay published between 1962 and 1993, reprints of my editorials for *The Scientist* (1986-1993), and numerous reprints of articles by me and others. The *CC* essays published in 1994 and 1995 are listed separately on my Web site and are hot-linked to the *Institute for Scientific Information* (ISI) Web site (www.isinet.com/hot/essays). These electronic bibliographies have created a renewed interest in the Essays volumes that are still sold by ISI in Philadelphia.

It is a sad fact that most of the libraries that purchased the first five or six volumes in this series are not aware that they were continued annually until 1993. Even though they all have ISBNs, the Library of Congress (LC) failed to catalog them. Hopefully LC will catch up with Amazon.com in the near future. Amazon now lists and sells all 15 volumes.

While access to my bibliography is an important first step, we had long ago decided to make all *CC* editorial material available in full text electronically. So we contacted Crossaig, Ltd., a subsidiary of ISI in Helensburgh, Scotland. This firm is in the business of converting journal files from print to electronic format. They, and many other firms worldwide today, "digitize" printed journals and other documents en masse.

To facilitate scanning, a copy of the Essays volumes was torn apart and the loose pages were fed mechanically into a commercial scanner to create Portable Document Format (PDF) files. A PDF file provides a "facsimile" version of the page scanned. So when you call up a PDF file of an article on the Web, it looks identical to the original. Free Adobe Acrobat Reader software permits you to read and print out a copy. With Optical Character Recognition (OCR) software Crossaig also provided us with a "dirty" ASCII text that is hidden but resides behind the facsimile image. (Essentially ASCII is a text version of the document that does not retain any typesetting codes necessary to produce the stylized look of the typeset original.)

The accuracy of the hidden ASCII text, which is used for full-text searching, depends upon the quality of the original copy, the resolution at which it is scanned, and the discrimination of the OCR software. While it is almost impossible to obtain 100 percent accuracy in OCR scanning, it is possible to achieve a level of accuracy that is acceptable for searching since there is considerable redundancy in most scientific or scholarly texts. Perfect transcription requires human proofreading and keying.

The ASCII text is also used to create the HTML version that is used by Web browsers. So when you access most scientific journals online today, you have the option of printing reprints from PDF original format via the Adobe Acrobat software, or you can quickly print the less formatted, but bulkier, HTML version. Creating the HTML version also involves the arduous task of hot-linking footnotes and superscripts. So when you encounter a cited reference, there is a hypertext link to the full text of the cited essay or other work.

It is the utopian expectation of those who live in cyberspace that eventually most researchers will create Web sites containing the full text of all their papers and books or links to them wherever they reside. The social, economic, and scholarly impact of this development has major consequences for the future. So, before you write me for a reprint, please check my Web site and then contact me at egarfield@the-scientist.com.

Eugene Garfield is president and editor-in-chief of *The Scientist.*

1. D.J.D. Price, Little Science, Big Science ... and Beyond, New York, Columbia University Press, 1986, page 1.

2. E. Garfield, "Reprint Exchange. 1. The multimillion dollar problem ordinaire," *Essays of an Information Scientist*, 1:359-60, 1972. http://165.123.33.33/eugene_garfield/essays/V1p359y1962-73.pdf

3. E. Garfield, "Is the 'free reprint system' free and/or obsolete?" *Essays of an Information Scientist*, 1:10-11, 1965. http://165.123.33.33/eugene_garfield/essays/V1p010y1962-73.pdf

4. D.J.D. Price and D.D. Beaver, "Collaboration is an invisible college," *American Psychologist*, 21:1011-8, 1966.

5. E. Garfield, "Can reprint requests serve as a new form of international currency for the scientific community?" *Essays of an Information Scientist*, 307-15, 1977. http://165.123.33.33/ eugene_garfield/essays/v3p307y1977-78.pdf

6. S. Borman, "Advances in electronic publishing herald changes for scientists," *Chemical & Engineering News*, 71[24]:10+, June 14, 1993.

The Scientist 13[4]:14, Feb. 15, 1999