University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal) | Libraries at University of Nebraska-Lincoln

4-1-2012

Building an Iran Web Archive in the National Library and Archives of Iran: A Feasibility Study

Farzaneh Shadanpour
Member of Research Board of NLAI, fshadanpur@nlai.ir

Saeideh Akbari Dariyan
Azad University, sakbaridaryan@gmail.com

Reza Shahrabi Farahani
General Director of IT Department of NLAI, r-shahrabi@nlai.ir

Soudeh Seirafi
Non book materials cataloguer in NLAI, soodeh_seirafi@yahoo.com

Alireza Vazifehdoust
Tehran University, a-vazifehdoust@nlai.ir

Follow this and additional works at: http://digitalcommons.unl.edu/libphilprac

Part of the Library and Information Science Commons

Shadanpour, Farzaneh; Dariyan, Saeideh Akbari; Farahani, Reza Shahrabi; Seirafi, Soudeh; and Vazifehdoust, Alireza, "Building an Iran Web Archive in the National Library and Archives of Iran: A Feasibility Study" (2012). *Library Philosophy and Practice (e-journal)*. Paper 726.
http://digitalcommons.unl.edu/libphilprac/726

StatCounter.com RSS Add to Google

http://unllib.unl.edu/LPP/

***Library Philosophy and Practice* 2012**

ISSN 1522-0222

Building an Iran Web Archive in the National Library and Archives of Iran: A Feasibility Study

Farzaneh Shadanpour
MA of LIS (Instructor)
Member of Research Board of NLAI

Saeideh Akbari Dariyan
PhD Student in LIS (Azad Univ.)
Deputy of General Information Department in NLAI

Reza Shahrabi Farahani
MA in LIS
General Director of IT Department of NLAI

Soudeh Seirafi
MA in LIS
Non book materials cataloguer in NLAI

Alireza Vazifehdoust
PhD candidate in Software Engineering (Tehran Univ.)
IT Department of NLAI/p>

Introduction

Internet and Web have provided content producers with different form of media; and a big deal of intellectual products of different countries are only found on the internet. Loss and change are among the obvious features of the Web ; and there is no guarantee that the information now existing on the Web will remain stable and usable in the distant future. Content change has different reasons ranging from the inclination of the creators towards the change and edition of parts of content to accidental changes during transfer, for instance, change into different formats. Even domain names and addresses fall prey to change and omission. As Berners-Lee put it: "There are no reasons at all in theory for people to change URIs (or stop maintaining documents), but millions of reasons in practice"(Berners-Lee,1998).

Institutes like national libraries and archives that, according to their own responsibilities, are obliged to collect publications and intellectual productions of any country deal with more complicated aspects and different scales of the issue. And the definition of the term "digital preservation" seems to suffice for its purpose. Identifying and harvesting the intellectual content of a nation from the interwoven world of the Web and the internet, organizing and making accessible this intellectual property, all, belong to the conceptual scope of digital preservation, which, according to a brief definition, "combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change" (*Definition of Digital Preservation*, 2010). These aspects assume special importance considering the mission vested in national libraries to provide a reference model in professional information work and specialized

activities of librarianship.

In compliance with the articles of association of the National Library and Archives of IR of Iran (hereinafter referred to as NLAI), establishment, making accessible, organization and preservation of the intellectual and cultural works and contents at national level, constitute the main responsibility of this organization [1].

With regard to production and publication of a large part of this content on the Web and awareness of the fact that Web resources are always changing, evolving and getting lost, development of an organized archive of Iranian Web resources is considered part of NLAI responsibilities. Such archive put the gathered Web resources at present and future researchers' disposal; and hence it can be considered a rich data center for the country [2]. Since any scientific project should be aware of scientific investigations and as there exist special internet and Web issues in Iran, present paper is the report of a feasibility study, which has dealt with the possibility of development of Iran Web archive in NLAI as well as its possible problems, necessities and requirements so that expert advices and information become the basis for the future decision makings.

Study Aspects

In Iran, various individuals, groups and institutes get involved in the internet-related policies. A plethora of technological, specialized, legal and statutory facts in Iran and the world influence the internet and Web , which should be logically investigated in this study. This study first analyzes the existing status and facilities including the internet in Iran, characteristics of the Iranian Web , legal aspects, the status of the "Iranian Legal Deposit Law", hardware and Telecom infrastructure, the software packages used in the NLAI, metadata and their application in the present NLAI system. Then, the comment section mentions executive and managerial techniques, proposed approaches for resource collecting, national and international cooperation, information management, legal aspects, information management standards, the specifications of the software and equipments required and some points on the budget and personnel.

Status of Internet in Iran

In Iran, the internet emerged from academic environments, and according to Rahimi (2003): "Internet use in Iran was first promoted by the government to provide an alternative means of scientific and technological advancement during the troubled economic period that followed the Iran-Iraq War". The internet was introduced into Iran through The Institute for Studies in Theoretical Physics and Mathematics (IPM) [3] in 1993 [4]. Connection to the internet was given to the academic users through IPM. Neda Rayaneh, for the first time, provided the public with the use of internet electronic mail via BBS. BBS was the first computerized virtual space in Iran, through which users exchanged information and files [5].

Iranian Ministry of Post, Telegraph and Telephone began the public sale of the internet access as of 1998. From the viewpoint of technology provision and infrastructure development, the internet, from the very beginning, has had several main custodians in the companies affiliated with the Iranian Ministry of Post, Telegraph and Telephone, which changed its name into Ministry of ICT in 2003. Such companies have been responsible for policymaking on equipments, commissioning, introduction and promotion of technology and development. These custodians have fallen prey to organizational and structural changes and transitions in the kind and scope of their responsibilities during these years.

In Iran, network infrastructure and access to the internet is conducted through two main networks: Public Switched Telephone Network and National Data Network. The switching Network, called IRANPAC with X 25 Protocol, which connects most of the cities together, is provided. Intercity Switching Network exists in more than 100 points in the country, and the international fixed and mobile communications is made through international switching centers [6]. International connection is provided by Zirsakht Company (Telecommunication Infrastructure Company (TIC)) through submarine optical fiber cable to the Emirates, which connects to FLAG (Fiber-Optical Link Around the Globe), Asia-Europe Optical Fiber, which goes from the Azerbaijan of Iran to Turkmenistan, Georgia and Republic of Azerbaijan, the microwave connected to Turkey, Azerbaijan, Afghanistan, Turkmenistan, Syria, Kuwait, Tajikistan and Uzbekistan and finally the Satellite Stations 13 (Intelsat 9 and Inmarset 4) [7].

Access to the internet is provided in compliance with instructions and regulations compiled and communicated by the Communication Regulatory Commission (CRA).

In November 2003, Iran signed the Declaration of Information Society Principles in Geneva; and since 2004 Iran has started the Rural ICT Project determining executor and executive instructions. Until the date of this report, according to the statistics of the official Web site of Telecommunications Company of Iran (TCI), 10,000 villages have been equipped with Rural ICT [8].

According to the statistics of ITU, there have been 27,914,700 internet users in 2009. This well indicates that the internet and its various applications, which actually constitute the virtual space basis and Web space, were widely welcomed by the Iranian people. Internet national network is underway as a national project in the Ministry of ICT. The idea behind the project is to develop the domestic internet network, which provides not only specified access points to the World Wide Web in essential communications but also data traffic in the domestic use centers at high speed and without any need to pass through the data centers outside the country, which, in turn, requires domestic data centers (Riazi, 2008).

Iranian Web

As stated earlier, Web content, which stands for the resources and pages existing in the internet and being accessible through search engines, can be studied from different views making use of different methods and devices. The content studies are among the most important research approaches in the Web , which may be conducted with various research objectives, and categorized based on different models.

A plethora of Iranian Web content, for instance, is constituted by Web logs, which make use of non-Iranian and even Iranian blog services with a non ".ir" top-level domain. This range of Iranian Web content, which is mainly in Persian, reflects a plethora of Iranian thoughts, lifestyles and views on the Web and the internet. The research project titled "the Internet and Democracy" conducted by Kelly and Elthing (2008) in Berkman Internet and Society Center affiliated with the University of Harvard, gave a sociological but criticized outlook on an important part of Iranian Web , "Persian blogsphere". The research deals with Persian blogsphere with a combination of survey, computational, social network and content analysis making use of computer and human and obtained the following results: Persian blogsphere is a colossal public space for individual view exchange and discussion. It includes 60,000 active Web log, which are regularly updated and have four main network hubs. Each hub has certain subsets. These hubs are as follows: 1. Secular/Reformist, 2. Religious Conservative, 3. Persian Literature and Poem, 4. Miscellaneous networks. The filtering of the blogs by the government is well below what was first supposed by the researchers and major part of blogsphere is accessible (Kelly and Etling, 2008).

As a complement to the above-mentioned research, the writers provided the changes made in the Persian blogsphere and declared (not with certainty but in doubt) that, one year after the research, the space belonging to the religious conservative section has a very quick growth possibly thanks to the Web logs developed by the Basijis. It is of course obvious that there are some signs of the influence of Shiite reflection and beliefs on the people outside Iran in the virtual space (Kelly & Etling, 2009).

Another approach is to evaluate parts of the Web making use of search engines. Here, the following points are noteworthy:

- The search engines only index publicly available Web sites
- The search engines only index the "visible Web " and will not index dynamic pages, many proprietary file formats, etc.
- The search engines may duplicate findings, by not be able to differentiate between identical pages
- Results from search engines can fluctuate due to the load on the search engines (Day, 2003).

Therefore, part of the Iranian Web pages with the Domain Name ".ir" was analyzed to find answers to the following questions:

1. In which domain names under ".ir", more pages are indexed by Google or Yahoo Search Engines?

2. Which file formats are used in the Web pages of ".ir" domain and its subsidiary domains?

3. How is the dispersion of all formats in each and every domain name?

The method used in this section is Web ometrics, and its tools are two search engines of Yahoo and Google, which, according to Alexa Site [15], were most widely viewed search engines at the time of the research. The next ranks belong to Windows Live and MSN. It is noteworthy that the search results for the last two items are far lower compared to that of the first two search engines. Here, the recall in retrieval from the view of the pages' maximized cover is intended. The first two search engines were considered enough. In the Google search engine, the search term formats "site:.[domain name]/ and –site:.ac.ir/ site:.ir/ and site:.org.ir/…" are used for the number of pages indexed under each domain name and the top-level name ".ir" respectively.

In the Yahoo search engine, the search term formats "domain:. [domain name]/ and domain:.ir/ NOT (domain:.ac.ir/ OR domain:.org.ir/ OR domain:.gov.ir/ OR…)" are used for the number of pages indexed under each domain name and the domain name ".ir" respectively. The results are shown in Table 1.

Table 1. The pages indexed under each domain name in Google and Yahoo

No.	Domain Name	Google	Yahoo
1	.ir/	23,300,000	23,103,052
2	.ac.ir/	1,470,000	1,590,075
3	.org.ir	140,000	168,012
4	.gov.ir/	2,840,000	378,003
5	.co.ir	73,100	109,000
6	.sch.ir	11,500	18,300
7	.id.ir	1600	13,900
8	.net.ir	1280	1510
Total		27,837,480	27,181,417

Although the results do not show considerable differences between the search engines of Google and Yahoo - regardless of the domain name of ."ir", which naturally appropriate the highest number-pages appear in largest numbers in the domains ".ac.ir/" and ".gov.ir/" but not in the domain names ".id.ir" and ".net.ir".

Only the Google search engine with the search term format "site: [domain name]/filetype:[html….]" was used to extract the type of file format being used in the ."ir" pages and its subsidiary domains; and the search term format "site:.ir/filetype:[html…]-site:.ac.ir/ -site:.gov.ir…" is used for the domain of ."ir".

The results are summarized in Table 2.

Table 2: File formats used in the Web pages of ".ir" domain name

File types	/ir.	/id.ir.	/sch.ir.	/net.ir	/gov.ir.	/org.ir.	/co.ir.	/ac.ir.

HTM	1,040,000	99	760	42	13,200	3160	4560	60,300
HTML	2,450,000	1240	196	787	56,200	10,900	17,800	13,6000
ASPX	11,200,000	24	4190	314	627,000	18,600	9190	310,000
PHP	3,840,000	1080	3180	71	66,200	66,100	12,400	487,000
PDF	187,000	6	240	75	4850	4760	3380	110,000
ASP	2,900,000	0	974	117	10,500	10,600	1720	132,000
SWF	19,600	9	1260	7	107	104	443	1870
DOC	17,100	4	455	115	469	460	458	17,400
SHTML	4780	0	0	59	0	0	7	19
PPT	2640	0	3	6	133	131	138	4710
JSP	12,200	0	143	0	239	235	15	4190
XLS	1920	0	3	4	422	402	105	1360
XML	43,400	0	26	1	152	152	65	365
TXT	32,600	0	2	0	12,200	14,000	34	387
PS	116	0	0	0	0	0	0	361
CFM	747	0	0	0	2	2	129	276
PPS	654	0	8	0	22	22	43	386
MHT	407	0	1	0	4	4	8	172
CGI	6730	0	104	0	864	865	109	210
PHTML	10,200	0	0	0	0	0	11	294
RTF	406	0	0	0	0	0	0	232
PL	3480	831	0	0	0	0	3	470
DAT	384	0	0	0	0	0	2	105
VBS	2	0	0	0	0	0	0	1

SHTM	5	0	0	0	0	0	0	4
XSLT	7	0	0	0	0	0	0	2
PHP3	657	0	0	0	0	0	9	6
SQL	28	0	0	0	0	0	0	7
XHTML	295	0	0	0	0	0	0	0
EPS	9	0	0	0	0	0	0	263

File formats asp, html, htm, php and aspx have been used more than other formats in the Iranian Web pages. Regarding what was stated above, the results shown in the tables somewhat help describe the Iranian Web in the domain of ."ir", but they should be cautiously used because of the problems peculiar to this type of research from reliability viewpoint. The most conspicuous flaw arises from the instability of the Web space. Notwithstanding these problems, Webometric is recognized as a method for studying the Web ; yet the point is, the results should not be considered final and unchangeable, and should be used proportional to the objectives of the research to be conducted.

Legal Aspects: Author's Rights and Legal Deposit in Iran

The bill of support for the rights of authors, writers and artists was ratified by National Consultative Assembly on Jan.1, 1970 in Iran. The second independent law in this regard was the "Act for Translation and Reproduction of Books, Publications and Audio Works", which was ratified by the National Consultative Assembly in January 1974. As the laws related to intellectual ownership lacked the perspicuity and comprehensiveness required for software products, therefore, in 2000, considering the requirements of the day and development of electronic media, the "Act of Support for Computerized Software Packages" was ratified by the Islamic Consultative Assembly. With the Act ratified, the national support for intellectual ownership developed in Iran. Nowadays, it almost applies to all necessary fields. Meanwhile, the draft of the "Iranian Comprehensive Act of Copyright and Related Rights" has been prepared and is supposed to be presented by the Cabinet as a bill to the Islamic Consultative Assembly to be changed into an act [11]. The said draft was prepared in accordance with Sample WIPO Act, in the light of the joining of Iran to WTO and considering the necessity for coordination with the laws of copyright in other countries.

In Iran, the "Iranian Legal Deposit Law" ratified by High Council of Cultural Revolution on 1989 requires publishers to deliver some free copies of their publications to the State Depository Libraries. Another ratification of this council, dated May 11, 1999, Session 441, requires all governmental and nongovernmental producers of non book materials to deliver two free copies of their works to the National Library and receive depository number. According to this ratification, non book materials are:

"Electronic publications: electronic books, illustrated electronic books, both motionless and moving, talking books, etc" (Act of non book materials legal deposit, 1999). Yet, deposit of internet resources has not been directly mentioned in this act.

The Status of ICT Infrastructure in the NLAI

Telecommunication infrastructure of the NLAI is optical fibers between the building of the library and the relevant telecommunications center. For outside communications, the access to the internet with the bandwidth of 10 megabytes/second as well as the National Internet Network with the bandwidth of 50 megabytes/second has been provided so that the users may receive their required information from the site of the NLAI and the existing portals at highest speed.

At present, RASA and Digital Library, are two main software packages used for the management of the library as well as management of digital content. Before presenting a brief overview of the stated software packages, it should be reminded that there are generally four main parts in a Web archiving

software as below:

1. Crawler: for collecting data

2. Preparation and archive of data: including classification, filter, addition of metadata and other relative tasks

3. Information representation and search: interaction with user for meeting information requirements

4. Information processing for establishment of value added for the data collected, specially by data mining procedures

With this in mind, the general architecture of a Web archiving software package is discussed for both software packages.

1. RASA Software: At present, the software is used as one of the main tools for providing services in National Library. Its conspicuous specification is the comprehensive implementation of IRANMARC.

2. Digital Library Software: Digital Library is actually a software package used for non-Web digital items. Although the interaction between Web archiving software, RASA, Digital Library Software Packages and especially metadata is possible, from software engineering viewpoint, that a new software package is used as an extension of an old software package is not acceptable. The most important reason is the advances made from the view of implementation method since the time of development of old software packages that, from economic viewpoint, make continuation of software development by old methods and tools irrational.

Metadata and Standards

Rasa Software supports IRANMARC, which is based on UNIMARC. IRANMARC was developed for storage, retrieval and exchange of data based on ISBDs, taking into consideration the specifications of Persian language cataloguing. IRANMARC provides four databases: bibliography, authorities, holdings and classification. At present, authorities and holdings databases are in use in RASA Software; and National IRANMARC Committee has considered the use of classification database one of the next priorities.

Meanwhile, RASA system includes over 50 type library of materials such as papers, serials, images, stamps, diagrams, flashcards, logos, flags, etc. One of the objectives of cataloguing is to determine the bibliographic relation between the work under cataloguing process and other relevant works. Establishment of this relation is made possible using Fields of Block 4. At present, some of the fields of Block 4 are used to demonstrate such a relation in the system. Any country using one of the MARC Standards (MARC 21 or UNIMARC), can customize it considering its national requirements, considering that Permanent UNIMARC Committee has paved the way for its users to make use of number "9", e.g. -9, -9-, --9 as well as subsidiary fields for their national and local needs. Meanwhile, National IRANMARC Committee has added some items to UNIMARC in bibliography, authorities and holdings databases within the vested powers (Akbari Daryan; Yaghoub Pour Nargesi, 2008).

Authority control is possible in RASA, that is to say, the system of referrals has been used in the software. In the metadata of each bibliographic record of the National Library, it is necessary to establish a link to Iran Web archive. This way, the Web archive resources accompanied by other resources are recovered. In the following, some special fields that should exist in Web resource worksheet are described. It should be noted that the Field 856 can only establish the link to Iran Web archive resources but cannot keep and archive them (Table 3).

Table 3: Special Fields of Website Worksheet in RASA

.No	Tag	Field Label
1	135	:Fields of coded data
2	$200b	General Name of Materials

3	210	Web Resource Development Status
4	330	Introduction of Web Resources
5	337	Record of technical details
6	856	Electronic Location and Access

RASA System has already implemented Z39.50, Version 2 and is able to exchange data with many libraries, which have implemented the standard.

Results and Recommendations

Considering the whole findings in this regard, development of Iran Web archive in the NLAI is feasible because:

- There is a content rather considerable, but in need of further meticulous appraisals
- The growing technical, specialized and telecommunication infrastructures of the country confirm such facilities notwithstanding the following difficulties and constraints.
- The proposals of the study for development of Iran Web archive in different respects are as follows:

1. Large-scale management: Formation of Strategic Management Council

Considering that the Web archiving is a large-scale and national project, it should undergo a weighty and dynamic managerial supervision, which can lead the project in all technologic, research and executive respects. Development of such a strategic management council for Iran Web archive is therefore proposed for content selection, technical and legal policy makings as well as supervision over planning and execution of project within the framework of the abovementioned policy makings. In the light of national scope of the project, the presence of representatives from governmental bodies involved in production, preservation and promotion of content and the affairs related to intellectual ownership (especially in the Web space) is recommended in the said council. The council can facilitate the national and international communications and cooperation. The paper specifically proposes the presence of the following representatives from governmental bodies in addition to the executor(s) and managers of the project in the said Strategic Management Council:

1. High Council of Information Technology

2. Ministry of ICT

3. High Council of Informatics

4. Ministry of Culture and Islamic Guidance

5. Super Council of ICT

6. Faculty member(s) of the universities in the specialized fields of Librarianship and Information, Information Technology, communications and cultural studies

Proposed Approaches for Collecting Iranian Web Content

Initiatives of Web archiving in the world have adopted different methods in harvesting Web content. The harvesting approaches ranges from the most idealistic (mixed approaches such as automatic harvesting with crawlers in the national scope as well as thematic, selective and preferential crawls, deposit etc) to limited scopes, such as thematic selection.

As we mentioned above, the Strategic Management Council is the best authority to make decision on the content to be collected. In case of organizing the Council, it may conduct some studies, measurements and discussion to propose, from the viewpoint of kind and subjective theme, a special

yet limited spectrum of Web sites for a pilot phase.

Considering the objectives and responsibilities of the NLAI in the national level, it is appropriate, as far as possible, not to omit any Web resources that, from content viewpoint, can be included in the framework of these objectives and responsibilities. On the other hand, regarding two phases considered for the implementation, it is proposed, at the pilot phase, to harvest and archive non governmental Iranian News Agencies Web sites that do not have any print formats. These resources, though limited from the quantitative view, enjoy multi-media content, diversity of the file formats used, interactive characteristics and dynamic pages application and are seemed to be appropriate for testing the used software system. At the main phase, Iran Web archive should be able to commission the archiving without any data volume limitation. This study, therefore, proposes the mixed approach for the main phase.

3. Cooperation

In the Iran Web archive, cooperation is proposed in several levels:

a) General policymaking: at this level, cooperator institutes and organizations proposed for participation in the Strategic Management Council were stated in 1.

b) Research: in addition to all governmental and non-governmental scientific and research institutes and bodies, which may intend to accept the projects proposed by the NLAI in this regard or support some studies on their own in this field; the present study proposes the use of technical and scientific ability of the Research Institute for ICT (ITRC), especially the Information Technology faculty of this institute (thanks to its special position in Ministry of ICT) for qualitatively supervising and executing the research projects in this field.

c) Development and implementation: Development of a Web resources archive not only requires technical and telecommunication facilities in the NLAI but also needs a connection with broad bandwidth and without filtering; and hence, this project is not feasible without the cooperation of Zirsakht Company. As for the problems related to the cyber space information exchange security, cooperation with Information Technology Company will be necessary. Of course, considering the changes in activity scopes and legal responsibilities of the companies, the occurrence of changes will be feasible.

Participation in the international information exchange space requires certain technological preparations and global standards. Therefore, cooperation and/or establishment of relations with International Internet Protection Consortium (IIPC) as the most important international institute in this regard and exchange of experiences with its members will appropriately reflect the scientific and specialized activities of the country in the world. Therefore, contingent upon the priority of customization, and in compliance with the framework of the state general policies, cooperation with the consortium may be useful.

Ministry of Culture and Islamic Guidance's Project for Website Registration (Samandehi) may be used for collecting Web sites as the case may be.

4. Proposition on Legal Deposit Law

The Iranian "Non book Materials Deposit Act" does not address online electronic publications. It is proposed that online electronic publications be added to the appropriate clause of the Act.

5. Metadata and standards

As none of the metadata standards, according to special functions, are not solely able to meet special native and local requirements, the study first proposes MODS, MADS, METS, RDA, DOI, MIX, TextMD, PREMIS, OAI-PHM, SRU/W, Z39.50 and MARC as standards appropriate for Iran Web archive. The study is also of the opinion that it is necessary to incorporate a permanent committee for investigation and customization of metadata standards in the NLAI. It also proposes that National Digital Object Identifier System should be implemented in the National Library. Archive and retrieval of the collection levels should be conducted in RASA and IRANMARC, in such a way that the records of Iran Web archive may be browsed in the collection level next to other librarian resources. And with link between URL and Iran Web archive holdings being established, user accesses every single record of the collection.

5. Web archiving system software

Considering four main parts of Web archiving, that is, harvesting (crawling), preparation and archiving, providing information and browsing, information processing as mentioned earlier ,the specifications of each and every said component are proposed as below:

a) The crawler of the system should have the following capabilities: widespread, focused, continuous and deep crawling; supporting No-Robot Standard; controlling crawling depth; duplicate detecting; providing information in file formats capable of being indexed; capable of receiving information from protected servers; enjoying modern programming language and modular design; supporting API; type of license (open or limited text), supporting multithread function; maximum automation and minimum human intervention; capable of gathering metadata; high efficiency; fault tolerance; configurability, scalability and flexibility; maintaining records; user friendly interface; maintaining numerous copies; archiving all contents of Web sites.

b) The crawler should be selected from among the open-source software packages. Following the real tests conducted, the study gives the first priority to Hertrix Crawler for evaluating the efficiency of this system.

c) Weka Multipurpose System may be used in order to prepare the information for archiving, which includes quality-evaluation, automatic classification (which facilitates user access to their required information) and metadata addition processes. At present, there is no software that can be entrusted with the whole quality evaluation process in an automatic manner.

d) The system should support core services including distributed index, system file management at high volume storage, security at the program, operating system, network … levels, recovery and system design in a crash-safe manner, load equilibrium, damage management and appropriate data provision.

e) Either Lucene or Lemur Systems are proposed for Web archiving system. Considering the open-source nature of these two systems and regarding the fact that the licenses of these software packages are AS Is, the quality and quantity of their capabilities should be validated in comparison with real data. As these two systems differ from certain substantial facilities such as distributed information retrieval, possibility of implementation on hardware cluster, XML retrieval etc, system selection will be feasible based on final facilities appraisal.

7. Hardware and telecommunication requirements: As for the architecture considered for the required software packages, the following servers are required:

a) Processing resources required for crawlers

Table 4: Specifications of the processing resources for crawlers and preparation of data

Services			Crawlers and Preparation of Data						
No	CPU	Memory	Load Balance/Cluster	Storage			Server Type		
				HDD	SAN (HBA)	SAN (iSCSI)	VM	Blade	Dedicated
2	2*Quad 3000 MHz	32 GB	ü	û	û	500 GB	û	û	ü

b) Processing resources for searching the indexed data

Table 5: Specifications of processing resources for managing data index

Service			Management of Data Index		

No	CPU	Memory	Load Balance/Cluster	Storage			Server Type		
				HDD	SAN (HBA)	SAN (iSCSI)	VM	Blade	Dedicated
2	2*Quad 3000 MHz	32 GB	ü	û	û	500 GB	û	û	ü

Table 6: Specifications of processing resources for database of the archive

Service			Database of Web Archive						
No	CPU	Memory	Load Balance/Cluster	Storage			Server Type		
				HDD	SAN (HBA)	SAN (iSCSI)	VM	Blade	Dedicated
3	2*6-Core 3000 MHz	32 GB	ü	û	2TB	û	û	û	ü

c) Processing resources for a Web based search portal for providing users with services

Table 7: Specifications of processing resources for search portal

Service			Search Portal						
No	CPU	Memory	Load Balance/Cluster	Storage			Server Type		
				HDD	SAN (HBA)	SAN (iSCSI)	VM	Blade	Dedicated
2	2* Quad 3000 MHz	32 GB	ü	200 GB	û	û	û	û	ü

Regarding the approximate estimation of number of sites existing in the ".ir" domain as mentioned earlier and the estimation of the volume of the information in every single database, the memory required for archiving such information is proposed to be two terabytes for archiving the contents produced during a year. Therefore, at least, a 10-terabyte memory is required for archiving the contents produced during 4 years. At present, there is EMC CX3-20C in the datacenter of the national library; and considering its extensibility, an enclosure with 15 hard discs, fiber channel, each of which has one terabyte memory for archiving the gathered information.

Apart from the online storage resource, a tape library is required for maintaining archived data such as backups. A tape library with Adic iScalar 500 and the capacity of 36 LTO4 Tape, 30 terabytes storage capacity, should be prepared. The existence of a SAN System is necessary for the archive.

In order to harvest and update information, it is necessary to allocate an appropriate internet bandwidth (at least 10 megabytes/sec) for this service.

Considering the exertion of filtering by Iran Telecommunications Company, a proprietary internet line without filtering should be allocated to the NLAI. This requires coordination with Iran Telecommunications Company.

The archive's backup copy should be kept in a place other than the NLAI in order to maintain the security. Selection of such a place and the institute in charge of it requires some all-inclusive studies by the Strategic Management Council. In order to provide on-line backup, mirror of all services and a powerful datacenter are required that, in case of causing problem, provide similar services. This study proposes establishment and commissioning of datacenter of the NLAI in Shiraz, Capital of Fars Province.

8. Budget and Personnel

In this project, the budget estimate levels consist of the expenses of technical equipments amount to Rls. 2 billion (approximately US$ 200000), expenses of personnel that will not include extra expenses by relying on the present personnel, and the most important part of the project is outsourcing and development of the software of Iran Web archive, which may not be estimated at present due to quick changes in price and necessity of public tender.

9. Human Resources

Notwithstanding the fact that the present employees may be used for Iran Web archive, some expert employees are required according to Table 8.

Table 8: Personnel and specializations required

No.	Field, Scientific Degree and Specialty	Number
1.	LIS/ MA or PhD: Expert in digital resources cataloguing, identification and selection of resources	2
2.	Software engineering/ Bachelor or MA	2
3.	LIS/ Member of the faculty, Expert in metadata standards, supervising quality control of descriptive metadata standards	2
4.	Coordinating affairs and secretariat	1
	Total	7

10. Execution Phasing

The study proposes the following phases for execution following the feasibility study:

1. Establishment of Strategic Management Council of Iran Web Archive

2. Compilation of RFP for archiving system

3. Organizing of public tender in order to develop the Iran Web Archive System

4. Operation of trial version

5. Debugging, receiving feedbacks and revision

6. Operation of the principal version

Notes

1. http://www.nlai.ir/Default.aspx?tabid=98

2. It should be noted that, in order to develop collections of digital resources, Digital Library Project was implemented in NLAI in 2008, in the trial phase of which a selected collection of digital but non-Web resources were made available to the public via the NLAI's Web site. This paper is in fact the first phase of "Iran Web Archive Project".

3. The name of this center was changed into "Institute for Research in Fundamental Sciences" in 1997.

4. http://www.ipm.ac.ir.fa/network

5. http://www.neda.net/nsit/home_about.asp

6. http://www.tic.ir/Web /content/621

7. http://www.cia.gov/library/publications/the-worldfactbook/docs/profileguide.html

8. http://tci.ir/s40/P_1.aspx?lang=fa

9. http://www.itu.int/ITUD/ICTEYE/Indicators/Indicators.aspx

10. www.alexa.com/topsites

11. http://www.scict.ir

12. http://rc.majlis.ir/fa/law/show/100644

References

1. Act of non book materials legal deposit, 1999.

2. Akbari Dariyan, Saeideh; Yaghoub Pour Nargesi, Tahereh. 2008. Criticism of IRANMARC and where it is heading for; Study of IRANMARC in three stages; *Faslname-ye Ketab* , vol. 75 (Autumn), pp. 321-329.

3. Berners- Lee, T. 1998. *Cool URLs don't change.* Retrieved Nov., 23, 2009 from http://www.w3.org/Provider/Style/URI

4. Day, Michael. 2003*. Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Welcome Trust*. version 1.0, 25 February. Retrieved July12, 2009 from http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

5*. Definition of Digital Preservation.* Retrieved April 22, 2010 from: http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm

6. Kelly, John; Etling, Bruce. 2008. *Mapping Iran's Online Public: Politics and Culture in Persian Blogsphere*. Retrieved Mars 3, 2010 from http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Mapping_Irans_Online_Public_Persian.pdf

7. Kelly, John; Etling, Bruce. 2009. *Mapping change in Iranian Blogsphere*. Retrieved July 13, 2010 from: http://blogs.law.harvard.edu/idblog/2009/02/12/mapping-change-in-the-iranian-blogsphere/

8. Rahimi, Babak. 2003. Cyberdissent: the Internet in Revolutionary Iran. *Midddle East Review of International Affaires*, Vol. 17, No. 3 (Sept.). Retrieved March, 15, 2010 from: http://meria.idc.ac.il/journal/2003/issue3/rahimi.pdf

9. Riazi, Abdolmajid. 2008. *National Internet Network: the Project for compilation of the State Integrated Information Technology*. Retrieved March 23, 2010 from: http://www.matma.ir/matma/images/files/INN_Summary_2.pdf