Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Mining the Web: How Useful is the Global Public Library?

Rudolf Albrecht[1]

Space Telescope European Coordinating Facility, European Southern Observatory, Karl Schwarzschild Str. 2, Garching, Germany
ralbrech@eso.org

Peter B. Boyce[2]

PBoyce Associates, Nantucket, MA, USA
pboyce@aas.org

Abstract. The Web has matured into the most universal source of information. At this point in time it still suffers from the fact that finding the pertinent information, even if available, is difficult for a variety of reasons. This paper explores the usefulness of the Web for professional scientists and the interested public. Using examples, we examine the reliability and completeness of the information on subjects that are well known, and on cutting edge science. A recent survey by the AAS[3] shows that only about 1% of the articles that astronomers read are found through standard Web search engines. Specialized services like the ADS are more popular (20%), but information passed on by colleagues is the most popular among astronomers (21% of the articles are found that way).

1. Introduction

The World Wide Web was not adopted by astronomers, nor, in fact, by any other group of users, on the basis of a detailed analysis of the requirements, a careful study of the available options, and an eventual implementation in the most cost-effective manner. Instead it seems to have filled a need which the users did not know existed. It shares this property with many other significant innovations: the photocopier, the fax, laptops, email, etc. Had the users been asked, before the introduction of these devices, whether they needed them for their work, the answer would have been "no". But in all cases, the capabilities became indispensable within a short time after they became available.

---

[1]Space Telescope Division, Research and Science Support Department, European Space Agency

[2]Senior Consultant to the American Astronomical Society

[3]Supported in part by a grant from NASA to the AAS

This is especially true in the field of astronomy. In fact, during the early days of the Web, around 1994, astronomers were, for a short time, the largest single group of users. This fact was noted with some amazement by the Wall Street Journal. It seems to have exceeded the imagination of the journalists.

The lesson in all this is that while it is usually good management practice to identify requirements and to produce and implement solutions in a controlled manner, there are cases when it is more efficient to quickly react if and when the market produces a solution. This situation still exists with the Web: any attempt on the part of the astronomers to deviate from the de-facto standards represented by the major browsers, any attempt to produce astronomy-specific one-of-a-kind solutions would lead to a dead-end development. Thus, the challenge is to attempt to maximize the usefulness of the available tools to the scientists, either by using them in an appropriate manner, or by utilizing public services and off-the-shelf software.

2. Status Quo

Given the market forces that drive the Web - mainly entertainment and commercial advertising - the Web is ideally suited for the dissemination of information, and, among professionals, for documentation and publications. In fact, if an organization is not present on the Web it "virtually" does not exist. Even secret societies like the Freemasons and the Ku Klux Klan have web pages.

In the scientific community we have come to taking for granted access to services like data archives and on-line information servers. Various computing services, for example exposure time calculators for scientific instruments, have been implemented either on servers (through various scripts) or on the client (through Java). This development will continue to be driven by what is now called the GRID.

Astronomy has also pioneered special services like the ADS and preprint delivery services. This has been heralded as proof that astronomers are somehow more internationally minded than other professionals. It might, however, just be caused by the lack of commercial interest in astronomical information.

3. Doing Science

All recent discoveries can be found on the Web in full: Gamma Ray Bursts (59,000 pages), Extrasolar Planets (23,000 pages), Neutrino Oscillations (25,500 pages). These discoveries happened after the introduction of the Web, so their full history has been documented on web pages.

Even fields that have been around longer are well represented. While not fully documented, we see review pages of historical perspective. "Active galactic nuclei" yields 36,000 pages, while quasars, are represented by 96,000 pages.

However, "Astrometric solution", an example of early 20th century astronomy, is represented by only 4800 pages, and, as this is true for other top-ranked pages, they mainly contain applications of the technique rather than an explanation of it. However, there is a growing body of textbook web sites.

An increasing number of scientists maintain personal web pages that hold information on their research activities. Thus, more so than at any time before, it is possible to find out who is doing what, and to locate and to contact experts in the different fields instantly. Interaction among scientists has become easy.

Virtual Institutes, in the case of astronomy Virtual Observatories, are the next logical step. Starting from the simple consideration that the data that have been collected by expensive instruments must be fully exploited, and driven by science requirements like epoch coverage and the need for pre-discovery data, distributed science data archives are in the process of being interconnected. The goal of this connection is to make it possible for any computer to access any data. This, in combination with publicly available data analysis software, will make it possible to do meaningful science even for scientists who have been traditionally disadvantaged because they, or their countries, had no access to adequate observing facilities. Virtual Observatories have been started in Europe and in the US.

We conclude that now astronomy can be studied, astronomical data accessed and analyzed, and meaningful astronomical research conducted by exclusively using the Web.

4. Electronic Publishing

E-journals are increasingly being made available on the Web. A separate opinion survey of AAS members illustrates just how valuable they believe the electronic journals to be for their work, both for keeping up with current developments as well as for obtaining definitive information. Seventy-two percent of them rate electronic journals as either 'very useful' or 'essential' for keeping up with recent developments. When seeking definitive information, astronomers value the e-journals even more highly. Virtually all astronomers (96%) rate e-journals as either "essential" or "very useful" for delivering definitive results. This overwhelming approval rate reflects the effectiveness of the whole electronic information system used in astronomy, particularly seamless links between the e-journals and the highly effective NASA/ADS (an A&I service and a database of historical full text journal articles). The same survey indicated that 97% of AAS members knew about the ADS and over 50% of them use it at least every other day. Twenty-seven percent of AAS members use the ADS every day. ADS usage statistics confirm this level of activity.

When asked in which manner the readers found the last article they read, the answers were: Found by browsing 20.4%, found because it was cited in an article or e-print 13.0%, found by searching 42.5%, or "A colleague told me about it" (including mention in a colloquium) 21.1%. Of the "Search" group, the largest fraction (20.2%) used the ADS. Only 1.2% used standard Web search engines such as Google.

5. Challenges

The most obvious challenge of the Web is information overload. Clearly, if a topic is represented by several thousand web pages it is totally impossible to read through all of them and to distill the pertinent information from them.

The ultimate challenge of the Web is languages: one quarter of the world's population is Chinese, and soon many of them will be on the Web, so Chinese will be the largest single language group. However, to expect the rest of the world to study Chinese is not realistic. To expect the Chinese to learn English is a bit more realistic (the scientists and the professionals already speak English), but,ultimately, the problem will have to be solved in a different manner: it might just happen that the problem of machine translation will be solved because of the need to translate Chinese web pages.

Junk or irrelevant content is another problem. Garbage information clogs the net and hides information.

Search engines are as old as the Web. In fact,the speed and ability to locate information has become very impressive. Using different feedback mechanisms, the pages found are ranked in order of approximate relevance, in most cases eminently successfully. This is another example of a solution produced by the market that would have been impossible, or too expensive, to produce in any other way.

6. The Future

There is a need for special search engines, capable of sorting by content. Experiments based on neural nets have been done. The Vivisimo clustering engine (http://vivisimo.com/html/products.html), which takes the output of search engines and spontaneously organizes the results into a meaningful hierarchy of folders, is already being used with promising results.

There is an obvious need for a personalized "agent" to scan the Web for interesting content. While sounding far fetched, this selection is already possible for special purposes, such as scanning the news on the CNN Web pages.

Other mechanisms (moderated Web sites, or a system of Rating of Web sites based on the usefulness of their content) are desirable, but are difficult to implement.

Today's e-journals are the first steps to full Web publishing. By this we mean more than just replacing paper by electrons, we mean abandoning the classical concept of publishing in the form we know it. In astronomy and most sciences, information is packaged as an article, i.e. a body of prose, formulae and graphics, arranged by the author. Web publishing will, instead, be done by making original contributions to one or more tightly linked databases that hold all the information for a field of science. Such contributions may take the form of traditional articles, but are more likely to be information cast in other formats.

This change in publication will be paralleled by significant changes in our scientific libraries. In spite of the technological advances, today's libraries are operated in much the same way that libraries have been operated for centuries. Libraries house collections for preservation and access. Both of these require stability, but this stability is at odds with the need to capture dynamic information. The librarians, like their colleagues before them, are expert in organizing collections of bulk media, but they are much less aware of their content and its relevance. In small specialties such as astronomy, the librarians, while not being scientists themselves, can, nevertheless, become remarkably proficient at

understanding the importance of the various materials in their care and the information they contain.

In the future, librarians will become information managers, who operate the local access points to the global data base, and to other data bases, and who keep the system updated to reflect the state of the art of the field.