Online Research Methodology: Using the Internet and the Web for Research and Publication

Tarun Tapas Mukherjee
Bhatter College, Dantan, Paschim Medinipur, West Bengal

Abstract
The advent and introduction of digital technology has created a challenging situation for all and slowly but steadily we are experiencing its immense impact on education. In the wake of this digital revolution, we witness the rise of a converged platform in the form of World Wide Web, which has become favorite destination for information seekers. With the platform theoretically available anywhere anytime, the task of a researcher has become on the one hand easier and on the other very complicated. This paper is an attempt at understanding the context and tries to formulate certain methodology for making effective use of a new medium for scholarly research and publication.

[Keywords: Web, Internet, online methodology, access, search tools, reference management, e-journal]

“The human mind…operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature.”

Vannevar Bush

Introduction
The web is one of the services running on the internet using the hypertext and certain other protocols for electronic access and communication. It has been possible because the web exploits reproductive and regenerative capacities of computer. After 20 years of its invention by Tim Berners-Lee, we see how the web has become a favourite destination for researchers. However, the paradigm of a networked world of knowledge independent of any physical location "that would supplement, add functionality, and even replace traditional libraries was not new; the idea was used first by H.G. Wells” in his vision of the “world brain” (Harter 11). Then in 1945 Vannevar Bush published “As May Think” in the Atlantic Monthly, in which Bush conceives of a ‘memex’ machine, which inspired much of the early application of computers to information retrieval. The vision of a fully computer-based library began to emerge in terminology in the mid-60s with J. C. R. Licklider’s “The Library of the Future, a book that that contributed much to the construction of the internet” and the web later on. The digital library became possible in 1971 with Michael Hart’s Project Gutenberg; but still it remained confined to a comparatively very small location. The construction of a worldwide network of information and communication became possible only with the invention of the web.
Now in the second decade of this century we have already seen how ICT is revolutionizing the field of research and scholarly communication. With the arrival of web 2.0 or the semantic web and various networked reading devices like the Kindle, iPad, e-book readers and even mobile phones, we are becoming more and more dependent on a networked system like the web. The success of the web as a scholarly platform and tool can be understood in Lesk’s ambitious claim in 1997 that "we will have the equivalent of a major research library on each desk. And it will have searching capabilities beyond those Bush imagined" (Lesk 270).

With the advent of the portable hand-held devices like the tablets and smartphones powered by operating systems of considerable power—greater than the computer which was used on Moon Mission, people are moving towards mobility and portability for information retrieval and networking via apps through the internet protocols. This shift towards apps—which are created on the principle of object-orientation has created such situation in which Chris Anderson, the Editor-in-Chief of Wired Magazine recently made a hypothetical announcement: “The Web is Dead. Long Live the Internet”. Whatever the situation we will have in future, at present we don’t have the killers apps which could replace the web and for scholarly purposes the web will continue to be used.

Approaching the new medium

The online medium is fundamentally different from the print medium and a researcher faces certain problems in using it because of the virtual and volatile nature of the contents. In view of those problems this paper proposes certain research methods for using the web for research. For structural convenience this paper approaches online research methodology from three key areas:

- Exploring the resources on the web
- Accessing the resources on the web
Organising the web resources
Publishing on the web

Exploring the Web

The web has been conceived of and created as an interconnected network of resources and that is why we get the protocol ‘URL’ as Uniform Resource Locator. While explaining URL Tim Berners-Lee and his team defined a web resource as

“…anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., “today’s weather report for Los Angeles”), and a collection of other resources.” (Lee)

So theoretically we can make use of any resource available on the web for our research having an ‘identity’. Though web resources are found in unorganized forms, for our convenience we can divide them into certain types:

Digital libraries: Digital libraries are being created as a full-fledged alternative to the traditional physical library system for accessing a variety of materials (original texts, creative works, movies, paintings, music albums etc) in various formats. Notable examples are: Project Gutenberg (www.gutenberg.org), The Perseus Project (www.perseus.tufts.edu), ILEJ: Internet Library of Early Journals (www.bodley.ox.ac.uk/ilej) etc.

Online Archives: Just like digital libraries, archives are also being created online as an alternative to traditional archives. However, an online archive may function just like a digital library and the difference may be just in name. Famous archives on the web are: Internet Archive (www.archive.org), The Oxford Text Archive (http://ota.ahds.ac.uk), Poetry Archives (http://www.emule.com/poetry).

Full text databases: A full-text database is a compilation of documents or other information in the form of a database in which the complete text of each referenced document is available for online viewing, printing, or downloading. In addition to text documents, images are often included, such as graphs, maps, photos, and diagrams JSTOR (www.jstor.org), ARTstor (www.artstor.org), Project muse (http://muse.jhu.edu), EBSCOHost (www.ebscohost.com)

Independent scholarly sites: The publishing technology of the web has facilitated the rise of many e-zines and e-journals. Many of them has transferred from the print to the web edition. Even it has become a practice for many established newspapers, magazines and journals to being out web editions.

Format specific repositories: Because of the worldwide demand for certain types of resources, many format specific large sites have come up with special services; for example, Flickr and Picassa (photo sharing sites), Youtube (video sharing sites).

Social networking sites: In the early days of social networking certain sites like Myspace, Orkut, Facebook were avoided by scholars because of unscholarly nature of the contents generated there. But now some of the materials can be used for scholarly purposes; for instance, the post of a famous writer or a communicated message.

Personal sites: Many authors and critics now maintain personal sites or blogs for communication with readers or for advertising. A researcher can make use of those resources.

General websites: Depending upon the kind of research information available on general websites can be used for the purpose of research.
Wiki Sites: There are many sites like Wikipedia which run on the Wiki software for collaborative publishing. Researchers may consult those sites but should avoid citing them as source of research because the wikis are frequently updated by writers of dubious identity and intentions without proper control of an editorial authority. Of course it records the history of edits, but still they cannot be used for citations because of the lack authority.

The resources available on those location can be used both for primary and secondary sources depending upon the kind of research. But the problem is that a particular resource may not be scholarly at all or may have dubious existence. More importantly, the web resources may not have stable existence and may change or may disappear altogether. MLA Handbook specifically says,

“Whereas the print publications…are generally issued by reputable publishers, like university presses, that accept accountability for the quality and reliability of the works they distribute, relatively few electronic organizations currently have comparable authority.” (p. 34).

For this, the MLA Handbook emphasizes the “need to assess the quality of any work scrupulously before using and citing it” (p. 33) and recommends carefully checking out three aspects of a source: authority, accuracy and currency while evaluating resources (p. 34). By authority the MLA suggests considering the credentials of the persons responsible generating the contents and the authenticity of the content: ‘author’, ‘text’, ‘editorial policy’, “publisher or sponsoring organization” (p. 36). As for accuracy of the resources on the web, what is applicable for a print resource, is equally applicable for an online resource. According MLA Handbook accuracy is important for verifiability of the information and sources. For evaluating a web resource it is necessary to check out currency of a resource—“how current the author’s scholarship is”. (p. 37). In other words, a researcher must look for the dates of publication of the sources cited.

Accessing Web Resources

The concept of access to information has evolved, as Borgman (2000, 79) shows, from the varied areas such as library system, telecommunication system and so on. According to her (Borgman 2000, 57), access to information is a process, through which the user is able to retrieve the information s/he seeks from the internetwork of computers provided that,

i. The user has the basic technical knowledge and skills,

ii. The technology is viable,

iii. The information is relevant and usable.

The whole process of access to digital libraries is dependent on these three factors: the knowledgeable user, technology and nature and quality of data. In other words, the user should have a minimum level of technical knowledge for better access in terms of quality of the retrieved data. This paper is not about the technical knowledge a scholar should have, but the most basic things are explained.

The Search Models

Generally speaking, we come to across the following information models with the digital libraries on the internet:

- Boolean Model
The Vector Space Model
The Probabilistic Model
The Natural Language Processing Model
The Hypertext Model

The first three models function by matching search terms with index terms to generate search results. “One of the major criticisms of them is”, (Gobinda Chowdhury and Sudatta Chowdhury, 2003), “that they look at individual search terms; they do not consider the search or index terms as part of a sentence or document.” That is why the last two models are put forward to tackle the limitation of the previous models.

**Boolean Search Model**

This search model is the oldest and functions in accordance with set theory and Boolean algebra. It operates by matching a set of search terms against a set of index terms. Multiple search terms are processed on the basis of logical product (AND logic), logical sum (OR logic) and logical difference (NOT logic). The processes of its functioning are described later in this chapter.

**The Vector Space Model**

This model is based on the calculation of binary weights. It functions by assigning non-binary weights to index terms in queries as well as in documents and computing the degree of similarity between each document in a collection and the query based on the weight of the terms. Thus a ranked list of output can be produced with items that fully as well as partially match the query. While this model produces a ranked list, the major weakness of this model lies in its assumption that index terms are mutually independent.

**Probabilistic Model**

Probabilistic models are based on the principles of probability theory and they treat the process of document retrieval as a probabilistic inference. Similarities and associations are computed as probabilities in order to determine whether a document is relevant for a given query.

**The Natural Language Processing Model**

This model (also known as computational linguistics) is an attempt at processing search items not simply in terms of keywords, but also in terms sentences, taking into consideration syntactic, semantic and pragmatic analyses. Webopedia defines it as “a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages" (Webopedia). In other words, it tries to make computer understand how human beings learn and use language.

**The Hypertext Model**

This model evolved as a system to overcome the limitations of the fixity and linearity of the conventional documents. It does so by putting in hyperlinks to other parts of a document (sentence, paragraph or the entire document on a local machine and to other domains and sub-domains on the web. The hyperlinks are made indexable and search able by search programmes. For the flexibility of this model it has played a major role in the designs of the websites and in the functioning of the internet. It should be noted that hypertext model has been largely instrumental in the making of Hypertext Markup Language (HTML) and Hypertext Transfer Protocol (http).
How to Search Effectively

It has been generally found that teachers and students search the web for resources just by using the major search engines and through certain keyword or phrases, which lead them to particular digital libraries and web resources. Since they are not familiar with the search techniques, they cannot get optimum access to the resources. Added to this is their deep-seated phobia of viruses and distrust of unknown sites. While the virus threat can be effectively minimised by using a good anti-virus software, better access can be achieved by being familiar with the ways the digital libraries and the web function.

Boolean Search

Boolean search employs special logic to produce search results. Without knowing its basic functions, a user cannot apply the logic to retrieve information in the digital environment. The search operators may vary with different libraries, but the basic function is very intuitive and simple. For instance, if a user applies the logical product (AND logic) and enters the search terms “Shakespeare and fool”, it will retrieve all those documents where both the terms appear. The second ‘OR logic’ “allows the user to combine two or more search terms in order to retrieve all those items that contains either one or all of the constituent terms” (p. 188) Following this the search terms “Shakespeare or Marlowe” will retrieve all those documents i) where the term ‘Shakespeare’ occurs, ii) where the term ‘Marlowe’ occurs and iii) where both the terms occur. By using this logic, search broadens its scope. On the other hand, ‘NOT logic’ is used to restrict the search results to specific terms and exclude particular term. For instance, “Elizabethan dramatist not Marlowe” will retrieve all the records except Marlowe.

Truncation

Truncation sends signals to a search engines to retrieve the information relating to the different terms having the same common root. The user can perform this kind of search by placing operator like ‘*’ or ‘?’ (which may vary with different search engines) in the left hand side of a root, in the right hand side of a root or in the middle of a world. For instance, “*logy” will result in retrieve terms having ‘logy’ at the end like ‘philology’, ‘psychology’, ‘biology’ etc. Right-hand truncation like “philo*” will produce search results having the same characters in the beginning like ‘philosophy’, ‘philology’, ‘philomel’ etc. Similarly middle-truncation (humo*r) retrieves the terms matching characters (like ‘humour’ ‘humor’).

Proximity Search

This type of search is performed in order to specify the distance between two terms in the retrieved results. In principle, this is similar to the Boolean ‘AND’ search, but the difference is that it makes the search more restricted and more user’s query-oriented. The use of operators for this varies with different digital libraries. In the ACM digital library (http://portal.acm.org/dl.cfm) the ‘NEAR’ is used to retrieve terms which will have close proximity to each other.

Field or Meta Tag Search

This search is performed when a user wants to restrict searches to more specific results. This is done by selecting an appropriate given field (area) before proceeding to search a particular item in the collection. This is called field or meta tag search because the fields in digital collections are specified by meta tags. For instance, in the “Advanced Search” wizard of the Project Gutenberg library, the user can restrict search results by selecting appropriate fields from ‘Language’, ‘Category’, ‘LoCC’ and ‘File Type’, where the items are expected to be found. In the Batleby library the user is given the option of choosing a particular field in “Select Search” option before performing a particular search.
Limiting Searches

A digital collection in a particular library may contain many items with similar index terms. In this a particular simple search may result in hundreds of retrieved items. In such cases, it is necessary to limit searches by choosing appropriate criteria such as language, year of publication, type of information, file type etc. This type of action is also useful in searching the entire web.

Organising web resources

A researcher has to acknowledge his/her “debts to predecessors by carefully documenting each source, so that earlier contributions receive appropriate credit and readers can evaluate the basis for claims and conclusions.” (MLA 126). This is a daunting task for any researcher to keep a record of the sources. In case of online sources the difficulty multiplies because of the unstable nature of some of the web resources and their location, and difficult nature of location names.

Keeping Records and Reference Management

Reference management has become very important now-a-days. A university advises its research scholar in its website this:

“A critical part of the student and faculty research process is keeping track of relevant literature—journal papers, books, web pages, images, quotations, etc.—so that they can be utilized and properly cited in the writing process of research.”

While many of the print journals are migrating to the online format even if they keep up the print issues, For this it directs recommends “reference management programs” which “can assist by:

- Collecting references 1) from online sources…databases, web pages, and other sources; or 2) by manual input.
- Storing and managing these references in searchable folders.
- Capturing related PDFs, web pages, files, or images; or linking to available fulltext.
- Adding personal notes and indexing PDF fulltext.
- Generating standalone bibliographies or inserting references into papers composed in Microsoft Word, OpenOffice and other word processors and automatically formatting them in a required publication style, e.g., MLA, APA , or CSE.
- Creating user groups and sharing references for class and other collaborative research work.”

It is vital to keep record of the online sources offline in a local computer in a convenient organized way. This can be done by making separate folders for a specific type of resources. For instance one can make a folder for the resources and under this folder create another folder for web resources and then make separate folders for separate materials. Unlike the PDF documents HTML documents are not directly downloadable. One has to download the whole page for offline use. For this it is necessary to make a separate folder for a single HTML document. For convenience of research, one can modify a file name and add metadata; such as, short name of the article, date of access, site name etc. All these sources can be locally tracked from a single document file which may be in the form of a locally hyperlinked
Online Research Methodology: Using the Internet and the Web for Research and Publication

bibliography so that one can easily keep track of the sources and verify it and modify the bibliography with additions.

For reference management it is better to seek the help of those word processors which have extra plug-ins for generating bibliography and end notes; for instance, Microsoft Word, OpenOffice etc. There are many paid and free reference management softwares and services available on the web for online or/and offline use; among the paid softwares Reference Manager® by Thompson Reuters (http://www.refman.com/), EndnoteX4® by Thompson Reuters (http://www.endnote.com) and Mendeley® (http://www.mendeley.com) by Mendeley can be mentioned. Among the open source and GPL licensed softwares Pybliographer (http://www.pybliographer.org/) by Pybliographer Developers, Aigaion (http://www.aigaion.nl) Aigaion developers can be mentioned. Citeulike (www.citeulike.org) allows users to make a personal library of the online materials.

Publishing on the web

Publication of research findings is a crucial part of any research work itself. Trends and surveys show that researchers and publishers are steadily moving towards online formats for a number of reasons like ease of access, worldwide visibility, ease of payment, low cost, currency of publications etc: “Most of the literature describing the recent growth in electronic journals emphasise three important factors; money, technology and convenience, and speed.” (Umeshareddy Kacherki*and Mahesh J. Thombare, 24). The PDF format duplicates the print format and most scholarly journals stick to it. However apart from PDF and HTML other formats are being used and created for hand-held devices, and online journals and magazines are making effective use of the format. While the traditional, HTML regular site designing is well enough for hosting a journal following the principles followed in print journals, certain softwares have been created for exclusively hosting online journals in order to bring in certain new functionalities like collaboration, inter-operability, real-time status-checking—which can be utilized only in online format. Mention may be made of Open Journal System which has been “made freely available to journals worldwide for the purpose of making open access publishing a viable option for more journals, as open access can increase a journal’s readership as well as its contribution to the public good on a global scale” (Public Knowledge Project). It offers certain features, some of which are unique and not possible for a print journal:

“OJS Features

1. OJS is installed locally and locally controlled.
2. Editors configure requirements, sections, review process, etc.
3. Online submission and management of all content.
4. Subscription module with delayed open access options.
5. Comprehensive indexing of content part of global system.
6. Reading Tools for content, based on field and editors’ choice.
7. Email notification and commenting ability for readers.
8. Complete context-sensitive online Help support.” (Public Knowledge Project)

It biggest strength is that it “assists with every stage of the refereed publishing process, from submissions through to online publication and indexing.” (Public Knowledge Project)

Apart from OJS other standalone softwares are also available, some of which are Open Source and some are proprietary ones. Whatever the software and platform however and whether an author has to pay for publication or not, it is important for researchers to select journals to publish with very carefully. It is not just sufficient to publish on journals having only
ISSN\textsuperscript{xiii}. There are certain other criteria—which good journals must fulfill. Scholars must consider

- The value of a particular publishing company/organization/institution;
- Whether it is indexed and abstracted in notable directories and databases like MLA, Elsevier, DOAJ, EBSCO, Thompson Reuters etc;
- Whether it is generating citations and has got considerable Impact Factor\textsuperscript{xiv}.
- Whether it is a peer-reviewed journal with an editorial board consisting of renowned scholars;
- Whether it provides detailed review report/s;
- Whether it is hosted on standard website platforms like regular HTML site, OJS and other journals systems (blogging platforms are considered not so scholarly);
- Whether it has established itself as scholarly space.

These precautions have become necessary as many dubious journals have appeared online, which do not follow standard procedures and their only aim is to make money out of the loopholes of certain academic norms.

**Not to conclude**

The web and other related technology may be said to be still in infancy, and nobody can anticipate their future. But in view of open and closed access to all forms of information on a single converged medium a different type of system would be necessary for handling digital objects in post-Gutenberg period. Certain proprietary specialized services are already available in various forms on subscription basis. But given the open and generative nature of the web we can hope for specialized open services through open source systems:

While we cannot be sure exactly where the Internet will lead, we are confident that its influence on our personal and professional lives will only increase in the next decade. Researches need to be actively engaging with the issues it raises.” (Chris Mann and Fiona Stewart, 218.)

**Notes**


\textsuperscript{ii} The ‘memex’ was conceived of as a microfilm-based machine which would include links between pieces of information in a research library combined with personal notes and notes of colleagues, anticipating the ideas of both hypertext and personal information retrieval systems.

\textsuperscript{iii} A fundamental pioneer in the call for a global network, J. C. R. Licklider, articulated the ideas in his January 1960 paper, “Man-Computer Symbiosis” in Transactions on Human Factors in Electronic: “A network of such [computers], connected to one another by wide-band communication lines [which provided] the functions of present-day libraries together with anticipated advances in information storage.

\textsuperscript{iv} Internet as ARPANET became possible on 29, October 1969.

\textsuperscript{v} Chris Anderson explains this in the following way: “You wake up and check your email on your bedside iPad — that’s one app. During breakfast you browse Facebook, Twitter, and \textit{The New York Times} — three more apps. On the way to the office, you listen to a podcast on your smartphone. Another app. At work, you scroll through RSS feeds in a reader and have Skype and IM conversations. More apps. At the end of the day, you come home, make dinner while listening to Pandora, play some games on Xbox Live, and watch a movie on Netflix’s streaming service. You’ve spent the day on the Internet — but not on the Web. And you are not alone.” The editorial is available at http://www.wired.com/magazine/2010/08/ff_webrip.


\textsuperscript{vii} For more please visit the website of the Humboldt State University Library at http://library.humboldt.edu/~rls/references.html.

\textsuperscript{viii} Ibid.

\textsuperscript{ix} Reference Manager is available for $239 though sometimes discounts are given.

\textsuperscript{x} EndnoteX4 is available for $229.

\textsuperscript{xi} Mendeley is available for $79.

\textsuperscript{xii} The Public Knowledge Project was founded by John Willinsky in the Faculty of Education at the University of British Columbia in 1998 and operates through a partnership among Simon Fraser University, the School of Education at Stanford University, the University of British Columbia, the University of Pittsburgh, the Ontario Council of University Libraries and the California Digital Library. It aims at “improving the scholarly and public quality of research”.

\textsuperscript{xiii} ISSN does not certify the quality of contents and the standard of publication. ISSN International Centre clearly states that the “ISSN…is an eight-digit number which identifies periodical publications as such, including electronic serials. The ISSN is a numeric code which is used as an identifier: it has no signification in itself and does not contain in itself any information referring to the origin or contents of the publication.” (ISSN International Centre)

\textsuperscript{xiv} Thomson Reuters defines Impact Factor as “a measure of the frequency with which the “average article” in a journal has been cited in a particular year or period. The annual \textit{JCR} impact factor is a ratio between citations and recent citable items published. Thus, the impact factor of a journal is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years”. For more visit http://thomsonreuters.com/products_services/science/free/essays/impact_factor

\textbf{Works Cited}


---

*Tarun Tapas Mukherjee is Assistant Professor in English, Department of English, Bhatter College, Dantan, Paschim Medinipur, West Bengal, India. Email: ttm1974@gmail.com*