Distr.
LIMITED
E/ESCWA/ICTD/2003/WG.1/CRP.29
3 February 2003
ORIGINAL: ENGLISH

ESCWA

Ministry of Telecommunications

infoDev

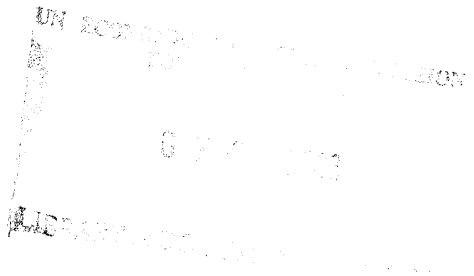
UNESCO

UN ICT Task Force

ITU

ECONOMIC AND SOCIAL COMMISSION FOR WESTERN ASIA

Western Asia Preparatory Conference for the World
Summit on the Information Society (WSIS)
Beirut, 4-6 February 2003



STANDARDIZATION RELATED TO ARABIC LANGUAGE USE IN ICT

Hassan Diab
Department of Electrical and Computer Engineering
Faculty of Engineering and Architecture
American University of Beirut
Beirut, Lebanon

03-0119

The information provided in this paper is the result of the work carried out [1] at Information and Communications Technologies Division (ICTD), Economic and Social Commission for Western Asia (ESCWA), Beirut, Lebanon during September 2002. The Internet is quickly emerging as a new defining line in society. Access to information transmitted electronically and to the growing online marketplace available through the Internet will be essential for the economic development of communities, cities, regions, and nations. The Arabic language population constitute over 5% of the non-English population in the world, yet the corresponding percentage of Internet users is slightly over 1%. This is highly attributed to the absence of unified ICT standards related to Arabic language use in information society applications.

One of the major problems that face the use of Arabic on the Internet with respect to standardization is either the diversity of standards for specific areas (e.g. multiplicity of character sets) or the complete absence of these standards in other areas. This study addresses the issues pertaining to ICT standards related to Arabic language use in information society applications. There is an urgent need for formulating and adopting standards for the use of Arabic language in ICT. Thus, the importance of Arabization on the Internet and, therefore, the needed ICT standardization is discussed.

Accordingly, existing ICT standards related to the Arabic language as well as the Arab and International organizations concerned are surveyed. Identification of Arabic language standardization issues are highlighted such as character encoding, e-mail, Web browsing, speech recognition/synthesis, as well as search and indexing on the Internet. Furthermore, legislative and regulatory issues for information and knowledge management with respect to the use of IT terminology and Internet security are also considered.

The display features of Arabic text necessitate specialized text-display algorithms. One of the most important server-processing issues for Arabic text is the problem of searching and indexing. Solutions have started to emerge with browsers and mail programs building on new Internet standards such as Multipurpose Internet Mail Extensions (MIME) and HTML 4.0. The trend towards Unicode also helps to standardize the exchange of Arabic on the Internet. Issues pertaining to character sets, bi-directional displays, Arabic e-mail, Arabic Web browsing, as well as searching and indexing of Arabic text on the Internet are also discussed. Finally, an action plan is proposed to fill this obvious gap that is inhibiting the efficient utilization of Arabic on the Internet.

INTRODUCTION

Despite the obstacles faced, Arabic is being increasingly used on the Internet. However, a major impediment facing the use of Arabic on the Internet is the lack of unified standards, particularly in the field of character sets. Other obstacles (1997 survey, carried out in Saudi Arabia) include: weak telecom infrastructure, lack of Arabic content on the Internet, and lack of Arabic Internet access programs for the Web and for e-mail.

The support required for Arabic on the Internet can be categorized in the fields of:

- content, Arabic textual content relates to representing the data itself and to formatting it through Internet standards such as HTML in the case of the WWW pages and RFC 822 and MIME in the case of e-mail messages.
- transport, The Transport protocol is HTTP for the Web and SMTP for e-mail.
- client processing, includes generating, displaying, & interacting with Arabic text
- server processing, storing, processing, searching, & providing Arabic content

One of the major problems that face the use of Arabic is transporting Arabic text over the Internet due to the multiplicity of character sets. The display features of Arabic text set it apart from other languages in several ways:

- Arabic text is cursive, and the shapes of its characters depend on their position in the word.
- The directionality of Arabic text is peculiar: While Arabic text is written right-to-left, Arabic numbers are written left-to-right.
- One of the most important server processing issues for Arabic text is the problem of search and indexing.

Figure 1. Number of online people in each language zone

	Internet access (M)	% World online pop	2003 (est./M)	Total pop.1 (M)	GDP2 ($B)	% of world economy
English[3]	228	40.2%	270	567	$13,812	33.4%
Non-English	339	59.8%	510	5633	$27,590	66.6%
Arabic[4]	4.45	0.8%	6	3006	$678	1.6%
Total European Languages (non-English)	192.3	33.9%	259.3	1,218	$12,550	30.3%
Total Asian Languages	146.2	26.1%	254	1,658		
WORLD TOTAL	560		762	6,2007	$41,400	

[1] http://global-reach.biz/globstats/refs.php3#languages

[2] http://global-reach.biz/globstats/refs.php3#gdp

[3] http://www.glreach.com/gbc/en/english.php3

[4] http://www.glreach.com/gbc/afr-mideast/arabic.php3

[5] http://global-reach.biz/globstats/refs.php3#20

[6] http://www.forbes.com/global/2002/0401/027.html

[7] http://www.popexpo.net/eMain.html

Global Internet statistics8 can be found categorized by language. Figure 1 shows the latest estimated figures of the number of people online in each language zone (native speakers), which classifies by languages instead of by countries, since people speaking the same language form their own online community.

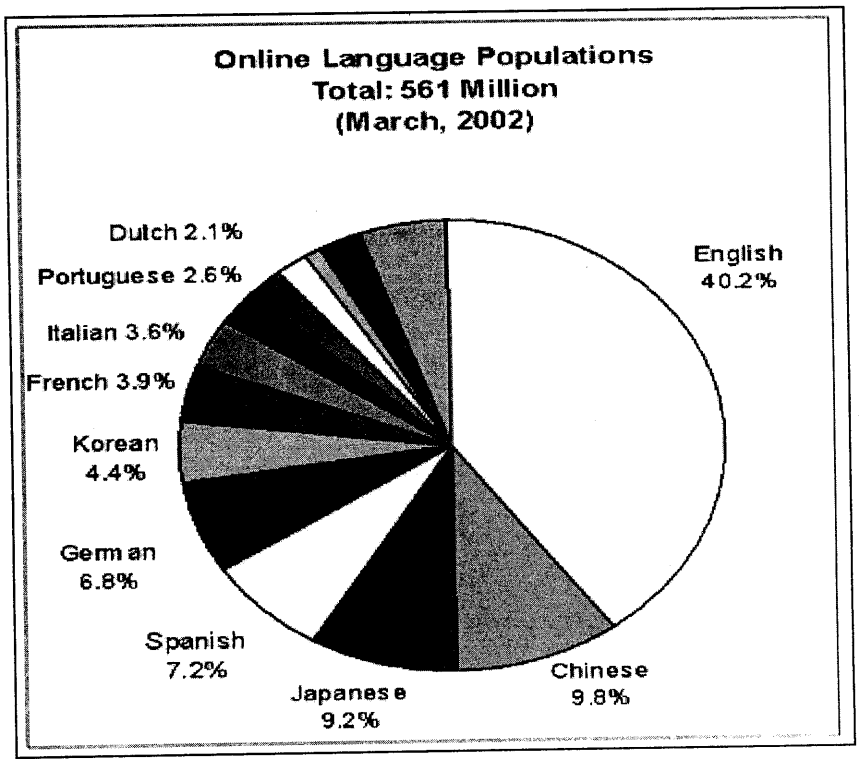
As can be seen from Figure 1, the estimated number of Arabic language Internet users is 2.4% of total Asian lanuages users, 1.2% of total non-English users, and 0.8% of total world users. However, the Arabic language population constitute 18.1% of total Asian languages population, 5.3% of total non-English languages population, and 4.8% of total world population (Figure 2).

Figure 2. Percentage of Arabic language online population and Internet users relative to that of Asian, non-English, and the world

	% Arab Language		
Language	Asian	N	World
% Internet Use	2.4%		0.8%
% Population	18.1%		4.8%

The following figure shows the distribution of the online language population totaling 561 million (March 2002 est.).

Figure 3. Distribution of the online language population



The English language has an important presence on the WWW. Current estimates of language occurrence indicate that approximately 70-80% of existing Web pages are written in English. There are a number of reasons for the dominant position of the English language online, the most important of which is

[8] http://www.glreach.com/globstats/index.php3

the high percentage (around 55% in 1998 according to Computer Industry Almanac) of Internet users that are based in the United States.

The number of users on line who speak a particular language is of crucial importance to the amount of content available in that language. The more content there is online in a language, the more reasons speakers of that language will have to go on line. Some of these new users will produce their own content, increasing the overall content available in that language, and so on in a form of a positive feedback loop.

Indeed, whilst it is true that the English language is predominant on the World Wide Web, there is evidence that the number of non-English speaking Internet users is rising steadily as penetration rates in non-English speaking countries continue to rise. According to Computer Industry Almanac9, the number of Internet users surpassed 530 million in 2001 and will continue to grow strongly in the next few years. Most of the growth is coming from Asia, Latin America and parts of Europe. By year-end 2005 the number of worldwide Internet users will exceed 1 Billion. An increasing portion of Internet users will be using wireless devices such as web-enabled cell phones and PDAs to go online. The wireless devices will be supplemental to PC Internet access for most users in developed countries. In countries with low Internet penetration many wireless Internet devices will be the primary Internet access devices.

I. ARABIZATION OF THE INTERNET

Arabic is the mother tongue of over 300 million people in 22 Arab countries. While many Arabs use English or French as their preferred language on the Internet, the majority of Arabs, particularly in Saudi Arabia, Egypt, United Arab Emirates, Kuwait, Bahrain, Qatar, Oman, and Syria use Arabic. If the Arab world is to be a knowledge based society in which all of its organizations and all of its peoples can participate then it is essential that as much as possible of the digital communication network can be accessed and used via the Arabic language [2]. In this respect, bridging the digital divide is progressing on three fronts. Firstly there is the Arabization of the Internet. Work has started via ICANN which is partly responsible for the regulations, that govern business through the Internet and which controls domain names, and grants licenses to the companies allowed to register domain names. The AKMS has established the Arab Internet and Domains Association (AIDNA) with the objectives of entering the knowledge world and developing the Arabic Internet, consolidating the relationship and cooperation among holders of electronic addresses and domain names, and Arabization of the Internet and dissemination of Arab culture. The Arab Club for Information (Arabcin) also has an important role here in getting those in the Arab world who have an interest in these developments to share information and to build up a unified data bank.

Secondly there is a need to stimulate the attention and interest of governments in creating the environment in which a knowledge society is able to function. The Arab Digital Economy Initiative, and the Muscat Declaration are both designed to get Arab governments to work together in order to create this right environment, to foster innovation, and to develop business, technical and legal programs that will achieve this.

The third front is that of publications. Making use of the best technology requires that other basic material must be available and that it should be in Arabic. There were, and still are, major gaps in the Arabic material available [2].

The idea of developing multilingual domain names arises from the need to eliminate the last language barrier that prevents non-native English speaking population from using the Internet. By allowing the intuitive use of native languages of each country, the Internet will be more accessible. As a result, more people will be able to enjoy the Internet experience and be exposed to the variety of opportunities it offers. The Arab Internet community will no longer be forced to create strange domain names in English with a

[9] http://www.c-i-a.com/

meaning in Arabic and they will no longer have to improvise or guess the spelling of the domain name in English. Moreover, existing sites can be given new names in Arabic, with no alteration required.

ICT includes the specification, design and development of systems and tools dealing with the capture, representation, processing, security, transfer, interchange, presentation, management, organization, storage and retrieval of information. Each language has its particularities which consequently have to be taken into consideration in standards. For the Arabic language, there is a technical committee (TC-8) belonging to AIDMO-CSM. Examples of ICT standards that should be addressed:

- Input of information: Keyboard layout, character-set, phonemes set, etc.
- Processing of information: compression standards of text and of speech, Natural Language Processing (NLP) algorithms, control codes, etc.
- Transfer of information over networks, such as the Internet: standard transfer protocols, standard Markup Languages, etc.
- Output: displayed or printed character sets, page formatting, etc.
- Application software standards: e-commerce, e-documentation, e-publishing...
- Terminology standards.
- Standards for testing procedures to assure and control the quality of software.
- Guides for unification of regulations for issuing conformity to standards certificates. Unification of procedures for accreditation/certification of testing.

Standardization is an important component of the science & technology agenda of any nation. Standards for Arabic in ICT are particularly important especially that the ICT sector in the Arab world is rapidly growing, and expected to form an essential part of its economy. This sector has several particularities which represent a challenge and an opportunity to the Arab world:

- Economically, ICT is considered one of the largest sectors in the world with high value added. Its growth rate is higher than the average growth rate of industry as a whole.
- ICT has penetrated homes, schools, offices and factories, thus it does not only have an economical impact but also cultural and social ones.
- The Arab software industry could have a comparative advantage in applications related to the Arabic language or Arabic culture.

Consequently, standards for Arabic in ICT have distinctive criteria compared to standards in other sectors:

- The absence of regionally unified Arabic standards will result in ICT systems in the Arab world that cannot communicate between each other easily and efficiently.
- The absence of regionally unified Arabic standards in ICT will potentially result in 22 fragmented ICT markets. Whereas the existence of these standards will foster ICT industry in the Arab world.
- The absence of regional Arabic standards in ICT could result in lower quality software products.
- Standards for Arabic in ICT, if produced or approved by Arabic regional NSBs (AIDMO-CSM) on time, will assure the respect of the Arabic language peculiarities and avoid unsuitable de facto standards imposed by certain companies.
- When Arabic national or regional standardization bodies deal with standards in ICT, they are faced with a much more difficult job, compared to formulating standards in other sectors. To formulate ICT standards, they have to do special studies involving linguists, computer scientists,

electronics engineers, standardization specialists, etc. It is not possible to translate or adapt international standards for this sector, as is the case usually in other sectors.

- An important role to be played by NSBs in the field of IT is to enforce, as they do in other sectors, certain standards that assure issues such as: transparency, portability, efficiency, bilingualism and respect of the specifics pertaining to the Arabic language.

Some of the existing arab language standards include:

- BS 4280:1969(1983): Transliteration of Arabic Characters.
- ISO 233:1984: Documentation - Transliteration of Arabic characters into Latin characters.
- ISO 233-2:1993: Information and documentation - Transliteration of Arabic characters into Latin characters - Part 2: Arabic language - Simplified transliteration.
- ISO 233-3:1999: Information and documentation - Transliteration of Arabic characters into Latin characters - Part 3: Persian language - Simplified transliteration.
- ISO 639:1988: Code for the representation of names of languages.
- ISO 6438:1983: Documentation - African coded character set for bibliographic information interchange.
- ISO 8859-6:1987 (ASMO 449E): Information processing - 8-Bit single-byte coded graphic character sets.
- ISO 9036:1987 (ASMO 449): Information processing - Arabic 7-bit coded character set for information interchange.
- ISO/DIS 11822: Information and documentation - Extension of the Arabic alphabet coded character set for bibliographic information interchange.
- ISO-10646 (Unicode)

II. ARABIC AND INTERNATIONAL ORGANIZATIONS

There are several international organizations involved in ICT standardization, such as the International Standardization Organization (ISO), International Electrotechnical Commission (IEC), International Telecommunication Union (ITU) and the International Internet Engineering Task Force (IETF). ISO and IEC have a joint technical committee on IT called JTC1. There are 18 major sub-committees (SC) within JTC1 and two independent working Groups.

On the other hand, ISO has more than 8 Technical Committees related to IT, other than JTC1, they are in the following fields: TC10 (Technical drawings), TC37 (Terminology), TC46 (Information and documentation), TC145 (Graphical symbols), TC154 (Processes, data elements and documents in commerce, industry and administration), TC171 (Document imaging applications), TC173 (Technical systems and aids for disabled or handicapped persons), and TC215 (Health informatics).

The European Telecommunications Standards Institute (ETSI) also plays a major role in developing a wide range of standards and other technical documentation as Europe's contribution to world-wide standardization in telecommunications, broadcasting and information technology. The development of telecommunications in Europe and the Arab States has received a boost with the signing of a Memorandum of Understanding (MoU) during 2002 between ETSI and the Arab Telecommunications Council of Ministers/League of Arab States (ACTM/LAS). Recognizing the importance of standards in the overall development of telecommunications and global harmonization, several major Arab telecom players have begun to sign up as Associate Members of ETSI.

The following are some of the Arabic and International organizations that deal with aspects pertaining to ICT Arabization and ICT standards related to Arabic language use in information society applications.

Through its industrial information center, the Arab Industrial Development and Mining Organization (AIDMO) provides information, statistics and publications pertaining to the various activities of the Organization. It also builds up advanced databases to serve Arab states in the fields of industry, mining and standardization. It further develops an Arab Industrial Information Network (ARIFONET) composed of sub-networks and focal points in Arab countries in addition to Arab and international centers and data banks. The purpose of ARIFONET is to promote an exchange of industrial and technological information among Arab countries and the rest of the world. This exchange is made possible by connecting the information centers in the field of industry, mining, standards and measurement, electrical energy associations, and technology transfer centers, industrial research centers, statistics bodies, to the Organization's main information center and to regional and international information centers through the net.

AIDMO works within the strategy of Arab common economic action initiated by the Arab summit conferences. Its major objective is to achieve the following:

- To ensure Arab industrial co-ordination and integration
- To contribute to the development and promotion of the Arab economy and its support in the fields of industry, energy, mining and standardization with a view to strengthening its productivity, quality and competitiveness.
- To plan for the support and elaboration of local, national and regional industrial projects and to encourage investment in mining and industry in the Arab world.
- To establish Arab unified standards in order to facilitate trade among Arab countries.
- To promote technical, technological and industrial co-operation among Arab states, and with foreign developed and developing countries.

TC-8 is an AIDMO-CSM technical committee that handles standards pertaining to the use of the Arabic language in ICT systems.

In 1981 the Arab Standardization and Metrology Organization (ASMO), with the head-quarter in Amman, formed TC-8 (Arabic Character in Informatics). ASMO was one of the Arab League organizations till 1989 when it was dissolved and replaced by a center attached to The Arab Industrial Development and Mining Organization (AIDMO), called Center for Standardization and Metrology (CSM), with the head-quarter in Rabat. The secretariat of TC-8 was given to SASMO in Damascus till 1992 when it was transferred to Tunisia until 1998 and later affiliated to SASMO again. Its name was changed to: TC on Usage of Arabic in IT. The main activities of TC-8 are the following [3-4]:

- Prepare standards on the Arab League level for the use of Arabic in IT.
- Coordinate with other standardization organizations such as the European Computer Manufacturing Association, Arabic Task Force (ECMA-ATF).
- Participate at meetings of international TCs and WGs on IT.
- Cooperate with R&D institutes and universities active in IT.

TC-8 has held 12 official meetings since it was formed in 1981, and participated in many meetings of ECMA-ATF, ISO and ALECSO. The last three meetings were the 11th held in Damascus on 6-8 December 1996, the 12th held also in Damascus on 17-19 November 1998 and the 13th meeting held in Cairo on 5-6 April 1999 at the GITEX IT Exposition venue. Table 1 shows the standards developed by TC-8, some of which has been adopted as ISO standards.

Table 1. Standards developed by TC-8

Arabic Standard	Year	Description	ISO Standard
ASMO-445		Bilingual 5-bit code (telex)	
ASMO-449		7-bit Arabic character-set SASO-429(1986), GSMO-50	ISO/9036 (1987)
ASMO-584		Trans-coding ASMO-449/ ASMO-445	
ASMO-662	1985	8-bit Arabic character-set	
ASMO-663	1987	Arabic keyboard layout GSMO-596(1995)	
ASMO-708	1986	8-bit Arabic/Latin character-set. SASO-1139(1996), GSMO-653(1996)	ISO/8859-6 (1987)
ASMO-968	1988	Trans-coding ASMO-662/ASMO-445	
CSM-969	1992	Displayed and printed Arabic character-set	
CSM-1021	1992	Trans-coding AASMO-708/ASMO-445	
		16-bit Multilingual character-set	Unicode
		32-bit Universal Character Set (UCS)	ISO/10646

The following is a summary of some of the Arabic and International organizations that deal with aspects pertaining to ICT Arabization and ICT standards related to Arabic language use in information society applications.

ARABIC AND INTERNATIONAL ORGANIZATIONS

		ORGANIZATION	COMMENT
International	ISO	International Standardization Organization	ISO has more than 8 Technical Committees related to IT
	IEC	International Electrotechnical Commission	ISO and IEC have a joint technical committee on IT called JTC1
	ITU	International Telecommunication Union	
	IETF	International Internet Engineering Task Force	
	ICANN	Internet Corporation for Assigned Names & Numbers	Management of the Internet's domain name
Arab	AIDMO	Arab Industrial Development & Mining Organization	Unify standards among Arab States. TC-8 handles standards pertaining to the use of the Arabic language in ICT systems
	AKMS	Arab Knowledge Management Society	Utilize modern management and technology to effectively develop Arab capabilities to derive social and economic value from an organization's knowledge resources
	TAGI	Talal Abu-Ghazaleh International Society	Foster innovation for enhanced creativity in the knowledge-based global economy
	AINC	Arabic Internet Names Consortium	Internationalization of the Internet. Arabize domain names.
	ALECSO	Arab League Educational, Cultural & Scientific Organ.	Promotion & co-ordination of educational, cultural & scientific activities at the regional level
	Arabcin	Arab Club for Information	Promotion of Arabic content generation, digitization, and dissemination
Arab Banks	BDSM	Bahrain Directorate of Standards & Metrology	(1975) NSB in the Ministry of Commerce
	EOS	Egyptian Organ. for Standardization & Quality Control	1957 NSB in the Ministry of Industry & Mineral Wealth
	GSMO	GCC Standardization & Metrology Organization	1982 Regional Organization, for the Gulf Cooperation Council (GCC Countries); which include: Bahrain, Kuwait, Oman, Qatar, Saudi Arabia and UAE
	ICOSM	Iraqi Central Organ. for Standardization & Metrology	1963 NSB in the Ministry of Planning, the Planning Board
	JISM	Jordan Institution for Standards & Metrology	NSB in the Ministry of Industry and Trade
	KSMD	Kuwait Standards & Metrology Department	1977 NSB in the Ministry of Commerce & Industry
	LIBNOR	Lebanese Standards Institution	1962 NSB in the Ministry of Industry
	ODGSM	Oman Directorate General For Specs & Measurements	1976 NSB in the Ministry of Commerce & Industry
	PSSE	Palestinian Standards & Specifications Establishment	
	QDSMCP	Qatar Dept of Standards, Measurements & Consumer Protection	1972 NSB in the Ministry of Finance, Economy & Commerce
	SASO	Saudi Arabian Standards Organization	NSB in the Ministry of Industry
	SASMO	Syrian Arab Organization for Standardization & Metrology	1969 NSB in the Ministry of Industry
	SSUAE	Directorate of Standards & Metrology of UAE	1976 NSB in the Ministry of Finance & Industry

III. IDENTIFICATION OF ARABIC LANGUAGE STANDARDIZATION ISSUES

A. DOCUMENTATION OF STANDARDS

It is important to ensure the translation of existing standards, especially pertaining to ICT, into Arabic. Some effort has been carried out in this respect. For example, the translation into Arabic of 249 ISO standards adopted in the 17th meeting of the Higher Consultative Committee for Standards, AIDMO, which was held during 12-13 May 2001. However, there is still a long way to go with regards of translation of existing standards pertaining to ICT systems and protocols into Arabic. Although there are many translation tools today that support translation across language pairs, very few include the Arabic language as an option.

B. INTERNET

Some of the problems faced when using Arabic on the Internet include the following:

- Character Set: There are two main reasons why using non-English languages pose problems. One is because there are different computer types - DOS, Windows, Mac, Unix - on the Net, and these differ in how they handle non-English. The other is caused by the network itself, and the specific restrictions that are built into it, which hurts in particular non-English languages that require more characters than unaccented A-Z.
- E-mail: Some mail software supports Arabic almost correctly with some minor glitches. As for alignment, there is an unfortunate but unavoidable side-effect of the fact that the may be different across systems. Ay period, colon or similar punctuation in an Arabic text from Windows (or Unix) will cause the line to be broken in two, before and after the period. The text before the period will be displayed to the left, and the text after to the right of the period (as if it had been English).
- Usenet News: Usenet News is the place on the Internet where we can gather, question, and discuss our experiences within a wide variety of topics. It is a giant, worldwide bulletin board which facilitates free exchange of information for everyone ad gives everyone an equal opportunity to participate in the discussions. Arabic isn't really much used on Usenet news yet.
- WHOIS: Used for a variety of important purposes, including identifying and verifying online merchants, investigations by consumer protection and other law enforcement authorities, determining whether a domain name is available for registration, enforcement of IPRs, and addressing cyber-attacks and otherwise resolving technical network issues.

C. INTERNATIONALIZATION OF WWW (HTML, URL, HTTP)

Internationalization of HTML (Hypertext Markup Language)means that HTML should be able to deal with non-Western characters in the text, such as Arabic, Chinese, Thai, etc. The non-Western characters should be represented in a HTML document properly. Further, it should support their display and other operations correctly, since a HTML document will contain text fragments in multiple languages.

XML (eXtended Markup Language) 1.0 is a relatively new markup language in the Web. It becomes easy to define document types which can be globally shared on the Web. XML is based on ISO 10646/Unicode. XML requires that any XML processor accepts both UTF-8 and UTF-16 (Unicode encodings). It supports also the attribute (xml:lang) which indicates the language of the contents.

URLs are Web resource addresses and are limited to ASCII. However, an address is the facility and "guide" to find someone or something. This restriction to ASCII forces non-English users to provide their addresses in ASCII. Among diverse solutions discussed, the use of UTF-8 seems to be the preferred character encoding for URIs.

HTTP (Hypertext Transfer Protocol) is an 8-bit protocol. Special characters can be transferred properly. HTTP uses language tags within Accept-Language and Content-Language (RFC1766). However, a few compatibility issues, such as MIME type (Multipurpose Internet Mail Extensions) while transmitting Unicode text, should be addressed in the future.

Using MHTML (Multilingual-HTML) Server users can display and search multilingual documents from any Java-enabled browser. The MHTML system consists of two components, MHTML server and MHTML viewer applet. The MHTML server converts on the fly a HTML document into a MHTML document and sends it to a client. Once the viewer applet receives the MHTML document, it will display it on the client browser.

D. HARDWARE

The standardization issues pertaining to computer hardware include unifying standards mainly for displays, keyboards, and other computer peripherals.

- Displays: The issue of how Arabic characters will appear on displays is not clearly standardized. Also, a problem with Arabic text is that it is written from right to left, and mixed Arabic/Latin strings include text in both directions presented on the same line. Numerals are mainly handled in the same way as Latin languages. Numbers are read from left to right (although Arabic text is read from right to left) with the highest order digit on the left side.
- Dual keyboard management: In most cases, European language keyboards have one specific keyboard layout, including all needed Latin letters. Since the Arabic alphabet is different from the Latin character set and because a user must always be able to input Latin and Arabic characters from the keyboard (ISO 8859-6 includes both ASCII and Arabic characters), a dual keyboard management system is needed. The keyboard management system must allow the user to switch from one language to the other using a single keystroke.
- Optical Character Recognition: Because Arabic is a connecting letter language, it is quite difficult to use the same method and algorithms for Optical Character Recognition as for Latin languages. The main problem is the ability to extract a single letter from a word.
- Peripherals: Many peripherals such as printers, plotters, HMD (head mounted displays), etc. lack the Arabic language support through unified standards pertaining the the hardware and software aspects.

E. APPLICATIONS

Many applications lack the needed standards for Arabic language utilization. This includes applications in the areas of word processing, spreadsheets, search engines, chatting, cryptology, operating systems, and other software.

- Word processing: ArabTEX is a package extending the capabilities of TEX/LATEX to generate the Arabic writing from an ASCII transliteration for texts in several languages using the Arabic script.
- Search engines: There is an increased need for methods and tools to publish Arabic content information. Although Web pages with Arabic script constitute a very small portion of the World Wide Web, there is a need for tools to enable non-HTML based databases and textual data, with Arabic scripts, to be indexed, searched and published.
- Operating systems: The DOS market created most of the Arabic implementation standards. Microsoft Arabic DOS and Arabic Microsoft Windows have reinforced this aspect. Some of these PC solutions could not be implemented on the UNIX environment. For example, in Arabic DOS

character applications, users press Right Shift+Left shift to toggle their keyboard layout between Arabic and Latin mode. The UNIX operating system is unable to use this sequence to switch between two internal logical keyboard mappings.

IV. SELECTED PRIORITY ARABIC LANGUAGE STANDARDIZATION ISSUES

The priority issues that need to be addressed pertaining to harmonization of ICT standards related to Arabic language use in information society applications include:

- Character Sets: One of the biggest problems is the issue of multiple character sets for representing Arabic. The knowledge of the character set used is needed for the correct encoding at the transmitting end and the ability to decode the text at the receiving end. Specifying the parameters of a textual transmission requires both (1) a set of labels for specifying the character set (which can be done in the MIME headers), encoding scheme, and transfering syntax used, and (2) a procedure for attaching these labels to the data.

A review of the history of Arabic character sets can be summarized by the following:

- ✓ In 1981, CUDAR-U appeared as the first standard Arabic character set (which used 7 bits per character).
- ✓ In 1982, ASMO produced its first character set standard, AMSO-449 (7 bits). It became the basis for all subsequent standard sets. Thus, it has a role similar to ASCII for Latin characters.
- ✓ In 1986, ASMO standard 708 appeared (8 bits), and became the international standard ISO-8859-6.

- Arabic display issues: Difficullties pertaining to the display of Arabic characters include representation of Arabic text and the directionality of display

- Arabic e-mail: So the two problems in exchanging Arabic e-mail are (1) correctly transporting Arabic messages that are encoded using 8 bits, and (2) specifying the language and character set used in a particular message since transporting e-mail does not involve a prior exchange of information about content (as in HTTP). These problems are both solved by the MIME standard. MIME allows labeling and structuring message contents using RFC 822 headers because it introduces a new set of header fields that are added to the message header. This way, it allows the sending of binaries and non-ASCII text through e-mail by encoding them in ASCII.

- Arabic Web Browsing: The specification of the WWW system has two main components: The page transfer protocol (HTTP) and page description language (HTML). HTTP is an 8-bit clean protocol, meaning that it allows the transport of Arabic pages in 8-bit character sets. The major issues surrounding the use of Arabic on the Web are the labeling of the character set used and that of marking up Arabic pages in HTML.

- Arabic search and indexing: With the huge amount of information on the Internet, search and indexing tools are crucial for locating specific resources and organizing information. The search and indexing of Arabic text is more involved than other languages such as English. Searching and indexing Arabic text must rely on the root of a word and not merely on the final form. Further, the same word can have more than 100 combinations of prefixes and suffixes, which in English would be preceding stop words such as "with" and "for." Search systems for Arabic need to employ morphological analysis, which is an involved process and has its limitations. The large number of synonyms of Arabic words intensifies this problem greatly.

- Speech Technology:

Automatic speech translation

Some of the English to Arabic automatic translation products include (a) App-Tek, which has a PC product for English-to-Arabic translation, (b) Systran, which has English-to-Arabic software that runs on IBM mainframes.

Speech recognition

Particularities of the Arabic language such as geminate and emphatic consonants and the vowel duration are unanimously considered as the main root of failure of Automatic Speech Recognition (ASR) in systems dedicated to standard Arabic. Novel speech recognition models for Arabic have been explored in a Workshop on Language Engineering (http://www.clsp.jhu.edu/ws2002/groups/arabic/) held at Johns Hopkins University from July 15 to August 23, 2002.

Speech synthesis

A Text-To-Speech (TTS) system that supports Speech Synthesis will be responsible for rendering a document as spoken output. A survey of existing tools for developing synthesis of speech and for evaluating the quality of speech synthesis can be found at http://www.disc2.dk/tools/SGsurvey.html.

V. LEGISLATIVE & REGULATORY ISSUES FOR INFORMATION AND KNOWLEDGE MANAGEMENT

It is necessary for regulatory and law-drafting purposes to develop the basic concepts and principles of information and knowledge management, place them properly in the context of the legal system and revise the manner in which the legal order is described as a systematic entity. One of the regulatory challenges is the anonymity of legal transactions resulting from the development of ICT. The traditional categories and classifications of transactions and the corresponding regulatory regimes do not have any bearing on the new technical reality. One example, is the merger of the different types of media. This phenomenon creates a challenge to revise the systematic approach in various laws and other regulations and has, therefore, significant consequences for the practical business of law-drafting.

A. ICT TERMINOLOGY

It is vital to have a standard that unifies ICT terminologies used in all ICT-related Arabic language standards. As an example, the IEEE's Standards Coordinating Committee 10 (SCC10) is responsible for overseeing the use and development of terminology in IEEE standards. A similar committee should be formulated to unify ICT terminologies used in all ICT-related Arabic language standards.

B. STANDARDS FOR THE UTILIZATION OF ARABIC LANGUAGE ON THE INTERNET

An Internet Best Practices Committee for the Arabic language should be formed. An example of such a standard is the IEEE (Std 2001-1999) Recommended Practice for Internet Practices (Web Page Engineering and Intranet/Extranet Applications). This standard defines recommended practices for Web page engineering. It discusses life cycle planning: identifying the audience, the client environment, objectives, and metrics, and continues with recommendations on server considerations, and specific Web page content. This standard is intended to reduce site-management costs, reduce legal risks, facilitate user satisfaction, and increase the productivity of Web applications for both maintainers and users.

C. SECURITY ISSUES

Data security consists of three basic elements: confidentiality, integrity and availability:

- Confidentiality: means that data and data processing are available and the existence of data is disclosed only to duly authorized persons.

- Integrity: refers to the qualities of data and data processing being reliable, authentic in relation to the original data, structurally and logically correct, valid and complete. Integrity measures aim to protect the maintenance of the above-mentioned qualities of data and data processing. Integrity is at the core of maintaining the quality of information.

- Availability: Availability means access to data and data processing in real time and the utility of data in relation to the purpose to which data is collected and processed.

D. The Vulnerability of Technology

Developments in technology, technological dependencies and threats make it necessary for the modern society to implement measures to reduce the vulnerability in society. It is therefore important to establish a level of robustness in ICT infrastructures that makes it improbable that important functions in society will come to a stop in any normal situation. In order to achieve this goal, we must have a comprehensive strategy to reduce a society's ICT vulnerability:

- Partnership between private and public authorities: A partnership should be established between the authorities and companies that are either responsible for operating or are dependent on ICT infrastructures. This brings into focus the necessity of cooperation, mutual trust and exchange of information.

- The exchange of information: In addition to the need for information to be exchanged through a partnership, there is also a great need to develop flexible and dynamic mechanisms for the mutual exchange of information between various types of organizations, in the private as well as the public sectors. The systems and infrastructure have gradually become so complex that no one can have a full overview. In such a situation, it is important to have a forum for information exchange and cooperation for companies within and across sectors.

- Greater warning capability: Today the Internet is used extensively in production and operation of critical infrastructure services. If the systems are not sufficiently protected, it will be easy to penetrate and attack them. Many operators of critical infrastructures currently do not have sufficient information on neither vulnerabilities nor threats to be able to react quickly in a critical situation. It is therefore a need to implement measures so that companies can share technical surveillance, notification and handling of security incidents.

- Education and competence: In general, many organizations and companies have little awareness and knowledge of ICT vulnerability and security. A competence gap has arisen due to the growing complexity of the ICT systems and the challenges posed by the new threat scenarios. Efforts should be made to provide ICT security education.

- Research and development: With the exception of what takes place in individual companies and some technological environments, little research is being carried out in vulnerability analysis. It is therefore a need to implement measures to strengthen research and developmental activities in this area.

- Securing infrastructures of critical importance to society's other ICT systems: Some infrastructures are critical to a society's other ICT systems and should therefore be emphasized. This is primarily true of public telecommunications.

- Adapting laws and regulations: For the authorities, it will be a challenge to develop a legal framework that is relevant and forceful and adapted to the dynamics in society, created by ICTs. A key challenge in this area is how to balance individual freedom and the need for collective protection.

The following measures are proposed to realize the strategy:

- Center for Information Assurance: establishing such a Center based on a partnership between the public and private sector sectors. The center's main goals should be to coordinate some of the efforts to improve ICT security and to contribute to a more robust ICT infrastructure.
- Increased research and development: There is a need for co-ordinated R&D efforts in the area of ICT vulnerability and security in the Arab world. A strategic research program should be established in the field of ICT vulnerability and security. It is important that business and industry participate in the research work – in consistency with the partnership concept.
- Knowledge and education: Lack of awareness and knowledge in the field of ICT security is believed to be a major problem. It is a shortage of educated and trained personnel in the area of security and vulnerability, and thus a need to improve ICT training, with a focus on security, at all levels. This will help to improve the more general level of ICT security knowledge.
- Risk and vulnerability analysis: Risk and vulnerability analyses are becoming increasingly important as instruments for dealing with vulnerability and security issues. Efforts should be undertaken to make companies that depend on ICT systems use risk and vulnerability analysis. Government authorities having supervisory responsibilities should also use such analysis. Suitable methods and tools should be developed.
- User targeted measures: Even though measures to reduce vulnerability are implemented within individual infrastructures, they will never be sufficiently robust to cope with every possible challenge. It is therefore important that measures also are directed towards the end-users. The end-user target group ranges from private individuals to companies and public authorities. It is important to ensure that end-users at all times and to the greatest extent possible are aware of their vulnerabilities.
- Legislation and combating crime: Internet and ICT developments take place extremely quickly. The legislative process, on the other hand, is a thorough process that may take several years. The framing of legislation is in danger of lagging behind the society the laws are intended to regulate. It is therefore a need for mechanisms that will allow legislation to be prepared more quickly.

E. Digital Preservation Strategies

While digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preserving access to 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.

The use of standards is one strategy, which may be used to assist in preserving the integrity of and access to digital information. Adherence to standards can assist by facilitating the transfer of information between hardware and software platforms as technologies evolve.

There are different levels of standards with varying degrees of authority. The Guidelines on Best practices for Using Electronic Information (DLM Forum, 1997) define the following three levels of standards:

- De facto standards: these are standards that are commonly accepted by the marketplace; they are established by common practice or dominant market share.
- Publicly available specifications (PAS): these standards are developed when "several leading firms on the market join together in a consortium to define an interface standard". For example, specifications produced by the Internet Engineering Task Force.

- De jure standards: these are standards that are formally established by law, or by a recognized standards-setting body such as the International Standards Organisation (ISO).

There are standards for the many different aspects of storing and accessing digital information, including standards for: interoperability, data format, resource identification, resource description, data archiving and records management. Following are some examples of different sorts of standards, which are relevant to preserving access to digital information.

Interoperability standards: This type of standard allows communication between different systems, facilitating the discovery of and access to digital information. For example, ISO 23950 Information Retrieval Standard (equivalent to ANSI Z39.50) defines a standard way for two computers to communicate and share information.

Resource encoding standards: These standards define formats for the different types of digital information. They include standards for page description formats (e.g. postscript, PDF); graphics formats (e.g. TIFF, GIF); structured information (e.g. SGML); moving images and audio formats. Adherence to this type of standard allows data compatibility across a wide range of systems.

Resource identification standards: A way of uniquely identifying digital resources is desirable to ensure long-term and reliable access to resources while they are available over the Internet.

Resource description standards: Resource description standards can facilitate effective resource discovery. These include description standards such as AACR2, a set of rules used for describing library material, and the semantic aspects of Dublin Core, a descriptive metadata standard developed for resource description on the Internet.

Data archiving standards: The Open Archival Information System (OAIS) Reference Model is a model for an archival system for the long-term preservation of and access to digital information. OAIS has been developed by the Consultative Committee for Space Data Systems (CCSDS) and is expected to be released as a draft ISO standard in the future. Projects investigating the use of the OAIS model include NEDLIB in Europe, CEDARS in the UK, and PANDORA in Australia.

Records management standards: Records management standards provide guidance on how to implement records management strategies, procedures and practices. The AS 4390 series of standards produced by Standards Australia are an example of records management standards. This series includes standards relating to the following aspects of records management: responsibilities, strategies, controls, storage, appraisal and disposal.

VI. ACTION PLAN

Forming a Consulting Council to be composed of an Arab international work team with the participation of representatives from Arab countries, and which comprise representatives from the International Telecommunications Union (ITU), Word Trade Organization (WTO), UNCTAD and ESCWA, in addition to AKMS. This initiative aims basically at making the consulting council provide advice to governmental and private organizations in all Arab countries, for the following purposes:

- Establish an Arabic unified information network, through which the Arab communities can reap the fruits of technology, wherever they are.
- Establish Arab-consulting organizations that will cooperate with similar international organizations, in particular those related to the fields of economy, finance, industry and trade, and medical services.
- Establish an Arab organization for integrated solutions, and on the basis of international standards to render service to the Arab knowledge-based society.

Forming a technical committee in the above-mentioned Consulting Council to formalize a strategy for reducing ICT vulnerability addressing legislative and regulatory issues for information knowledge management and providing a security policy for the use of Arabic language on the Internet.

Forming a technical committee in the above-mentioned Consulting Council to unify ICT terminologies used in all ICT-related Arabic language standards.

Forming an Internet Best Practices Committee in the above-mentioned Consulting Council for the Arabic language. This standard defines recommended practices for Web page engineering. It addresses the needs of Webmasters and managers to effectively develop and manage WWW projects in order to reduce site-management costs and legal risks, facilitate user satisfaction, and increase the productivity of Web applications for both maintainers and users. Accordingly, the standard should provide guidance to Web developers by addressing the following issues: copyright, proprietary data declarations, indexing and content classification of pages, use of epoch transparent dates, multinational sensitivities, browser tolerance, bandwidth efficiencies, server operations, privacy and accommodation of physically challenged persons.

Strengthen the role of AIDMO-CSM's TC-8 in realizing its objectives by the Arab League of states.

NSBs in the Arab world should form ICT technical committees responsible for issues related to the use of Arabic language in ICT systems. These committees can cooperate with TC-8 of AIDMO-CSM.

Create policies with the aim of supporting R&D efforts in areas that enhances the objectives of this study such as in the area of Natural Language Processing (NLP).

The ICT technical committees formed by NSBs (recommendation f) above) in the Arab world should adopt the following proposed measures to realize a strategy for reducing a society's ICT vulnerability: center for information assurance, increased research and development, knowledge and education, risk and vulnerability analysis, user targeted measures, and legislation and combating crime.

Establishing a standard for expressing the security functional requirements in order to facilitate a common security evaluation criteria which provides a method to measure the capability of an installed system of trust or information security products.

Establishing unified standards for the many different aspects of storing and accessing digital information, including standards for: interoperability, data format, resource identification, resource description, data archiving and records management.

Systematize all the bodies, organizations, committees, technical sub-committees, etc. to some kind of an operational hierarchy to avoid repetition and promote efficiency in the task of establishing all relevant ICT standards in the Arab world.

AIDMO and ESCWA could cooperate, with other international standardization organizations to elaborate a standardization policy for the ESCWA region and the Arab world, and consequently issue recommendations for adopting standards especially in IT, trade facilitation, quality management and the like. Such policy should aim at promoting, adopting and harmonizing standards, on the one hand, and should strengthen the institutional capacity of NSBs, particularly their activities in implementation, accreditation and certification. The success of such a policy would make the environment conducive to standardization.

Awareness of the need for standardization is much needed in the region. It is recommended that curricula of the educational systems of the region include courses on standardization and quality management at all levels.

VII. CONCLUSION

The Arabic language is being increasingly used on the Internet despite significant obstacles. This study provides a brief historical overview of Arabic language calligraphy and script, addresses the Arabization of the Internet and relevant standards and stresses the importance of ICT standards for the Arabic language. It also provides information on some of the main players in this Arabization process. Major problems are identified and new international standards and Internet protocols that help alleviate the problems are outlined. One of the biggest problems is the issue of multiple character sets that represent Arabic. The use of MIME tags in specifying the name of the character set used has been discussed. The advent of Unicode will perhaps mean the replacement of other character sets and thus present a unified character set for all languages. Arabic has some particular characteristics that require specialized display routines, including the need for contextual analysis to select the appropriate character shapes and the incorporation of a bi-directional text display routine to order the text correctly. The Unicode standard provides a bi-directional display algorithm.

The two most important Internet applications, e-mail and the WWW, must work in Arabic flawlessly. MIME facilitates this for e-mail by allowing the specification of character set and by encoding 8-bit messages in 7-bits for safe transport. The HTML 4.0 specification also provides facilities for Arabic character set indication and for Arabic message markup. Alternate techniques for Web Arabization provide interim solutions until a simpler and more satisfying one is found. It is expected that when Web servers adhere to HTTP standards and include character set information in the header for page transmission, and when Web browsers use that information and set up the pages accordingly, the problem will be solved. The ability to search and index Arabic content on the Internet is crucial. Features of the Arabic language that relate to text search were mentioned and available systems were listed.

The ICT sector is growing in the knowledge economy and the Arab world can have an active role in this sector, especially in products developed in Arabic. Standards for Arabic in ICT are more needed on the regional level than on the national level. It is important for the Arab world to promote and accelerate the efforts on the regional level. AIDMO-CSM should be supported. Moreover, it should enhance its cooperation regionally and internationally. To achieve that, the following recommendations are proposed:

- NSBs in the Arab world should form ICT technical committees responsible for issues related to the use of Arabic language in ICT systems. These committees can cooperate with TC-8 of AIDMO-CSM.
- NSBs in the Arab world must adopt a regional and modular project to computerize the activities of NSBs in the Arab world. The project should work to unify procedures at all levels and for all activities of the NSBs. It will help connect these NSBs to the Internet and network them through linking their sites, establishing databases of standards at each NSB and making the catalogues available on the Internet sites in order to coordinate efforts and avoid repetition of work. This project would also help link NSBs to other institutions in each country (chamber of commerce and industries, Ministries, R&D institutes, etc.).
- AIDMO and ESCWA could cooperate, with other international standardization organizations to elaborate a standardization policy for the ESCWA region and the Arab world, and consequently issue recommendations for adopting standards especially in IT, trade facilitation, quality management and the like. Such policy should aim at promoting, adopting and harmonizing standards, on the one hand, and should strengthen the institutional capacity of NSBs, particularly their activities in implementation, accreditation and certification. The success of such a policy would make the environment conducive to standardization.
- Awareness of the need for standardization is much needed in the region. It is recommended that curricula of the educational systems of the region include courses on standardization and quality management at all levels.

- The private sector participation in the activities of standardization is vital and should be encouraged.

Legislative and regulatory issues for information knowledge management are discussed regarding the absence of ICT Arabic terminology, standards for the utilization of Arabic language on the Internet as well as security issues. Finally, an action plan is proposed to systematically attend to all the issues raised in this report that will lead to harmonization of ICT standards related to Arabic language use in information society applications with the potential of over 300 million users.

ACKNOWLEDGEMENTS

I would like to thank my colleagues at ESCWA for their help, encouragement, and contribution to this study. Special thanks go to Dr. Hasan Charif for his vision and continuous guidance and to Dr. Mohammad Mrayati for his numerous comments and suggestions. My gratitude also goes to Dr. Mansour Farah and Dr. Omar Bizri for their assistance in the many discussions we had. I also thank Dr. Abdulilah Dewachi for his observations and recommendations.

REFERENCES

1. H. Diab, "Harmonization Of ICT Standards Related To Arabic Language Use In Information Society Applications," ESCWA Report, October 2002.
2. TAGI & AKMS, Towards An Arab Knowledge Society. Report. April 2001.
3. M. Mrayati, "Standards for Arabic in Information Technology," Meeting on Standardization in the Arab Countries, 2-5 February 1999, Amman, Jordan.
4. AIDMO, "Use of Arabic Language on the Internet," A draft for Arabic standards, 2001, Tunis.