CHAPTER III

PRESERVATION AND ACCESS TO WEB RESOURCES

3.1 Access Versus Ownership
3.2 Digital Preservation
3.3 Access to Web Resources
CHAPTER III

PRESERVATION AND ACCESS TO WEB RESOURCES

This chapter introduces how the networked environment has brought about a change in the concept of the ‘collection’ of the modern library. It discusses the two options for building up a collection of digital resources, namely access and ownership or preservation. By late twentieth century, the Web had been flooded with a plethora of information resources on each and every topic. Serious thought began the world over regarding how to manage the growing Web-based resources. Two lines of action emerged in the decade from 1991-2000 – one was the initiation of Web preservation projects and the birth of digital institutional repositories in order to preserve the contents of this new, ever-changing medium for historical and scholarly research use. The second activity was the setting up of subject gateways to provide access to remote resources using links embedded in quality-controlled databases of resource descriptions. The chapter throws some light on digital preservation and subject gateways and traces the landmark developments in these activities the world over.

3.1 ACCESS VERSUS OWNERSHIP

The libraries of ancient times, such as the Alexandrian library promoted the idea that the primary purpose of a library should be the storage of knowledge. A large, renowned library was an indicator of a wealthy, educated and powerful society. The motto, “Bigger the better” has been the foundation of libraries throughout the centuries and into modern times. Ownership of comprehensive collections was considered necessary for establishing the identity of a library.

With the information explosion in the 20th century, it became impossible for libraries to purchase everything that was needed to satisfy their patrons. There was a shift in importance from books to periodicals as sources of information. However, prices of periodicals, especially the scholarly and research periodicals skyrocketed. In spite of
spending major parts of their budgets towards periodical subscriptions. Libraries could still not completely cover even the core periodicals. To offset the problems created by information explosion coupled with insufficient budgets and increasing costs, libraries were forced to pursue other methods of fulfilling the information needs of patrons. The principle of ownership began to undergo a change. Instead of subscribing to high cost, low use periodicals, libraries started preferring acquisition of specific articles on demand through inter-library loan or by document delivery from document supply centres. These were the earliest modes of providing ‘access’ without ‘ownership’, and entailed ‘access to material not purchased by libraries’.

In the late 20th century, online journals and databases proliferated due to advances in computer technology, communication technology and electronic publishing. The several advantages of electronic periodicals, such as speed of publication, better presentation, hypertext links to related research or multimedia sources, saving of library shelf space and cost of maintenance, simultaneous viewing by several users, etc. prompted libraries to start opting for them. Today, libraries have several options to choose from: subscription to a ‘bundle’ of journal titles, consortia subscription, subscription to an aggregator database, access to the electronic version for free or at very little extra cost along with print subscription, etc. ‘Pay per view’ or ‘sale by single article’ (provided by publishers, aggregators or secondary database services) is another alternative, wherein one can access and download a particular full-text article against payment. Under ‘perpetual access’ to electronic journals, the right to permanently access licensed materials paid for during the period of a license agreement is provided.

Thus, a new form of access, i.e. ‘online access’ or ‘remote access’ has come into being, which denotes ‘access to materials not physically housed in the library’. This access could be to material subscribed or leased by libraries (fee-based or paid access) or to material available in the open or public domain on the Internet (free access).

As a result, the entire approach and philosophy of a ‘library collection’ has undergone a change. The collection of the traditional library was more or less homogenous in format,
consisting primarily of various types of print documents physically housed in it. Whereas the collection of the LIC of today is multi-faceted. It is a mix of print documents, offline electronic documents and online electronic documents, leading to the ‘hybrid library’. The collection is not confined within the four walls of the library – the networked environment has broken down geographical and physical barriers, giving a totally new meaning to what constitutes a library’s collection or what it ‘owns’. Gorman (1) says, “In traditional librarianship, a collection consists of physical items that have been purchased. In the modern library, which relies heavily upon access, what constitutes a “collection” becomes somewhat metaphysical”. The changing nature of the collection has influenced all the other library functions. The library catalogue especially has undergone a revolutionary transformation, and is representative of the global, diverse nature of the library collection. From cataloguing of locally held resources, its coverage has expanded to include cataloguing of resources not owned by the library (e.g. free Web pages), or resources owned, but not held locally (e.g. paid online journal or database).

Townsend Kane (2) brings out how the concept of ‘ownership’ is undergoing a change in today’s context:

<table>
<thead>
<tr>
<th>Traditional model of ownership</th>
<th>Contemporary model of ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libraries as a warehouse of information</td>
<td>Library as a gateway to information</td>
</tr>
<tr>
<td>Librarian as a collector of information</td>
<td>Librarian as a gatekeeper to information</td>
</tr>
<tr>
<td>Bigger is better (emphasis on size of a collection)</td>
<td>Availability is key (emphasis on the availability and delivery of information)</td>
</tr>
<tr>
<td>Library as a stand-alone entity (with multiple comprehensive collection)</td>
<td>The library as a link in a network of shared resources</td>
</tr>
</tbody>
</table>

Table 3.1: Access versus Ownership - A comparison
The “Access Versus Ownership” debate has taken roots. Supporters of the access policy argue that most of the material in the future will be made available only in electronic form, and access is more cost-effective than ownership. A library that is based solely on ownership of information resources in print form would have difficulty in maintaining its collection, with no hope of developing it due to rising costs of library materials and insufficient budgets. It would have to make difficult choices (say between monograph collection and periodicals collection), and would simply have to purchase fewer items than it has in the past. The ‘just in case’ approach to collection development by physical acquisition of resources needs to be replaced by ‘just in time’ approach, under which information is accessed upon demand.

Supporters of the ownership policy point out that a library that is solely based on access would be at the mercy of those who actually own the information. It has been observed that though access is being considered as an alternative to ownership, it is mostly with regards to periodicals, mainly scientific and technical periodicals. Already, there are some fears/problems associated with subscribing to only the electronic version of a periodical – stopping of access to the archives once the subscription is cancelled, payment for access to many journals that are not relevant in order to get the few that are (due to the "all or nothing” arrangement under bundle subscriptions), prohibition of off-site access, etc. So also, it is predicted that the costs of subscriptions to online periodicals will heavily increase and the full-text e-journals presently available free of charge on the Internet will not continue to be so available.

A report by the Committee set up to formulate an Information Technology strategy for the LC (3) has identified four levels of collecting by a digital library, based upon where the information resides and what level of commitment the library makes to management and long-term preservation:

- Archived -The material is hosted at the library and the library intends to keep the intellectual content of the material available on a permanent basis.
- Served -The material resides at the library but the library has not yet made a commitment to keep it permanently.
• Mirrored - The library hosts a copy of material that also resides elsewhere and the library makes no commitment to maintaining the contents. At this level, another institution has responsibility for the content and its maintenance.

• Linked - The material resides elsewhere and the library points to that location but has no control over the information.

The first three levels can be related to the ‘ownership or preservation’ approach, while the fourth can be considered to be the ‘access’ approach.

3.1.1 Free Internet-based Resources

In case of free Web-based resources, the access versus ownership dilemma has to be viewed in a slightly different perspective, since this is not about choosing between print and digital formats. No cost in terms of subscription/license/purchase is involved. Here ownership could be viewed in terms of downloading and preserving on the local servers a copy of relevant Web resources, and providing continued access to them. This selective capture of relevant Web resources is required due to their vast numbers, lack of quality control, possibility of permanent removal from their host servers, etc. These problems are more pronounced in case of free resources, since their creators/owners are not under any obligation to ensure all this.

3.1.1.1 Ownership by Local Preservation

Atkinson (6) supports the storage aspect in the electronic environment saying that there is a need “to identify resources that are likely to be of greatest interest locally and downloading these to a local database - a kind of deferred collection development operation”. Holleman (4) makes an interesting point, saying that “ownership is the quickest form of access…. access without ownership is much more expensive than simple ownership”.

The advantages and disadvantages of ownership by preservation locally are:
Advantages:

- Resources are preserved on local servers, therefore risk of loss due to impermanence of Web resources is minimized
- Faster access to the selected resources
- No problems of slow or poor network connectivity or network failure
- Resources can form a part of the comprehensive full text database and become available seamlessly along with other resources such as paid electronic resources, locally produced digital resources, etc.

Disadvantages

- Large server space required by the LIC
- Tasks of selection, downloading and further preservation and maintenance of resources are required to be performed by the LIC
- The functionality of the Web resource is reduced or lost in some cases since it is lifted out of the hypertext Web environment (e.g. interactivity)
- Continuous efforts are required in order to replace superseded/earlier versions with new ones.

3.1.1.2 Access to Remotely-held Resources

Access to free Internet resources involves delivery of information from the location in which it is held (Web sites on remote hosts) to the user. The role of the LIC is that of an intermediary. The responsibility of preservation and ensuring access lies entirely with the owner/creator of the Web resource. Following are the advantages and disadvantages of this option:

Advantages

- LIC does not need large server space
- Time and efforts spent for downloading, preservation and maintenance are saved
- Users can make use of the full functionality of the Web resource they have been directed to (e.g., explore the links to citations given in journal articles)
• Access can be provided to different levels of resources – Web site, directory in a Web site, specific resource, etc. Hence the difficult decision of what specificity/granularity a Web site/page should be preserved is not needed to be made.

Disadvantages
• Web resources are not permanent, hence users could be deprived of relevant resources that had once existed
• Very good Internet connectivity is needed
• Access places a lot of demand upon computer technology and technical support
• Dependence on Internet network leads to inability to access resources on servers facing temporary network problems or being unavailable due to housekeeping, etc.

Thus, both access and ownership have their own advantages and disadvantages and rather than make a choice between ownership and access, libraries need to build a collection comprising both, materials that are owned and materials that are accessed. There is a need to balance the two and make access a viable partner to complement ownership. In the words of Crawford (5).

“The future means both print and electronic communication.
The future means both linear text and hypertext.
The future means both mediation by librarians and direct access.
The future means both collections and access.
The future means a library that is both edifice and interface.”

3.2 DIGITAL PRESERVATION

It has long been the responsibility of libraries and archives to assemble, organize, protect and preserve documentation of human activity. Librarians and archivists have carried out their social responsibility by increasing the chances that the evidence about how one lives, how one thinks, and what has been accomplished is preserved. Conway (6) says, “Preservation is the acquisition, organization, and distribution of resources to prevent further deterioration or renew the usability of selected groups of materials”. Traditional preservation has been an extension of the custodial function, under which importance was
given to preservation of the ‘physical form’ – manuscripts, books, sculptures, gramophone records, etc. In the digital world, this traditional preservation concept is undergoing a transformation, due to the special features of the digital resources.

Digital Preservation is concerned with ensuring that records which are created electronically using today’s computer systems and applications, will remain available, usable, and authentic in the future, when the applications and systems which were used to create and interpret the record would no longer be available. A few definitions that bring out the essence of Digital preservation are:

The definition adopted by UK CEDAR Project is widely used in Europe - “Storage, maintenance, and access to a digital object over the long term, usually as a consequence of applying one or more preservation strategies”. (7)

Short, medium and long definitions of digital preservation have been developed by the Preservation and Reformatting Section (PARS) of the American Library Association: The medium definition says, “Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time”. (8)

Research Libraries Group (9) gives the following definition: “Digital preservation is defined as the managed activities necessary: 1) For the long term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology”.

Thus, digital preservation involves preserving more than just the record’s bit stream. Efforts are also needed to interpret the bit stream - without interpretation, the bit stream is nothing more than a meaningless series of 0’s and 1’s. During preservation, questions of not only the context, content and structure of the digital resource need attention, but also
its appearance and behaviour, which are aspects that are peculiar to digital records. The record must be preserved so that it retains its integrity and is authentic and usable. Ensuring continual access involves keeping up with technological changes. It is necessary to identify the multiple aspects of digital document that must be preserved. Next, preservation measures should ensure that as many of these aspects as possible persist over time.

According to Graham (10), preservation of digital information needs to be looked at as three problems: medium preservation, technology preservation and intellectual preservation. The concern of medium preservation is the same as in the case of traditional print material, where preservation action is needed in order to prevent damage and decay of physical material or medium. So also, the physical media on which digital information is stored, such as tapes, disks, optical disks, CD-ROMs, etc. might degrade over time or may become corrupt, and hence preservation action is necessary. More serious is technology preservation. Digital information is stored in the form of bits – ones and zeros, which denote values in binary notation. These bits have no inherent meaning, but represent encoding of information in accordance with some predefined scheme. This information cannot be directly interpreted by a user, but rather requires the mediation of software capable of translating that information into human-readable form. The need for technology preservation arises due to technology obsolescence, which is the result of rapid changes in the means of recording, storage formats and software that allows electronic information to be used. Intellectual preservation addresses the integrity and authenticity of the information as originally recorded. Preservation of the medium and of the software technologies will serve only part of the need if the information content has been corrupted from its original form, whether by accident or design. The need for intellectual preservation arises because the ease with which an identical copy can be made is paralleled by the ease with which a change may undetectably be made.

**Aspects to be preserved:** Preservation of digital resources is much more complex than that of non-digital material such as paper. For example, when a book is preserved, all aspects of the book are preserved – its physical presence, format, layout and its content,
since all these individual elements are inextricably linked. Digital formats on the other hand can be easily separated into individual elements, and hence more efforts are needed to preserve them as a whole (one can retain the content of the electronic document while losing the layout, e.g. converting HTML to PDF format; or one can maintain its physical presence, i.e. the computer file, but fail to preserve its readability). Following options have to be therefore considered:

- **Preserve the physical presence:** Preserve the computer file, the series of “1”s and “0”s. Also called ‘bit preservation’
- **Preserve the content:** Since information content is the most crucial component, ability to access it at its lowest level, such as ASCII text, without the embellishments such as font variations or layout features would have to be considered.
- **Preserve the presentation:** The original format or layout may include different font faces and sizes, columns, margins, the use of white space, etc. that carry out functions such as emphasizing important parts of the text, improving clarity, etc. In many formats of digital documents (e.g. SGML, XML), the layout specifications are separate from the content. To maintain the original look of the document, these layout specifications must also be preserved.
- **Preserve the functionality:** Digital documents are much superior in functionality than their traditional print counterparts. They might contain multimedia components, exist in hypertext format, and have navigation functions such as toolbars or keyword search or interactive tables of contents. Special efforts are required to preserve the functionality.
- **Preserve the authenticity:** Digital documents can very easily be altered, copied or moved. They have to be secured against unauthorized changes. One must be able to distinguish the original digital document from other versions or editions.

### 3.2.1 Strategies/Techniques for Digital Preservation

While digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preservation of this information poses a significant challenge. Unless preservation strategies are actively employed, this
information will rapidly become inaccessible. Choice of strategy will depend upon the nature of the material and what aspects are to be retained.

**Refreshing** - Refreshing or copying information without changing it ensures that information is stored on newer media before the old media deteriorate beyond the point at which the information can be retrieved (e.g. transferring from CDROM to DVDROM). This strategy is in part migration, but involves transfer only from one storage medium to another.

**Migration** – Migration or conversion involves the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. This approach focuses on the digital object itself and aims at changing the object in such a way that software and hardware developments will not affect its availability. The underlying principle is that the formats and structure of data may be changed, but the semantics of the underlying content is preserved. File formats are converted into compatible new formats as soon as the original formats face the risk of becoming obsolete (for example, HTML 3.2 to HTML 4.01 or PDF 1.1 into PDF 1.7). This way, the digital publications will be prepared for use, until the current format faces the risk of becoming obsolete itself. At that time another migration procedure will need to be carried out. Sometimes, digital objects are migrated to a small number of standard formats that are considered to have a high longevity, at the time of ingest into an archive. This process is known as normalization. For example, images in formats such as GIF, JPEG, BMP, PNG might all be converted to TIFF. A major drawback of migration is that while converting documents from one form to another, some aspects of the document’s layout or even the data itself might be lost. If preserving the original “look and feel” of the document is important, or when one is dealing with dynamic objects or those containing active elements (e.g. computer programmes providing functions such as animation), then migration might not be the best solution. Migration offers one method of dealing with technological obsolescence. It is considered a very practical approach, and is one of the most popular strategies.
Emulation – This approach focuses on the environment in which the object is rendered. It involves retaining information about how a digital collection was created and accessed so that future access can be accurately and authentically reproduced. An emulation tool generates an authentic view by launching the original viewer in the context of the original platform, making them work in future environments. An emulator is a computer programme running on platform ‘B’ that takes the binary files that can be run on platform ‘A’ as input and runs them as they would have run on platform ‘A’ (for e.g., AppleWin emulator for Apple II for Windows). Thus, emulation retains the functionality, look, and feel of the original document. Different approaches to emulation are possible – software emulation, operating system emulation, hardware emulation, etc. A serious drawback is the complexity of developing and maintaining emulation tools (in the future, several emulation tools will have to be maintained). Costs could be high.

Technology Preservation – This involves preservation of the technical environment by conserving copies of the software and hardware. This requirement to keep every version of all software and hardware, operating systems and manuals, as well as relevant skills, does not make this a feasible strategy. The hardware and software for digital media changes so rapidly that it would be impossible to keep an up-to-date ‘technology museum’. Issues such as space and cost for maintenance of hardware, decreasing support for software and hardware, unavailability of hardware parts and skills required to operate the hardware and software make this an impractical solution for the long term. A more practical approach would be to select a few and map all content to them.

Conversion of Data to Standardized Format – Data could be transferred to ASCII format, which is widely used, backwards compatible when used with Unicode, and utilizes human-readable characters. It retains information, but not the structure information is presented in. For higher functionality, XML (eXtensible Markup Language) can be used. This is a text-based markup language for describing the structure and meaning of data. It is an open standard defined by the World Wide Web Consortium and is not tied to any particular type of hardware or operating system. Conversion of records to XML format can be seen as a particular type of migration approach. XML
could be particularly useful in storing metadata and linking that metadata to the data files making up a record.

**Encapsulation** – The technique that involves grouping together both, the digital object and anything else necessary to provide access to that object (e.g. the original software needed to access and interpret it) has been proposed by a number of researchers as a possible future strategy. The result would be a ‘digital tablet’, a self-contained device.

**Virtual Machine Software** - This is an extension of the encapsulation and emulation approaches. This addresses the problem of interpreting data files in the future by writing a programme to carry out this interpretation in the machine language of a “Universal Virtual Computer” (UVC). This programme or the emulator would be written in the UVC language at the time the record was archived (without requiring any knowledge of the future target machine) and would be preserved together with the record. This programme runs on a UVC Interpreter, i.e. a virtual machine. With this approach, the data can be stored in any format and the knowledge required to decode it is encapsulated in the UVC programme.

**Freeze it / Print to paper or microfilm** – This was one of the early approaches, where digital records were transferred to an eye-readable file like paper or microfilm. However, this is not a viable preservation method since the functionality of records in their digital form is totally lost. At the most ‘flat data’, such as text and some still images, can be printed to paper without loss of data but with some possible loss of functionality. This could at the most be used as an interim solution.

### 3.2.2 Challenges in Preservation of Web Resources

Libraries need to select and archive Web resources in the same way as they have print formats, so that users can be assured indefinite retrieval and delivery of what is otherwise an ephemeral format. Web resources have all the features of digital documents, plus some additional ones, leading to complexities in their preservation. Since Web resources reside on the server that hosts them, and have no other physical presence, their continued...
existence depends upon the policies of the host organization/individual. More efforts are required by LICs to ensure that they remain available by capturing and storing them on local servers before they disappear from the Web.

**Scope or Volume:** The total number as well as size of Web sites is growing rapidly. Even the sophisticated harvesting facilities of the Internet search engines can cover only a small percentage of all Web resources. Unlike for print documents, not many selection guidelines or selection tools are available for Web resources. Publishers are not fully aware of their responsibility of reporting new titles to preservation agencies, and legal deposit is not well established. In bulk collection using automatic Web crawlers, human intervention for selection is not required, the process is quicker and cost is much less. However, this could result in very large but poorly organized collections. While in selective collection, in which librarians or other experts choose Web sites or Web pages, the result is well organized collections with high relevance and high quality. However, the selective approach is labour intensive, slower and costlier. Anything that is not explicitly selected might be lost.

**Formats:** Another problem posed is by the wide variety of formats that Web resources exist in, with new formats and versions introduced each year. Different digital objects have different preservation requirements, depending upon the reason the record is being preserved, how long it needs to be preserved, its original format, etc. Duplicates, i.e. many copies of the same documents at different places also adds to the trouble. Storage of dynamic sites is a complicated task. Web sites might include many large files in special formats such as audio and video files in addition to text. They might consist of executable computer programs, related material such as style sheets, files of metadata. Embedded links (hyperlinks) also pose a problem.

**Data Storage:** Web archiving can generate very large volumes of data, which need to be stored and managed. The largest Web archives have already reached the petabyte level.

**Boundaries:** The interconnectedness of the Web is one of its key features. The boundaries of individual Web sites or Web pages are hard to define due to the hypertext
environment in which they exist. They might be compound objects created by assembling different media (multimedia object) or by linking resources from around a network (hyperlinks). Preserving a complex object such as a Web site involves two things – preservation of the individual components as well as preservation of the relationships between them. Whether one must preserve only the particular Web site or Web page, or also pages to which they link, how many levels of linked-to pages must be preserved etc. are questions that need to be answered. Very clear policies are required for fixing the Web resource as a discrete whole, i.e. deciding up to what level the Web resource should be preserved.

**Change:** Web sites and Web pages get created, change, and disappear at an astonishing rate. Mechanisms for version control and location such as unique identifiers and permanent naming are required. A continuous effort needs to be made to track and preserve new sites or pages. So also, decisions need to be made about how often the different generations or versions of a site or page must be downloaded and preserved, whether and for how long earlier versions should be maintained, etc. In case of preserving of entire Web sites, the time needed to download may be several hours, and if changes are made to the site while it is being collected, the snapshot may be incomplete or inconsistent. One must also cope with sites or pages that become inaccessible either due to being temporarily off-line, or due to the server/network connection being down, or because they have moved elsewhere.

**Technical evolution:** Print documents remain legible for hundreds or thousands of years without any special equipment or preservation treatment. However the ability to access and use digital documents depends upon the hardware and software with which they were created. Digital documents rely upon the software that was originally intended to interpret or display them. When that software becomes obsolete, the problem arises of how to read that record. Different versions of the application might not read the file in the same way, and this may result in a change in the interpreted record that affects its archival integrity and authenticity. Data may be lost or gained. There may be no way to compare a new version with the original, so changes may go unnoticed.
Authentication: Documents archived must remain unchanged and true to the original. Encryption, hashing, time stamping, watermarking, digital signatures are potential solutions.

Intellectual property rights or Copyright: Copyright is a more complex issue in an online environment. For example, in multimedia sites, several creators would be involved – authors of text, photographers, software developers, graphic artists, etc. For online journals, copyright might be held with individual authors rather than the publisher. Several Web pages are protected by copyright. Creating an archive of such materials without explicit permission of the owner raises issues of copyright violation. In some countries changes to copyright law to permit such archiving are planned. Countries that have enacted the legal deposit legislation or have a legislation process underway, covering both offline and online digital publications include: Canada, Denmark, Finland (expected in 2008), France, Germany, Iceland, New Zealand, Norway, South Africa, Sweden (have received a Government decree which permits them to harvest the Swedish portion of the Internet), United Kingdom.

3.2.3 Web Preservation Projects

There is no single organization (or set of organizations) responsible for the Web. It was developed in a decentralised way and has no governing body that can mandate the adoption of standards or Web site preservation policies. Instead, decisions about Web content and delivery are devolved down to Web site owners themselves. In the absence of one single organization that would be responsible for preserving the entirety of the World Wide Web, the responsibility for preserving defined subsets of the Web was bound to fall to a range of organization. At the forefront of the effort to preserve the ever changing, dynamic medium of the Web are the National Libraries, who realized that their legislated mandate of collecting printed documents must also be extended to electronic documents. Many documents are being published only in digital form, with the numbers increasing rapidly. If this information is not taken care of, a considerable part of the cultural heritage might be lost. The year 1996 saw a flurry of activity in the Web preservation area.
Several Web preservation projects began, with the intention of experimenting with how entire Web sites could be preserved.

By the early 21st century, several of the preservation projects, which had been started as pilot or experimental projects became well established and continue to exist. Some projects ended after publishing reports containing observations and recommendations. Taking cognizance of the fact that Web preservation is a mammoth task, and cooperation at various levels is necessary for its successful implementation, several collaborative projects came into existence. The following is a chronological account of some of the important Web preservation projects:

**National Library of Canada: Electronic Publication Pilot Project (EPPP)** - From June 1994 to July 1995, the National Library of Canada (NLC) conducted a pilot project to acquire, catalogue, preserve and provide access to a small number of Canadian electronic journals and other representative publications available on the Internet. This covered e-journals “published” in Canada, or sponsored or produced by a Canadian company or individual. The EPPP became part of the overall NLC Web structure in June 1995. A very comprehensive report, “Electronic Publications Pilot Project (EPPP): Final Report” (11) was published in June 1996. Acceptance of this report by the NLC Executive Committee formally completed the project. The Library and Archives Canada (Bibliothèque et archives Canada) continued to collect electronic publications on a voluntary deposit basis until its Act of 2004 extended legal deposit to online publications.

**National Library of Australia: PANDORA** – The PANDORA Project that started in July 1996 is now a routine part of selection, acquisition and cataloguing process of the National Library of Australia. PANDORA archive constitutes a strongly representative sample of Australian Web publishing by academic, government, commercial and community organizations. It is now built in collaboration with nine other Australian libraries and other cultural collecting organizations. The purpose of the PANDORA Archive is to collect and provide long-term access to selected online publications and Web sites that are about Australia, or are by an Australian author on a subject of
significance and relevance to Australia. A comprehensive archive management system called PANDAS (PANdora Digital Archiving System) has been introduced, which uses HTTrack as its default crawler. In addition to the selective approach, the National Library has undertaken large scale harvests of the .au domain in June – July 2005 and August - September 2006.

**Internet Archive** – This is a non-profit organization based in San Francisco, existing since 1996 with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format. Its origins lie in Alexa Internet, a Web cataloguing company set up by Brewster Kahle and Bruce Gilliat in the same year. Alexa harvest a snapshot of the World Wide Web every two months. These snapshots are donated to the Internet Archive, forming the main part of its collection. The archive is collaborating with institutions including the Library of Congress and the Smithsonian Institution. It has access to advanced technology and expertise in Web crawling, automatic indexing of Web sites and the techniques of managing massive collections. It has adopted the strategy of bulk collection. The Internet Archive includes text, audio, moving images, and software as well as archived Web pages. The public can access the archived versions through an interface, the “Wayback Machine”, which not only provides access to individual Web site snapshots, but also allows them to be browsed within their historical context.

**The Royal Library, National Library of Sweden (Kungl Biblioteket - KB): Kulturarw3 Program** – This began in 1996 and adopts a bulk strategy, covering Swedish electronic documents. It ranks among the longest-running Web archiving programmes. The collection covers electronic magazines and newspapers, static documents such as texts in electronic archives and dynamic documents with links. 12 sweeps have been carried out till 14 February 2005.

**Nordic Web Archive (NWA)** - The Royal Library of Sweden took the initiative in setting up the Nordic National Libraries' forum in 1997, and the other participants were the Royal Library of Denmark, Helsinki University Library of Finland, the National
Library of Norway and National and University Library of Iceland. The aim of the forum was co-ordination and exchange of experience in the fields of harvesting and archiving Web documents between the different national projects that were in place or about to start. In 2003, the individual members of NWA joined the International Internet Preservation Consortium. The NWA partners are developing WERA (WEb ARchive Access), an access system for Web archives as a part of the IIPC Toolkit.

Networked European Deposit Library (Project NEDLIB) – Under the leadership of The National Library of the Netherlands, eight national libraries, one national archive, two developers of information technology and three major publishers (Kluwer, Elsevier, and Springer-Verlag) from the European Union participated. They developed the NEDLIB Harvester, which has been used in a number of Web archiving programmes.

CEDARS project - This was funded by the Joint Information Systems Committee (JISC), UK and was carried out between 1998-2002. It aimed to provide guidance in best practices by developing practical demonstrator projects and sponsoring strategic working groups.

Library of Congress: MINERVA Project - Work began in 2000 at the LC to develop a prototype system to collect and preserve materials from the Web under Project MINERVA (acronym for Mapping the INternet: the Electronic Resources Virtual Archive). A prototype was developed using a few sites to gain insights into the practical issues involved in collecting and organizing selected Web sites. Based on the prototype, an Interim Report was published in January 2001 (12), with the Final report being published in September 2001 (13). Use of automated tools with skilled librarians establishing and monitoring the procedures, development of preservation programmes in partnership with libraries, archives, publishers and other organizations etc. were some of the recommendations.

National Digital Information Infrastructure and Preservation Program (NDIIPP) – Since 2000, the Library of Congress is working with partners from universities, libraries,
archives, federal agencies and commercial organizations to identify and collect at-risk born-digital content, build and support a national network of partners, and to develop tools and services for preservation.

Bibliotheque Nationale de France – The BnF has been developing and testing Web archiving techniques since 2000. It undertook two pilot projects, one to test methods for automatically identifying and harvesting sites using link ranking, and the second to investigate techniques to capture database-driven sites, leading to the development of the DeepArc tool.

International Internet Preservation Consortium (IIPC) - In July 2003, the IIPC was established at the Bibliotheque Nationale de France with 12 participating institutions including the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA). The mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations.

UK Web Archiving Consortium (UKWAC) – Set up in 2003, this comprises six UK organizations – the British Library, the National Archives, the National Libraries of Wales and Scotland, JISC and Wellcome Trust. The UKWAC Web archive was publicly launched in May 2005. A notable feature of the consortium has been the development of thematic collections, to which individual partners contribute sites within their sphere of interest.

The German National Library (DNB): Kopal program (Co-operative Development of a Long-Term Digital Information Archive) - The German legal deposit law has recently been expanded to include born-digital materials, including e-journals, online academic publications, e-newsletters, and Web sites. In order to fulfill the new deposit law, the DNB has set up a digital long-term archive under the project called Kopal in 2004. The German National Library is working with other academic and non-academic partners for long-term archiving and preservation of digital documents.
Thus, there is a wide diversity in the scale and approaches adopted by Web archiving programmes. Organizations of all sizes, from major national libraries and archives to consortia and university departments have played a vital role in Web archiving.

### 3.2.4 Digital Repositories or Archives

Preserving the Web in the bulk mode or even selective mode is a daunting task, as shown by the Web preservation projects, and needs high investments in hardware, software and related infrastructure. Technicalities are involved since Web sites increase or decrease in size, the depth and breadth of information contained in them may change, and information might be modified, amended, increased, or decreased. Naturally, the job can be undertaken only by large organizations with high funding and advanced facilities, that too in collaboration with other similar organizations.

During 2002, the development of digital repositories emerged as a new strategy in the continuing networked information revolution. JISC (14) defines a digital repository as, “a digital repository is where digital content, assets, are stored and can be searched and retrieved for use later. A repository supports mechanisms and workflows to import, export, identify, store and retrieve digital assets”.

Universities took the lead by establishing institutional repositories, which are digital collections providing access to the full text of scholarly material produced by its members. Lynch (15) defines an institutional repository as follows: “A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.”

Thus, institutional repositories are digital repositories that capture and preserve the intellectual output or content generated by the communities of the institution - published
articles, books, preprints and working papers, conference papers, teaching materials, student theses, data-sets, etc.

Heery and Anderson (16) bring out the following characteristics of an institutional repository, which are also applicable to digital repositories in general:

- Contains content deposited by owner, creator, or third party
- Repository architecture manages content as well as metadata
- Offers a minimum set of basic services – e.g. put, get, search, access, control
- Repository must be sustainable and trusted, well-supported and well-managed

Soon, other institutions such as research laboratories, publishers, libraries, and commercial organizations were creating digital repositories that address the entire lifecycle of information—from supporting the creation and management of digital content, to enabling use, re-use, and interconnection of information, to ultimately ensuring long-term preservation and archiving.

Another variation of digital repository is the ‘Subject Repository’, which collects and manages material relating to one or more related subject areas. A subject repository may be an independent collection or formed by partitioning one or more Institutional Repositories by discipline.

There is a lot of inconsistency in use of the terms digital archive, digital repository, and digital library. A digital archive can be considered to be a term broader in scope than digital repository. The Task Force on Archiving Digital Information 1996 defines digital archives in functional terms as “repositories of digital information that are collectively responsible for ensuring, through the exercise of various migration strategies, the integrity and long-term accessibility of the nation’s social, economic, cultural and intellectual heritage instantiated in digital form.”
The Tufts Digital Repository Program run by the Tufts University (17) brings out the difference between a ‘digital library and ‘digital repository’ thus: “It is useful to think of the digital library as ‘presenting’ tools and digital content from the repository for access and use, and to think of the digital repository as ‘preserving’ digital content for the long term… the digital library is just one application that can access and use the content in the digital repository and provides tools to access digital content in the digital repository, as well as viewing and using those resources”.

Maureen Pennock (18), in “Curating Emails” tries to bring out the difference between digital repository and digital archive. She says, “the meaning of the term ‘digital repository’ is widely debated…this requires not just software and hardware, but also policies, processes, services, and people, as well as content and metadata. Repositories must be sustainable, trusted, well-supported and well-managed in order to function properly. A digital archive has the same essential characteristics but with an additional and express commitment to preserve the integrity of the contents for the long-term, and provide for long-term storage and access to its contents”.

Thus, both, repository and archive have preservation as a goal. In an archive, the services are tailored for periodic deposit and infrequent delivery. They are not necessarily required to provide for a high-traffic delivery environment or daily end-user access. They are optimized for trust, not speed. The system driver is preservation, leading to access. While in a repository, frequent access to the stored digital content is catered to.

The basic steps in creating a repository or an archive are given by NASA’s Reference Model for an Open Archival Information System (OAIS) (19). It splits archiving into six functional entities:

1. Ingest: Getting records into an archive, including the capture of appropriate metadata.
2. Data Management: Controlled editing of data input into the system.
3. Storage: Physically storing records in an archive, including the creation of an appropriate backup policy, regular media migration etc.
4. Access: This covers two related aspects: finding records within the archive and disseminating them to consumers. This includes ensuring that the appropriate information is only disclosed to appropriate users of the system.

5. Preservation Planning: Ensuring that the contents of an archive remain more than just a meaningless bit-stream.

6. Administration: Running of the system itself including its maintenance.

In 2005, the Coalition for Networked Information (CNI) surveyed its academic member institutions to examine the current state of institutional repositories in the US (20). Though initially developed for organizing e-prints of the works of the faculty, it was observed that repositories were already including or planning for inclusion in the near future a wide range of materials such as electronic theses and dissertations, digitized reports, multimedia, course materials, datasets, digital images, digital audio, software, etc. Thus, any digital content that is worthy of preservation and access is becoming a part of repositories, and what goes into a repository is more a policy decision made by each institution or administrator.

3.3 ACCESS TO WEB RESOURCES

While on one side experiments and pilot projects were being taken up at different levels to preserve the contents of the Web, on the other side thought was being given to designing systems for identifying Internet-based information resources and enabling users to access them regardless of their physical location.

As the amount of resources available through the Internet has grown, various information retrieval tools have been developed. One approach is by automatic indexing of the full text of Web resources, the method used by search engines. Automatic indexing is done through robots that traverse the Internet, automatically extracting content (irrespective of context or quality) and submitting it to a database (e.g., AltaVista, Lycos). Automatic indexes or search engines have certain advantages. They are useful for locating specific resources where all or some details are already known. They are popular with users since
they use natural language. It is a cheap means of producing a large catalogue of searchable material, and hence can be much more comprehensive in coverage. Every word on every page of every item can be indexed.

The biggest drawback with search engines is that they do not differentiate in terms of quality of resources they retrieve. The quality is variable; resource descriptions are unintelligible as they are just extracted from the resource. Commercial Web sites, student home pages, high quality research papers and entertainment and sports pages all have the same, undifferentiated status. The results can be overwhelming, unmanageable and full of irrelevant references.

Wells, et.al. (21) say, “While these search engines (Yahoo and Alta Vista) and others like them have strengths, their weaknesses are well known: a high percentage of non-authoritative content mixed with quality content that, when indexed together, makes locating relevant information serendipitous at best”.

The other method for indexing the Internet-based resources was derived from the traditional functions carried out by libraries. A new access tool or service was developed, known as a ‘subject gateway’. Here, the subject experts and information professionals select, classify and catalogue Internet resources to aid search and retrieval for their users. What results is access to a quality-controlled collection of resource descriptions which users can search by keyword or browse by subject area. The description helps users to assess the origin, content and nature of Internet resources, enabling them to decide if it is worth investigating further. Links are checked periodically to ensure that they are working.

Since resources are only included if they fulfill certain criteria, the subject gateway databases are smaller and much more focused than the huge search engine databases. The results are specific, and linked to relevant documents. Thus, one of the primary objectives of subject gateways is to assist information seekers to make sense of the “chaotic
repository for the collective output of the world’s digital printing presses”, as described by Lynch (22).

### 3.3.1 Subject Gateways

Some definitions of a Subject Gateway that bring out its purpose and activities performed are:

Lorcan Dempsey (23) says that, “Subject gateways are Internet services which support systematic resource discovery. They provide links to resources (documents, objects, sites or services) predominantly accessible via the Internet. The service is based on resource description. Browsing access to the resource via a subject structure is an important feature”.

The definition given by the IMesh Toolkit project (24) is, “A subject gateway is a Web site that provides searchable and browsable access to online resources focused around a specific subject. Subject gateway resource descriptions are usually created manually rather than being generated via an automated process. Because the resource entries are generated by hand they are usually superior to those available from a conventional web search engine.”

The DESIRE project (25) has defined subject gateways as, “Selective subject gateways on the Internet are characterised by their quality control. The core activities of resource selection and description rely on skilled human input (by librarians, academics and experts) and are not activities that lend themselves to automation.”

According to Traugott Koch (26), “Quality-controlled subject gateways” are Internet-services which apply a rich set of quality measures to support systematic resource discovery. Considerable manual effort is used to secure a selection of resources which meet quality criteria and to display a rich description of these resources with standards-based metadata. Regular checking and updating ensure good collection management. A main goal is to provide a high quality of subject access through indexing resources using
controlled vocabularies and by offering a deep classification structure for advanced searching and browsing”.

Thus, the characteristics of a subject gateway can be summarized as:

- An online service that provides links to numerous other sites or documents on the Internet
- Manual creation/intervention, often by information and/or subject specialists
- Selection of resources according to published quality and scope criteria
- Intellectually produced content descriptions, ranging in length from short annotation to review.
- Search and browse access
- Collection management policy, supported by maintenance and updating procedures.

A subject gateway can be considered as a bibliographic database, since it consists of a database of resource descriptions or catalogue records. Since a subject gateway represents Web resources that are dynamic in nature, it itself also has to be dynamic. Maintenance is of utmost importance, if the subject gateway has to be efficient and effective. One part of maintenance involves the ongoing catalogue development by identifying and adding new Web resources. The other is maintenance of the resource catalogue, which is very crucial. Under this, tasks involved are validating records, link checking and updating resource descriptions. Validating records is necessary to ensure that the record is accurate, i.e. there are no spelling mistakes, the cataloguing guidelines are adhered to, etc. Dead links are among the most persistent problems associated with collections of free Web resources, and therefore link checking is required to ensure that the resources are still physically available at the given URL. Dead links will have to be culled and the corresponding catalogue entries removed. For a gateway's collection to maintain its integrity and usefulness, the resource descriptions must also reflect the changes in the resources. Care needs to be taken to see that they still adequately reflect the content of the Web resources, since they are subject to rapid and unadvertised change of various types. The content of Web pages may change, their virtual locations (URLs) can change. IP
addresses can expire or move to another organization. Content of resources should be reviewed periodically, and cataloguing records should be revised accordingly. Catalogue records and links need to be deleted when a site can no longer be found, or its content no longer conforms to the collection scope and policy.

### 3.3.1 Landmarks in the Development of Subject Gateways

**Follett Report:** The most significant step that led to the establishment of subject gateways is the Review of libraries and related provision in higher education in UK commissioned in 1992, and the recommendations given in the resulting comprehensive report, popularly known as the Follett Report (27). The purpose was to look into the needs of libraries across the sector. The Review Group was chaired by Prof Sir Brian Follett, and its membership was widely drawn, including nominees of all four funding bodies, and other representatives from both within and outside higher education.

The Review group met nine times between October 1992 and October 1993. The report was brought out in December 1993 and is one of the most influential of recent times and has resulted in several initiatives. The report consists of 8 chapters. The first chapter summarizes the Review Group’s general conclusions. These are developed in more detail in chapters 2 to 7, which contain a number of detailed recommendations. These recommendations are in turn summarized in chapter 8. Chapter 7 of the Report identifies areas where Information Technology might be employed to underpin significant shifts in the ways library services operate, and which, if implemented, would provide an impetus for change in the library sector.

In this chapter, the Report draws attention to the difficult task of classifying (networked) information resources, uniquely identifying them and then enabling users to identify and locate the information that they require. Discussing the concept of the electronic or “virtual library, the Report observes that librarians would essentially be enablers of information access, regardless of its form and location. The Report brings out the advantages of the shift in emphasis from ownership to access in the words, “As far as
libraries are concerned, the trend towards adoption of an access rather than a holdings strategy ... underlines the advantages which electronic delivery of articles and other documents can bring. It would enable material stored in one place to be readily transmitted elsewhere, and would thus make available the resources of the major libraries to other institutions which do not themselves have extensive holdings in traditional formats”.

**eLib Programme:** The Electronic Libraries Programme (eLib) grew out of the effort to implement recommendations given in Chapter 7 of the Follett Report, leading to the creation of subject gateways. It was an initiative of the Joint Information Systems Committee (JISC) of UK. One of the projects set up during the first phase of the eLib programme was Access to Network Resources (ANR) in 1995. The aims of ANR were to encourage the development of networking navigation tools and local subject based tools and to raise awareness of networked information resources. Under ANR, funding was provided to establish the eLib subject gateways – SOSIG (Social Science Information Gateway), ADAM (Art, Design, Architecture and Media Information Gateway), EEVL (Edinburgh Engineering Virtual Library), OMNI (Organised Access to Medical Networked Information) and BizEd. SOSIG slightly predated eLib, and acted as a model for other gateways.

**Resource Discovery Network (RDN):** Since the subject gateway projects were successful, there was a desire to maintain these gateways on a more sustainable long-term basis. At the same time there was recognition of the need to co-ordinate the activities of these gateways. A new service, the Resource Discovery Network (RDN) was started in 1999 under funding from JISC. It worked as a co-operative network consisting of a central organization, the Resource Discovery Network Centre (RDNC) established as a partnership between King's College London and UKOLN at the University of Bath, with support from the University of Hull. RDN was a free Internet service dedicated to providing effective access to high quality Internet resources for the learning, teaching and research community. Each Web site and resource in the core database of resource
Resource Discovery Network was renamed “Intute” in July 2006 (28). Intute is a composite word - derived from 'Internet' and 'Tutorial' and is intended to convey the experiences of guided learning and online resource discovery. Funded by JISC, it is a consortium of seven universities working with several partners. It is a free online service built on the key principles of quality, consistency, and interoperability. Subject specialists select and evaluate the Web sites in Intute database and write high quality descriptions of the resources. The eight subject gateways under RDN were re-organised to create four major subject groups: Arts and Humanities (Artifact and Humbul), Science, Engineering, Technology (EEVL, GEsource and PSIgate), Health and Life Sciences (BIOME), Social Sciences (Altis and SOSIG).

**DESIRE (Development of a European Service for Information on Research and Education):** In Europe, a collaborative project named DESIRE was established among ten institutions from four European countries - the Netherlands, Norway, Sweden and the UK. It was funded by the European Union. The aim was to promote the use of the Web within the European Research Community. The DESIRE project ran from 1996-2000 as DESIRE I and II. Phase 2 finished at the end of June 2000.

In order to support the DESIRE vision of a network of European cross-searchable information gateways, UKOLN produced the DESIRE Information Gateways Handbook (29) under the Work Package 3. The handbook is designed to support libraries and other organizations who are interested in setting up large-scale information gateways on the Internet. It offers a step by step guide and points to tools, examples and documentation,
which can support the process of setting up a gateway. It is divided into three sections to reflect the separate managerial, information and technical concerns associated with building a gateway. Though primarily concerned with the development of larger scale gateways, the handbook can be equally applied to a gateway set up by a single person, since many of the issues are common.

**ROADS (Resource Organisation And Discovery in Subject-based services):** An impetus to development of subject gateways was provided by the ROADS project, a collaborative development project funded by JISC to design and implement a user-oriented resource discovery system and help people to catalogue Internet resources such as Web sites. The purpose of ROADS was to produce a software package for setting up subject gateways, to investigate methods of cross-searching and interoperability within and between gateways and to participate in the development of standards for the indexing, cataloguing and searching of subject-specific resources. The result was the ROADS toolkit, a freely available system of software and standards. This "toolkit" included tools such as an automatic link checker, statistical counters and record validators. ROADS templates were defined for 15 different resource types. Cataloguing guidelines were developed for the two most commonly used ROADS template types (Service and Document).

Many of the eLib gateways were constructed using ROADS, with the software being available from the ROADS software centre at http://www.roads.lut.ac.uk. The original ROADS project ended in July 1999. ROADS continued as open-source software for a few more years, maintained by an informal consortium of developers. As of July 2004, this project is no longer under active development (30).

**IMESH Toolkit Project:** “Imesh” is a contraction of “International mesh” and is a loose association of subject gateway activity. At the first Imesh Framework Workshop thirty-five delegates from twelve countries convened at the University of Warwick in Coventry, UK in June 1999 to discuss their shared interest in gateways. The Imesh Toolkit project resulted, which was funded by the National Science Foundation (USA) and JISC (UK) to develop a consistent framework for development of subject gateway software, building
on existing tools such as ROADS. One of the main aims of the project was to reduce the entry costs for new subject gateways, including reducing the effort required to support specialized or local functionality.

**PINAKES: A Subject Launchpad (31)**

The inspiration behind creating the Pinakes Web site comes from the catalogue of the ancient Library of Alexandria, “The Pinakes”, which was compiled by the poet Callimachus to locate relevant material from the evergrowing library. The Pinakes Web site was created with a hope to provide a similar function for Internet resources. It provides access to major, high-quality subject gateways on a wide range of academic subjects. It is hosted by the Heriot-Watt University, Edinburgh, Scotland.

This researcher had taken a snapshot of the Pinakes Subject launchpad in 04 May 2001 (site shows date of updation as 14 February 2001) and once again on 10 September 2007 (site shows date of updation as 13 August 2007). The snapshot in 2001 showed 38 subject gateways and 10 multi-subject gateways. The snapshot taken in September 2007 shows 41 gateways and 10 multi-subject gateways. A lot of changes can be observed from the gateways listed at the Pinakes site, with a few gateways having disappeared over the years, a few new ones established. Some have undergone structural changes (as in case of the RDN gateways, where the original 8 gateways have been merged to form 4 portals), while some have been renamed.

Out of the 41 gateways listed on the Pinakes site in 2007,

- 24 were also present in 2001 by the same name. Two of these (AGRIGATE and EdWEB), though still present on the Pinakes site, are not being maintained anymore, as indicated on the site. However, they still hold the databases, hence they are included in the list
- Five of those present in 2001 were renamed after merging into four Intute portals
- One was renamed from Human Languages Page: language-learning & linguistics to iLoveLanguages: language-learning & linguistics
• 12 were new
• Eight of the 38 gateways in 2001 had gone

Out of the 10 multi-subject gateways listed on the Pinakes site in 2007,

• Five were also present in 2001 by the same name
• Two of those present in 2001 were renamed (RDN named as Intute, Scout Report
  Signpost named as Scout Report Archives)
• One was not being maintained any longer (DutchESS)
• Two were new
• Two of the 10 multi-subject gateways present in 2001 were gone

Thus, a lot of additions, deletions, modifications are visible on the world subject gateway scenario.

During the literature review, it was observed that several terms were used interchangeably with the term “subject gateway” right from their inception: information gateways; resource discovery systems, subject-based information gateways (SBIGs), subject-based gateways; quality-controlled subject gateways, subject index gateways: clearinghouses; subject trees.

As subject gateways have matured, value additions are being done by adding some more services and they are being described by new terms: Hubs, Subways, Portals, Vortals or Vertical Portals, Virtual Libraries: Pathfinders, Web Search Portals. Information Services, Search Portals. Vertical Industry Portals. Internet Portals.

**Portals:** Several subject gateways have been further extended into portals. The subject gateway forms the core of the portal, with additional services such as access to complementary databases, current awareness, personalization, communication facilities such as bulletin boards and discussion groups, etc.

The Joint Information Systems Committee (JISC). UK defines a portal thus: "...a network service that brings together content from diverse distributed resources using technologies
such as cross searching, harvesting, and alerting, and collates this into an amalgamated form for presentation to the user... for users, a portal is a possibly personalised, single point of access where searching can be carried out across one or more than one resource and the amalgamated results viewed. Information may also be presented via other means, for example, alerting services and conference listings or links to e-prints and learning materials." (32)

3.3.2 Subject Gateways in India

**Syama Prasad Mookerjee Information Gateway of Social Sciences (SPMIGSS):** This gateway has been developed by the Indian Council of Social Science Research (ICSSR) for providing links to a variety of information sources in social Sciences. The gateway was not accessible at its URL (http://www.icssr.org/igss) after repeated attempts in 2007.

**The ICMR-NIC Centre for Biomedical Information (Indian MELARS Centre)** has compiled a list of useful web sites for the users to facilitate an easy access to the resources over the Internet. The service consists of a single Web page containing links to the resource/organization and a brief annotation. The gateway is available at http://indmed.nic.in/.

**Agriculture Gateway to India:** This is being maintained by AIM Lab. Main subjects are provided with subdivisions under them. Links are listed under the subdivisions without any annotation or description. The URL of the gateway is: http://aimlab.aces.uiuc.edu//diglib/india/default.htm.

**Gateway on Sustainable Development from India** has been developed by Sustainable Development Networking Programme (SDNP) of the Ministry of Environment & Forests. This was not accessible at its URL http://sdnp.delhi.nic.in after repeated attempts in 2007.

Thus, one finds that a few very rudimentary subject gateway services have been set up in India.
3.3.3 Subject Gateways Related to Defence

An examination of existing subject gateways has revealed that only one gateway devoted entirely to the defence technology field exists, namely AERADE (http://aerade.cranfield.ac.uk/)

**AERADE:** It has been developed by Cranfield University, UK to enable aerospace and defence experts to find quality, relevant information on the Internet. The service is a joint venture between subject specialists within the information and library services at Cranfield campus (aerospace) and Shrivenham campus (defence). Launched in November 1999, AERADE was developed from two pre-existing Cranfield University library initiatives, CRUISE and DEVISE, that were created in order to provide the institution's staff and students with access to quality Internet resources in their subject areas.

Since August 2000, the subject gateway component is referred to as ‘Aerospace and Defence Resources’, and AERADE has been converted to a portal by the addition of three new services – ESDU series abstracts (subscription based design data and methods for aerospace, and related engineering fields), DEVISE (Defence Virtual Information Service - collection of resources focussing directly on the information needs of the armed forces) and the Internet Aviator (an interactive tutorial for discovery, selection and use of Internet resources).

Other subject gateways might have covered defence related resources under one of the main classes. For example, Intute Science & Technology gateway has ‘Aerospace and Defence Engineering’ as a category under Engineering, with further subtopics (http://www.intute.ac.uk/sciences/engineering/). BUBL LINK has a subclass ‘Military and Nautical Engineering under the class ‘Engineering and allied operations’. (http://www.bubl/ac/uk/)

Since defence technology is a multi-disciplinary area, its component subjects such as mechanical, electrical, electronics and computer engineering, materials science, physics,
chemistry, etc. are covered in related subject-specific gateways, or in multi-subject gateways.

Thus it is seen that the last decade of the 20th century saw the initiation of several measures in order to safeguard the contents of the Web on one hand, and to improve resource discovery on the other. The Web had begun to be recognized as a serious medium for scholarly communication in addition to its other applications for personal and commercial purposes. Libraries, archives and publishers, all got involved in trying to adapt and apply the principles and techniques of traditional librarianship to the Web resources. Problems/challenges associated with the new medium were examined in detail by them. They developed tools and techniques for preservation of Web resources for posterity, and providing access to relevant Web resources through pilot studies, experimental projects and collaborative efforts.

The next chapter deals with the formulation of content-based quality criteria for the evaluation of Internet-based resources. It also gives the details of the development of a new set of evaluation tools or instruments for the three types of Web resources, namely HTML Web pages, PDF Web pages and Web sites.
REFERENCES


