CHAPTER II

LITERATURE REVIEW

2.1 Longevity of Web Sites and Web Pages
2.2 Evaluation of Web Sites Using Quality Criteria
2.3 Web Impact Factor
2.4 PageRank
2.5 Issues Arising from Literature Review
CHAPTER II

LITERATURE REVIEW

This research is concerned with methodologies that can be used in evaluation of Web pages and Web sites, with emphasis on their information content rather than aspects such as design or aesthetics. It aims at suggesting ways for an individual LIC to build collections of Web pages and Web sites (locally held or remotely accessed) that have been evaluated using selection criteria formally laid down.

The literature review covers research that has taken place in the area of longevity of Web resources, changes they undergo over time, and their risk assessment, all aspects necessary to be studied for their collection building. Also reviewed is research in formulation of quality criteria for this new type of information carrier in order to study which of the criteria established for evaluating traditional print resources can be applied and which new criteria to suit the new medium will be required. Keeping stride with the developments in Internet-based information, a new branch ‘webometrics’ has come up, which offers alternative methods of assessing the quality of Web sites. The related literature also forms a part of this review. Searches for relevant literature were carried out using LISA online database. Along with books and journal articles, a large number of Internet-based resources have been used in the present study, since the concepts studied are of relatively recent origin, and themselves deal with Internet-based resources.

2.1 LONGEVITY OF WEB SITES AND WEB PAGES

2.1.1 Web Page and Web Site Persistence and Change

Collecting and managing Web-based resources is a challenge, due to their dynamic nature – some might be available one minute and gone the other, and some undergo constant addition, deletion or modification of their information content.
Pioneering work to study the dynamics of Web sites and Web pages has been done by Wallace Koehler, following his MS Thesis titled *Web Page and Web Site Persistence and Change: A Longitudinal Study* submitted to the University of Tennessee in 1997. His study is a preliminary exploration into the mortality rates of Web pages and Web sites. He tries to answer questions such as how permanent they are, what is their death rate, how often they move, how often they are resurrected or come back after being absent for a while. Another characteristic studied is constance – how constant are Web pages and Web sites. Two types of changes are studied, content and structural (hypertext link structure). Whether different classes of Web documents show different permanence and constance behaviour is studied.

Koehler examined a sample set of Web sites and Web pages from time to time over a period of six years. Koehler (1) reports findings of a one-year study of Web sites and Web pages. A random selection of 361 URLs was done for the study. Web site data was collected thrice, relating to number and size in bytes of various types of Web objects - text documents, graphics, audios and videos. Web page data such as size in kilobytes (to capture change in content), number of new links and changed items linked to the target document (to capture change in structure), was collected weekly. URLs were explored since not only can they help identify authority, publisher and quality, but also may offer predictive value for permanence and continuity.

An overall increase in average Web page and Web site byte weight, or “byte creep” was the general trend. The main object types contained in Web sites were text (21.2%), file retrieval (17.4%) and graphics (13.1%). Koehler estimates the half-life of a Web site to be approximately 2.9 years. Larger and denser Web sites are more likely to persist than the smaller and dispersed. The permanence also depends upon the ‘publisher’ of the site. At the end of one year, 74.7% of the Web site sample was persistent. More than 97% of Web sites underwent some kind of change after six months, while after one year, more than 99% had changed.
Web pages disappeared at a rate of 0.5% every week, with Web pages on smaller Web sites becoming unavailable more often than those on larger Web sites. 69% Web pages were persistent at the end of the year. 98.3% Web pages changed to some degree after six months, and 99.1% after one year. Constance of Web pages also depended upon the publisher of the site. Content change was seen most often. Changes to existing hypertext link structure occurred in case of 10-20% of the sample, while new links were added in case of 5-10% of the sample.

The study of Web document behaviour that had begun in December 1996 was extended to 106 weeks, i.e. up to January 1999 (2). At the end of the two-year period, 66% of the Web site sample and 58% of the Web page sample was persistent.

The same Web pages collected in December 1996 were followed up to February 2001, or 214 weeks (3). Of the original sample, only 34.4% Web pages were persistent at the end of the four years.

Koehler further continued the longitudinal study up to May 2003 (4), i.e. over a period of six years. The original sample of persistent pages was reduced to only 33.8% of the size. Thus, further erosion in the persistent pages was seen. However, compared to the huge loss that was observed during the first four years of study, where nearly two thirds of the original sample disappeared, the loss seemed to have slowed down. Navigation pages showed greater resilience than content pages.

Koehler’s study was related to dynamic HTML Web resources. A study that examined the persistence of static Web objects was that by Nelson and Allen (5). They studied the object persistence and availability of 1,000 digital library (DL) objects – static, document-like objects or publications such as reports, preprints, theses, etc. Twenty DLs were chosen based upon their mixture of coverage, popularity, geographic locations, and data storage architectures. They included full text archives of preprints, reprints, reports, research papers, theses, etc. from USA, Germany, Japan, UK and France.
From each DL, 50 objects in PDF or PostScript formats were chosen at random. A script checked the availability of each object three times a week over a one year period. DLs were completely unreachable 2% of the time, while individual objects were unavailable 11% of the time. Objects changed more than one kilobyte in size 5% of the time and showed at least some change from their baseline 22% of the time. The study thus found that contrary to the expectation that objects placed in a DL should persist longer than an average Web page, 31 of the original 1000 objects (3%) were no longer available, or could not be located.

2.1.2 Non-permanence of URLs

While not much literature was found dealing with constancy and permanence of Web resources, the non-permanent nature of URLs has received much attention, with studies being carried out especially regarding the use of citations or formal references to information on the Web, which are becoming increasingly common. Many journal articles, including refereed articles, contain citations to Internet sources. Cited references need to be accessible and persistent, because otherwise one of the purposes for giving citations, namely linking the written work to a much larger community, is defeated. The use of Web references in research articles has been of particular concern, given the fact that many a times they cite invalid links. This is the result of lack of persistence of Web pages and Web sites.

Harter and Kim (6) were among the first to document the impact of the ephemeral nature of the Web on citations and citation systems, as a part of the research to assess the impact of electronic journals (e-journals) on scholarly communication. They examined 279 articles from 74 scholarly e-journals published primarily in 1995. These contained 83 references to online sources (web pages, electronic personal papers, email messages, ejournal articles, news group and listserv postings, etc.). An examination of the citations to online resources revealed that between the writing of the articles (published primarily in 1995, some as early as 1993) and their analysis in 1995, almost half the references (48.2%) were no longer accessible.
Benbow (7) explores the issue of changing URLs and provides a brief analysis of the degree to which change is occurring. She found an attrition rate for Web resources of 20% and 50% over two and three year periods. She examines the range of potential solutions and provides discussion concerning the reasons for outdated and inaccurate URLs.

Carol Anne Germain (8) studied the accessibility of sixty four URLs cited in thirty one academic journal articles. The study was conducted over three years. Nine tests were carried out. 26.5% citations could not be accessed after the first year, 37.5% after the second year and 48.4% after the third year. Thus, availability of URLs declined at 11% annually. At the end of the third year, 67.7% of the original articles contained citations that could not be found.

Steve Lawrence and others (9) examine URLs contained in computer science research articles, analyzing the volume of citations, the validity of links, and the detailed nature of invalid links. 67,577 URLs were extracted from 100,826 articles. Attempt was made to access each URL in May 2000. It was observed that the percentage of invalid links contained in articles depends upon how long back the article was published – it increased from 23% in 1999 to a peak of 53% in 1994. The authors attempted to try and locate a random sample of 300 invalid URLs using manual searches employing different strategies. They were able to reduce the number of "lost" URLs to 3%. They concluded that with good search experience and abilities, and persistence in the task, most invalid Web references can be located.

2.1.3 Risk Management of Web Resources

Another methodology that has been used for studying the lifestyle and habits of Web resources is “Risk Management”. In the post 9/11 world risk management has become a business in itself, and is being applied in all fields. Risks to Web resources can be detected using automated tools and libraries can accordingly plan preservation strategies.
Anne Kenney and others (10) describe Project Prism of the Cornell University, which applied risk management techniques to Web-based materials. The nature of preservation risks in the Web environment was characterized. A risk management methodology for establishing a preservation monitoring and evaluation programme was developed. Various stages of risk management such as Risk identification; Risk classification; Risk assessment; Risk analysis and Risk management implementation were identified.

In case of a Web page, the possible indicators of high or low risk could be tracked using automated tools. Tidiness of HTML formatting, conformance to standards, document structure, presence or absence of metadata tags may indicate the level of management. High frequency of 404 – page not found errors and slow response time would indicate higher level of risk. So also, page change could be tracked – sometimes, no change at all might indicate complete lack of maintenance. On the other hand, unpredictable and large changes might indicate chaotic management. Pages that change on some predictable schedule might indicate high-integrity management.

Hypertext links could help in deducing longevity risks to the Web page. A page that links to a number of pages on the same server (intra-linked page) might be more integrated into a site and therefore at a lower risk as compared to a page that only links to pages on other servers. Pages with no links going out or coming in might be considered highly suspicious. Link volatility, or addition or updation of out-links shows that the page is being maintained and is therefore at reduced risk. The URL of a Web page can tell about the page’s provenance (origin) and management structure.

For assessing the longevity risk of a Web site, algorithms would be required for aggregating the risk metrics of its individual pages. A Web site would be at lower risk when care is taken of the hardware and software environment, administrative procedures, network configuration and maintenance, backup and archiving, etc.
Botticelli (11) carried out a study under Project Prism, in which Web sites were monitored over a year and changes in their status that may indicate short and long term risks to contents were documented. Fifty-four Web sites of political and nonprofit organizations covering Southeast Asia formed the sample. Each of the Asia sites was crawled ten times in an eight months period between April 2002 and January 2003 using an automated tool.

Some dangers to content loss were seen in the form that four sites disappeared, sites lost up to 20% of pages on average across ten crawls, with five sites losing 25-40% of their pages in the last crawl. The sudden shrinkage in size of one site (91%), and sudden growth of two sites (34% and 56%) from the previous month could be a sign of organizational changes, which might put content at risk. 32 different formats for MIME type were discovered, out of which only a few are widely used. Use of non-standard formats leads to a possibility of higher risk.

2.2 EVALUATION OF WEB SITES USING QUALITY CRITERIA

2.2.1 Identification of Quality Criteria

During the course of the study it was observed that Web selection criteria are rarely found in professional journals and other print formats. Instead, they are commonly found as Web-based documents, embedded in scope statements and collection policies of subject gateways and in the background papers on resource discovery systems. Also, a number of libraries, especially academic libraries (university and college libraries) of USA and UK have information on their Web pages as to how to evaluate Web sites.

Web evaluation resources found on the Web were extremely diverse in their styles. They ranged from one-page checklists to detailed descriptions ranging in length from one to several pages. Some provided guidelines to be applied for evaluation, others listed questions to be asked while evaluating, and yet others provided worksheets to be filled in order to help in evaluation.
There seems to have been little concern about quality in the online environment until about 1990. Juntunen, et.al. (12) report that a review of the literature in 1989 found that there were no guidelines, lists of criteria, or other tools for evaluation. This appeared to be changing. In Jasco’s words, the 1990s had become the “decade of quality” (13). However, Jasco’s literature review does not refer to the World Wide Web at all.

 Barely three years later, a considerable amount of literature was generated, specifically on the evaluation of Web resources, as evident from the bibliography by Auer (14), one of the most comprehensive bibliographies on the subject. This is an indication of the world taking note of the importance of applying the tried and tested document evaluation techniques to the new Web-based resources such as Web pages and Web sites. The bibliography was originally prepared at Virginia Tech by Auer for a panel discussion at a regional conference in Wisconsin and grew with the increasing number of documents which address the problems and issues related to teaching and using critical thinking skills to evaluate Internet resources. Another comprehensive list is Kathy Schrock’s Guide for Educators (15). Schrock has designed a series of evaluation surveys, one each at the elementary, middle, and secondary school levels, to aid teachers. She also provides links to several Web pages on critical evaluation. Google Directory for Web Site Evaluation (16), created as a part of the Open Directory Project, is another source for Web pages dealing with the subject. The pages can be viewed as per the Google PageRank order or alphabetically.

 Two reports have given a comprehensive listing of criteria and there is a lot of commonality in the criteria laid down by them. One is the report, Selection criteria for quality controlled information gateways (17). This report is a study of the selection criteria employed for selective information gateways or subject gateways on the Internet. It concentrates on the subject gateways run by DESIRE (Workpackage 3) partners (EELS, KB, SOSIG) and the UK Electronic Libraries (eLib) Programme subject gateways (ADAM, EEVL, OMNI, RUDI, SOSIG). For generating a comprehensive list of selection criteria, a systematic review was conducted. The aim was to capture all the
selection criteria and quality attributes either currently being used by Internet services, or recently mentioned in the literature. Four main information sources were used in the review:

- Subject based selective services on the Internet
- Other selective services on the Internet
- Related literature
- User surveys from the subject based services

The list of quality selection criteria produced as a result serves as a reference guide for individual services for selecting the criteria that are appropriate for them.

The second comprehensive report, “Building Sustainable Collections of Free Third-Party Web Resources” (18) was prepared by CLIR. Its purpose is to identify and synthesize existing practices used in developing collections of free third-party Internet resources that support higher education and research. The report recommends those practices, policies and models that have proved to be particularly effective in terms of sustainability, scalability, cost-effectiveness and applicability to their stated purpose. The report describes how the nature and complexity of free Web resources comply with or challenge traditional library practices and services. Cataloguing, access and archiving, staff development, user training, and cost issues associated with collecting free Web resources are elaborated. It gives a synthesis and an amalgamation of criteria and practices for evaluating free Web sites described at sites developed by Internet Scout Report, Research Discovery Network, and several other sites from various countries. The online tutorial, Internet Detective and the DESIRE Handbook have also been referred to. This report has served as a guide for several institutions in defining their quality criteria.

The guide Evaluating the Content of Web Sites: Guidelines for Educators (19) is a reference tool for educators for preparing students to think critically on the use of the Web. The criteria laid out within this document come from a broad array of sources.

The most comprehensive attempt to develop Web evaluation criteria is the study undertaken by Gene L Wilkinson and colleagues (20). They used the consensus method
to arrive at a set of criteria for evaluating Web resources. They examined the criteria used by sources that review sites and compiled a lengthy list comprising 509 possible quality criteria. A final list of Web site developers’ evaluation criteria consisting of fifty-two elements was arrived at.

Rettig (21) draws analogies between reference book reviewing and Web site reviewing, delineating both similarities and differences. The author finds that importance is given by commercial services to the superficial aspects such as the “look” or the “feel” of the site, and the resulting reviews are not really evaluative. Though many of the established criteria can be used for reviewing electronic resources, the dynamism of the electronic medium, incorporation of related resources through hot links and design issues peculiar to the medium require development of criteria appropriate to Web resources. In another article, Rettig and LaGuardia (22) review the efforts undertaken by different academicians and reference librarians for identifying evaluative criteria for Web resources. After examining the existing lists of criteria, the authors propose the ‘Rettig/LaGuardia Canon for Reviewing Electronic Resources’ consisting of eight criteria.

The Web page created by Elizabeth Kirk (23) for helping the users of the Sheridan Libraries of The Johns Hopkins University discusses six criteria that can be used to assess information found on the Internet. The Web page gives a very strong message regarding importance of evaluating information:

“All information, whether in print or by byte, needs to be evaluated by readers for authority, appropriateness, and other personal criteria for value. If you find information that is "too good to be true", it probably is. Never use information that you cannot verify. Establishing and learning criteria to filter information you find on the Internet is a good beginning for becoming a critical consumer of information in all forms. "Cast a cold eye" on everything you read. Question it. Look for other sources that can authenticate or corroborate what you find. Learn to be skeptical and then learn to trust your instincts.”
Smith (24) surveys criteria that have been published on the Web and in the print literature. He amalgamates the criteria and proposes a set of six main criteria in the form of a toolbox that can be used by librarians and users to evaluate Internet information sources. The intention of a toolbox approach is to help librarians and others to choose those criteria appropriate for their needs.

Librarians' Internet Index (LII) (25) is a service that aims to provide access to reliable, trustworthy, librarian-selected Web sites. LII is guided by the five key criteria, and Web sites that do not meet these are not included in the LII database. These factors also provide critical guidance for weeding out the sites from the database.

Institute for Scientific Information (ISI), the pioneer in selection and indexing of literature published in peer-reviewed scholarly journals, books and proceedings is helping to establish a list of authoritative Web sites using specialized criteria (26). ISI advises that objective, current, referenced or professionally sponsored information is a mark of accuracy, while bias and density of advertising could be a mark of inaccuracy. As regards currency, it suggests that since the Web environment is an environment of change, a six-month interval is the maximum acceptable period between updates regardless of discipline.

Guidelines for Evaluating Web Sites prepared by ERIC (27) consists of a checklist that has been compiled from several sources and is a summary of criteria for evaluating Web sites.

Monash University has laid down some criteria that must be used by selectors for evaluating Web sites and free Internet resources to be included in its Voyager Catalogue (28). The recommendations are submitted on a form. Sites are included if they satisfy the criteria.
The Internet Scout Report (29) brought out by the Internet Scout Project is one of the Web's oldest and most respected current awareness services. Web sites are selected only when they meet the fundamental criteria laid down.

Alexander and Tate (30) have compiled different Checklists (Advocacy Web page; Informational Web page; News Web page; Personal Web page; Business-Marketing Web page) for evaluating different types of Web pages based on five common criteria. The book by Alexander and Tate (31) is based on these checklists. It provides tools and techniques to help the users to meet the evaluation challenges posed by the Web. The book also provides important guidelines to Web page authors so that the information they provide can be considered as reliable, accurate and trustworthy.

The CARS Checklist (Credibility Accuracy Reasonableness Support) (32) provides criteria that help to critically evaluate the material and to separate the high-quality information from the low. Credibility is satisfied when the source is trustworthy; the quality of evidence and argument is evident; the author's credentials are available; quality control is evident; it is a known or respected authority; it has organizational support. Accuracy means the source is up-to-date, factual, comprehensive. Absence of date, one-sided view means poor accuracy. Reasonableness is when the source is balanced, objective, there is no conflict of interest; there is an absence of fallacies or slanted tone. Support indicates that the source has references and contact information.

The Ohio State University Libraries (OSUL) has given the criteria for inclusion of Web sites in OSCAR online catalogue in very crisp statements (33). They act as guidelines for selection of Web sites.

A study that has brought out the disturbing fact that not just the students, but also the faculty needs to be trained about the importance of evaluating Web resources from the point of view of information content rather than mere appearance, is the work by Cottrell (34). Participants at a faculty workshop when asked to generate a list of criteria for evaluating Web sites produced a total of 46 criteria. 30 of the 46 items in the total list
concerned design or usability rather than content. Only 16 items reflected criteria such as identification of the source, credibility of the source, appropriateness or relevance, currency, inclusion of references and similar criteria related to information content. Some of the basic criteria for evaluating sources such as accuracy of information, objectivity of presentation, and coverage were not mentioned by any group.

Sweetland (35) compares various attempts at creating consensus lists of evaluation criteria for Web resources using reviews of Web sites published in Choice magazine, a well known reviewing medium and selection tool. The author observes that users and developers of Web sites are more concerned about criteria such as ease, variety of access and aesthetics than with traditional aspects of quality such as reliability, validity, and accuracy. He fears that such "old fashioned" criteria might cease to exist as a major element in selection. However, if Internet sites are selected by professionals (librarians) just like the other forms of material, these quality measures will be taken care of, so that the users need not be worried.

Toni Greer, et.al. in their white paper (36) give a new approach to categorizing evaluation criteria as External criteria and Internal Criteria. External criteria refer to the ‘who and where’ of information - who wrote the article, from where did it come, what are the academic qualifications of the author, etc. Examining the URL for identifying the domain type helps in knowing where the information came from. Generally, Web sites are found to be more credible if the length of the URL is short and the end of the URL contains .edu, .gov, and .org, specifically. Internal criteria mean using one’s own expertise, or independent knowledge, to determine if the information is accurate.

Stanford Web Credibility Research has compiled the Stanford Guidelines for Web Credibility (37). Ten guidelines have been prepared for building the credibility of a Web site, which can be translated into criteria for assessing its quality. The guidelines have been prepared based on three years of research that included over 4,500 people.
Matthew Ciolek (38) has given seven root principles which 'quality' Web sites adhere to:

1. Provide their OWN information
2. Are useful and informative
3. Are easy to find
4. Are universally accessible
5. Are well structured and organized
6. Are well formatted and edited
7. Are easy to establish, run, maintain and improve on

2.2.2 Application of Quality Criteria for Evaluation

Though the survey of literature pointed out a wide range of studies undertaken to identify the quality criteria that should be applied in evaluation of Web resources, the researcher found very few elaborate studies where Web sites or Web pages were actually evaluated by applying the criteria.

In order to bring out the most important factors that Web designers should take into account when developing a site for selling their products, Oppenheim and Ward (39) carried out research to evaluate the effectiveness of a number of Web sites selling chocolate. A series of criteria were developed and a scoring system associated with them was generated. A panel of eight volunteers evaluated ten chocolate-related Web sites. Aspects evaluated included presentation, content, accessibility, language, navigation and structure, transaction page, security, privacy and authority, and marketing factors. The resulting scores were then analysed and the best Web site was identified. It was observed that there was significant inter-scorer discrepancy in the scores assigned. The authors comment that there is a need to involve real users when developing criteria for the evaluation of web sites.

Fan, et.al. (40) evaluated five Web sites dedicated to Chinese archaeology, using credibility, accuracy, reasonableness, support, and uniqueness as criteria in an effort to
determine the value and quality of information. They point out six main characteristics of online resources for the subject area.

Bouchier and Bath (41) carried out a study under which 15 Web sites providing information on Alzheimer’s disease were evaluated. Four Web site evaluation tools were used, and the Web sites were ranked according to the scores they achieved with each tool. The correlation coefficient of the rankings of the Web sites for each pair of tools was determined. Several Web sites received a high rank in all the evaluation tools. In further continuation, Bouchier and Bath (42) developed a tool specifically designed to evaluate Web sites that provide information about Alzheimer’s disease. The specific tool was used to evaluate the same 15 web sites and the results were compared with those obtained using the four generic tools used in the earlier study. There was a poor correlation between the rankings obtained using the specific tool compared with the generic tools. Though some Web sites attained high scores using the specific tool, most of the Web sites scored lower using the specific tool.

Sasikala (43) examined forty randomly selected Web sites of different organizations from India, USA and UK in order to test their authenticity and usability. A set of questions framed by the Pace University Library (http://www.pace.edu/library/instruct/webevalworksheet.html) was used as a checklist to evaluate the content, structure and authoritativeness. All the Web sites gave an indication of author/producer, most gave contact details, 67% gave affiliation of author. Information regarding the date of posting the information and latest date of revision were clearly indicated by about half of the Web sites. 85% of the Web sites were informative in nature. Only 30% of the sites were found to be having alternative sources, hence majority could not be evaluated in terms of their accuracy.

A few studies related to evaluation of Web sites belonging to a particular subject area have been undertaken for M.L.I.Sc. dissertation projects at the University of Pune. Gambhir (44) conducted a study on evaluation of selected Web sites for children. Patankar (45) evaluated the Home page of selected special libraries of Pune. Gaikwad (46)
carried out an evaluation of selected Web sites in the field of Library Science. Tarahomi (47) carried out the evaluation of 45 Web sites related to ‘Philosophy’. Hande (48) studied and evaluated 27 Web sites in Law. Kottalgi (49) evaluated 45 Web sites related to ‘Electronics’. In these studies, Web sites were evaluated on the basis of broad criteria such as Scope, Content, Graphics & Multimedia design, Purpose & audience, Reviews, Workability and Cost. Researchers have made comments on whether or not the criteria are satisfied. No measures have been used to rate the Web sites. Some common observations are that scope, purpose, and audience are not clearly stated in most Web sites. Site map and FAQ are rarely provided. Contents of Web sites are logically organized.

2.3 WEB IMPACT FACTOR

Apart from the use of quality criteria similar to those used to evaluate traditional print materials for Web resources, another method of evaluation can be used, taking advantage of the link-based structure of the Web. The methodology belongs to the field of Webometrics, which is the bibliometric study of Web resources, a concept that came about in the mid-1990s. It makes use of the Web Impact Factor (WIF), a quantitative tool developed on the lines of the popularly used Journal Impact Factor. WIF is a link-based Web site impact measure, and can be used for ranking, evaluating, categorizing and comparing Web sites. The concept was first discussed by Rodriguez I Gairin in 1997 (50) in a Spanish language journal. However, it gained wide attention only after a paper by Ingwersen (51) was published in English. On lines similar to the methodology adopted by Ingwersen, namely the use of large-scale search engines for finding out the number of links (to calculate the WIF), several studies were undertaken for studying Web impact of Web resources having varied scope and characteristics.

Ingwersen undertook a case study to investigate into the feasibility and reliability of calculating impact factors for Web sites, called Web Impact Factors. The study analyzes a sample of seven small and medium scale national Web domains including Japan and six European countries, four large Web domains (.gov, .org, .com and .edu sectors mainly
from the USA) and six institutional Web sites over a series of snapshots taken of the Web during a month.

The advanced search mode of the Web search engine Alta Vista, which allows the use of elaborate Boolean search strings and provides the exact number of pages found, was applied. Impact factors associated with internal and self-link Web pages and external-link Web pages were generated. Norway showed the highest Web impact factor out of the European countries. Japan showed the lowest WIF. The results of the study demonstrate that analyses of Web impact factors for national, sectoral and larger Web segments or sites are reliable. They might be less dependable for smaller institutional sites since variation of the Web impact factor over different snapshots might be much more significant in their case.

Smith (52) studied the viability and reliability of Alta Vista search engine for webometric research, and found it very well suited. He surveyed the WIF of some South-East Asian Countries sites, Australasian university Web sites, Australasian electronic journals and National Library sites for Australia and New Zealand. High WIF was received by the National Library of Australia, possibly due to the large number of information resources provided and the larger constituency it serves. The results surprisingly showed a very high external WIF of Papua New Guinea and Indonesia (relatively insignificant Web information resources), and very low external and overall WIF for Australia (largest number of Web pages). WIF for Australasian university Web sites were compared with the number of publications by them as per the ISI citation databases. Most of the times, the domains having larger number of pages also had more number of publications.

In another study, Smith (53) compared the WIFs for Australasian Universities and Australasian electronic journals. He observed that for large organizations such as research institutions or universities, WIF seems to be a useful measure of the overall influence. However for smaller Web spaces such as electronic journals, the WIF is less reliable as a measure. Smith (54) also examined the use of citation counts and Web links to evaluate online LIS journals. The study compared the WIF of a sample of open access online LIS
journal Web sites with their conventional citation counts. There was a poor correlation between the rankings obtained using the specific tool compared with the generic tools. The ISI databases were searched for citations to these e-journals, and those that had citations were searched on Alta Vista to determine the overall number of links, and Web Impact Factor. The counts from ISI citation databases and from the AltaVista searches were compared.

In addition, the Google PageRank was also included for comparison. There appeared to be only a slight relationship between the number of ISI citations and the number of links to the online journal. Several journals with low numbers of ISI citations had high number of links, and vice versa.

Noruzi (54) has discussed the advantages and disadvantages of WIF, data collection problems, and validity and reliability of WIF results using large scale search engines such as Yahoo and AltaVista. The paper observes that though these search engines offer link search strategies to make webometric measurements (e.g. total number of pages in a web site and the total number of back-links to the web site), the strategies are unorganized. He also observes that the results obtained by applying the inverted but logically identical set operations differ slightly, as earlier observed by Ingwersen (1998 ibid). So also, different results are obtained at different times of searching. The article concludes that while the WIF is useful for quantitative intra-country comparison, application beyond this (i.e., to inter-country assessment) has little value. A wide range of factors can influence the WIF. Ease of access to Web sites, publication immediacy, site language, site updating and the type of material published on the site are all contributors to WIF.

Noruzi (55) attempted to establish a kind of academic ranking of Iranian universities by evaluating their WIF. WIF were calculated for Web sites of 205 Iranian universities using AltaVista. These WIF were then compared to study the impact, visibility, and influence of Iranian university web sites.
In another study, Noruzi (56) investigated the Web presence and Web Impact Factor (WIF) for country code top-level domains (ccTLDs) of Middle-Eastern countries, and sub-level domains (SLDs) related to education and academic institutions in these countries. Counts of links to the Web sites of Middle-Eastern countries were calculated from the output of Yahoo search engine. WIF were computed at two levels: top-level domains, and sub-level domains. The results show that the Middle-Eastern countries, apart from Turkey, Israel and Iran, have a low Web presence, and their web sites have a low in-link WIF.

The paper in Chinese by An and Qiu (57) compares the Impact Factors of Chinese engineering journals, published by Institute of Scientific and Technical Information of China (ISTIC) for the year 2000, with the number of external links to Web sites of these journals and Web Impact Factors (WIFs). The Impact Factors are found to correlate significantly with external Web link counts and WIFs, therefore the authors conclude that both, external Web link counts and WIF can be considered as important indexes of Web site evaluation.

2.4 PAGERANK

Yet another method that can be used to get an idea of the quality of a Web site or the impact created by it is making use of the PageRank facility of Google search engine.

Brin and Page (58), in the path breaking paper presented at the Seventh International Web Conference, WWW98 write about Google, a prototype of a large-scale search engine developed by them, which makes use of the structure present in hypertext. They introduce how Google search engine makes use of the link structure of the Web to decide quality ranking for each Web page. This ranking is known as ‘PageRank’.

In another paper, Page, et.al. (59) propose PageRank as a method for computing a global ranking for every Web page based on their location in the Web’s graph structure, regardless of their content. A mathematical description of PageRank is given. The authors
explain that merely counting the back-links (links coming to a Web page) on the lines of citation counts have a number of problems on the Web, mainly due to the characteristics of the Web which are not present in normal academic citation databases. An intuitive description of PageRank is given as - a page has high rank if the sum of the ranks of its back-links is high; this covers both the cases when a page has many back-links and when a page has a few highly ranked back-links. The authors also remark that the library community is concerned with what quality means on the net, and PageRank can offer some insights into this.

2.5 ISSUES ARISING FROM LITERATURE REVIEW

The following issues have been identified from the study of literature:

- Not much literature was found dealing with the constance and permanence of Web pages and Web sites.

- Koehler’s study of constance or change behaviour of Web pages did not document the ‘importance’ or semantics of change. It only took cognizance that some kind of change occurred, since it was measured using automated tools and was linked to change in byte size or weight. Any change in size was considered to indicate a change in contents. However, change in byte weight or size does not necessarily indicate that the information content or subject content that matters has changed. For example, an increase in byte weight could be the result of addition of a graphic that has no other effect than adding to the visual appeal of the Web page. On the other hand, significant change in subject content could have occurred without much affecting the byte weight. To understand how updated and well maintained a Web site is, it is necessary to study how often the actual information content of its Web pages is updated.

- Removal of Web pages from Web sites or removal of entire Web sites results in dead links or unstable URLs, which affects the validity of citations to Web resources. Some possible solutions such as URNs (Uniform Resource Names), URIs (Uniform
Resource Identifiers) and PURLs (Persistent URLs) are being considered as alternative solutions to unstable and transitory URLs.

- Libraries are exceedingly depending on Web resources they neither own nor manage (external digital assets), and are including in their catalogues and gateways, open-access Web resources that are not covered by licenses or other formal arrangements. These open access resources are managed with varying degrees of control - some might be available on an individual’s informally managed Web page, others on an organization’s highly controlled Web site. Risk assessment plays an important part in selection and long-term preservation of Web resources. Risk management techniques are being explored in order to implement appropriate and effective measures to prevent or mitigate loss of resources crucial to the organization/users. Several indicators can be used to identify and quantify risks to Web resources.

- Survey of literature related to development of quality criteria for evaluation of Web resources has revealed that none of the studies except one (that by Wilkinson and others) has differentiated between criteria for Web pages and Web sites. Though there is bound to be some overlap of criteria between the two, it is equally necessary to have two distinct sets of criteria – one specifically for Web pages and the other for Web sites.

- The survey also shows that there is a lot of duplication and lack of standardization with regard to Web evaluation criteria and the terminology used. There is a need to arrive at a standardized list of quality criteria after removing duplicates, adopting standard terms to represent the criteria and logically grouping the criteria under broad categories.

- It is necessary to develop a set of quality criteria that concentrate on evaluating the information content of Web resources keeping in mind the needs of the scientists/technologists. Their requirements and the purpose for which information is to be used is different from that of individuals who wish to use it for personal use.
such as entertainment, or students who wish to use it for academic projects, etc. The subject information delivered by the Web resources would be used by scientists for serious research endeavours, and cannot be taken at its face value merely because it comes from a site that is attractively designed or has aesthetic appeal. In depth and critical evaluation of the objectivity, authority and authenticity of information is of utmost importance.

- The significance of Web sites can be established with the help of quantitative measures such as Web Impact Factor and PageRank, that are based on the hypertext link structure of the Web.

The next chapter deals with the access and preservation issues related to Web resources. It discusses the concept of digital preservation, preservation strategies that can be adopted, and challenges faced in preservation of Web resources. It traces the development of Web preservation projects and digital repositories. The chapter elaborates upon the establishment of subject gateways, a new system for enabling users to access Web-based resources regardless of their physical location, and traces landmark developments in the area.
REFERENCES


http://www.dlib.org/dlib/january02/kenney/01kenney.html (Accessed on 06 Jan 2005)


25. Librarians’ Internet Index. LII selection criteria.


28. Selecting free Internet resources: Criteria for cataloguing free Web sites, e-journal, e-books and other Internet resources.

29. The Internet Scout Report: Selection Criteria


