DRTC Workshop on Information Management
6-8 January 1999

PAPER: EA

CHAOS! THY NAME IS INTERNET

ARD Prasad, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore-59 E-mail: ardp@hotmail.com

[*Of the recent developments there is hardly anything comparable to the Internet when it comes to the impact on public life. But again, it can at best described as immature. Firstly, it is too slow, has unreliable connectivity and is almost on the verge of collapse with more and more multimedia. (Partly, the World Wide Web was responsible for this, as it had eaten away other services on the Net like Archie, gopher etc.). Secondly, the Net is the best example of chaos. Yes! Internet has a long way to go.*]

1. CONNECTIVITY ISSUES

1.1 Data Transfer

The quality of the Internet connectivity reminds us of the early days of TV broadcasting, where things used and appear and disappear the next moment, with too many 'sorry' messages and horizontal lines popping up on the screen. Internet has many bottlenecks in the data transmission. If you have Internet connectivity via LAN, the data transfer can get slowed by LAN (as you might be using hubs rather than switches), your server may be slow, your telephone line has lot of noise resulting in more and more bad packets, your ISP has less band width or the server on the other side might be busy. Mostly, the problem is that of POTS!!(Plain Old Telephone System). There are three possible solutions to the problem

1. **Cable modems**: Using the existing cable TV network, it possible for the customers to have 30Mbs bandwidth, i.e. if there are 30 subscribers, each will get 1Mbps of transfer rate. An advantage is that it integrates TV and Internet technologies. The disadvantage is only one of them can be used at a time, either the TV or the Internet.

2. **Low Earth Orbit Satellites**: This technology uses a network of Low Earth Orbit satellites, which are placed about 1,000 Km above the earth. The hookup to the satellite is through a portable transmitter/receiver. The data from the system is sent to the satellite through the transmitter/receiver and the satellite sends the data to the closest similar LEO satellite among the satellites surrounding the earth. The data moves across the satellites until it is received by the satellite having link with the transmitter/receiver of the destination system. The advantage of the system is that any portable system can be hooked up to the Internet from any corner of the globe. The disadvantage is that it is costlier than the other alternatives. Although the connect speed is 2.4 kbps in the existing Iridium system (1), it is expected to be upto 64Mps in the case of Teledesic system (2).

3. **ADSL Modems**: Asynchronous Digital Subscriber Line Modem, although slower than cable modems have high connectivity speed, as the bandwidth is not shared. One advantage of ADSL modems is that it uses plain copper telephone wire by utilizing the vacant frequencies in the range of 0 to 2 MHz in the copper wires. It promises a 9 Mbps of download speed and 1.5 Mbps of upload speed. This kind of connectivity requires an ADSL modem on the user side and an ADSL switch at the ISP's location. But the disadvantage of the system is that the signal becomes weak if the ISP is more than 3 Kms away.

In addition, the ISPs bandwidth also has to be improved to a large extent for higher speeds in the connectivity. The existing speed of the backbones is 622Mbps which equals the capacity of 18,660 modems connected at 33.6Kpbs. But is expected that by

the year 2000, the capacity will be 2.5Gbps and by 2002 it will be in Terra bits per second and can serve 30 million modems!

1.2 IP Numbering

At present IP numbers constitute of a four-byte code. With the constant growth of the Internet, with thousands of servers and millions of users being added every year, soon there will be a dearth of IP addresses. In fact, one of the DEC's technology forecasting reports mentions that by 2020 there will be more IP numbers required than the human population on the globe! Ofcourse, it does not mean that everybody will be prosperous by then. It only means that many other machinery will be remotely controlled and they require IP numbering. Such machinery includes not only the industrial machinery, but also home appliances. At present, the number of Internet subscribers is growing by 69,000 per week and it is expected that the total number of Internet subscriber will be 137 million by the next millennium. Theoretically, the present IP4 protocol which allows a 4 byte address code (i.e. 32 bits) can have 4,294 million addresses, but in practice it is much lower than that. The problem with the four byte address code is that it is OP4 protocol dependent. It is expected that IP6 will soon replace IP4. IP6(3) allows 128 bit address code with better security options.

2. INFORMATION ISSUES

The most often heard criticism against Internet is that of questionable *authenticity of information* on the Net. But I feel that even the information in the print media is not authentic. How many newspapers carry the truth, and what is truth for that matter? We may mostly agree or disagree with some information, depending on our interests. If something suits us we accept it. In a way, Internet only provides a different medium for the information. With regard to authenticity, it only reflects us and our society, its good and evil, just as the paper medium. One may dare to question how authentic the articles

published in a technical journal. We know that journals carry information that varies from mediocre to good scientific work. On the whole, it is a misplaced criticism.

One of the serious problems with regard to the information on the Net is its euphemeral nature. "What-is-there-today-is-not-there-tomorrow" kind of situation. Preserving what was once there for posterity is of less concern. This problem exists even for printed documents, but of less serious proportion. We have a feeling that a printed document if not readily available, but with a little or more effort, it can be found sooner or later or for that matter never, though we continue to hope. With Internet everything appears to be current and if site is not maintained, we tend to say it is outdated. Unless, there is guaranty that what appears today on the net will be still available in the future, historical studies would become difficult. Unless, there are archival sites, which collect all the important information and archive them, we may loose the information forever. Ofcourse, we do not collect every printed information for posterity. I always wonder about the proportion of information that is printed and preserved for long. For the detectives of the past (historians) we know, even garbage yard of an ancient town provides lots of information on that society.

In humanities and social sciences, we have a tendency to quote a thinker or author verbatim and refer to an author's statement by page number. But the HTML based Internet documents do not have a page concept as the display depends on the screen size and window size, the exception being the PDF documents, which are very few in number.

This kind of uncertainty on Internet is not confined to the information; this exists with the very site. This situation leads to another problem with regard to referencing documents on the Net. In spite of the feeling that much of the content of Internet is of commercial nature (for that matter, it is said that 30% of e-mail is commercial), we still have a lot of technical information on the net on various subjects. At times, I feel we get more recent information on the net than in journals. It is not difficult do give a description of the document we referred, like its URL, but will that URL points to the same information we have referred after a period of time? And at times that URL itself might not exist any

more. There are a few attempts to overcome this problem by introducing various other methods like PURL (Persistent URL)(4), URN(5) etc. But, it requires a little more time to find out the best alternative. It is only a matter of time that we get used to Internet and accept it with its strengths and weaknesses, as with the case of many things in life.

Yet another problem is the very definition of a document on Internet. There is no particular approach followed in building an Internet document. For a particular piece of information some webmasters put the information in one page, some divide into a number of pages or files, and in some case it might be a distributed document residing on more than one server that are geographically miles apart, only giving the impression to the user that it is a single document. We do not know how to give reference to distributed documents having more than one URL (the documents has tendency of being amorphous). Even, if we mention more than one URL, the question is in what order the URLs are to be presented. Remember, after all they are hypertext documents and are very much non-linear. As the electronic media opens up a few more opportunities in presenting information, we use a web page in many a way. We have more or less some standard practices with regard to print media in identifying various types of documents like monographs, cartographic materials, serials etc. We have yet to identify various types of documents on the net. From the cataloguer point of view, we should be able to provide a bibliographic description of the Internet documents. In fact, the UNIMARC(6) came out with content designators of Internet documents, and makes a clear distinction between the documents covered by ISBD-CF or Computer Files of AACR2, as these deal with mostly the sources of information giving off-line data, although basically the information is in digital form.

Search Engines: It is well recognized that there are millions of documents on the Net. As the Net is non-authoritative, so is the information unorganized. Within a library, the library staffs organize the information following standards in classification and cataloguing. But if we perceive the Internet as a global digital library, it really requires high discipline to accept one single set of standards. As Internet happens to be a

distributed data system, expect such a mythical unified structure. There can only be lot of compromises.

There is lot of information on every conceivable topic on the Net. The problem is how to access right information within a reasonable time. The present day search engines and meta search engines can utmost be rated shoddy. There is much to be desired. Firstly, they may be good in recall and naturally very low in precision. If, thousands of records, or for that matter even hundreds of records are retrieved and of which almost equal number is useless or irrelevant, they can be hardly called search engines, however fashionable the word is. The lack of indexing knowledge is too obvious. Even the output of a search has no format of any kind. Generally, the robots of search engines collect the keywords from the first few lines of a web page; the history of indexing proves that context sensitiveness is very much important in indexing. If only the documents have a structure, this problem can be overcome.

The present HTML has a limited set of tags and is not designed keeping bibliographic descriptive elements in mind. However, XML (eXtended Mark up Language) promises to allow us to design our own tags(7). In future, hopefully there will be general understanding to follow a set of standard tags based on MARCs in preparing web documents, so that the search engines collect information from these tags to build indexes and also generate a standard format for the output.

3. CONCLUSION

The absence of the librarian is very much apparent in the Net. Wherever there is information, the librarian has a role in it. Dealing with massive quantities of information, the library profession has developed many tools and techniques and if the Internet ignores these tools, one day they have to reinvent them painstakingly. If the library people ignore Internet, their job will be done by somebody else. Hope there lies better future in the era of I2 (Internet 2)(8,9) for Internet and the library profession, as they can not do away with each other.

4. FOR SURFING

1. Iridium: Calling Planet Earth. www.iridium.com
2. The Teledesic Network. www.teledesic.com
3. Introduction to and History of IP6. www.tcm.hut.fi/Opinnot/Tik-110.551/1996/Story.html
4. Introduction to Persistent Uniform resource Locators. http://.purl.oclc.org
5. Uniform Resource Names: An overview. www.acl.lanl.gov
6. Presentation of UNIMARC on the Web: New fields including the one for Electronic Resources. www.ifla.org/IV/ifla64/110-161e.htm
7. W3C World wide web consortium. www.w3.org
8. Internet 2, by University corporation for Advanced Development: www.Internet2.edu
9. Next Generation Internet (NGI) Initiative. www.ngi.gov