MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

ICPC Multilingual Collaboratory: a Web-and Unicode-Based System for Distributed Editing/Translating/Viewing of the Multilingual International Classification of Primary Care

R. P. Channing Rodgers[a], Ziying Sherwin[a], Henk Lamberts[b], Inge M. Okkes[b]

[a] *Office of High Performance Computing & Communications (OHPCC)*
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda MD 20894
[b] *Department of General Practice/Family Medicine, Academic Medical Center/University of Amsterdam*

Abstract

The International Classification of Primary Care (ICPC) is a clinical classification containing 726 clinical concepts, available in over 20 languages, augmented by links to ICD-10 concepts. It is employed in clinical information systems in several European countries, Israel, Japan, and Australia. In translating ICPC, it has been challenging to manage the flow of multilingual information, maintain its quality, and optimize its portability, particularly in light of the numerous character encodings used to represent its content. The ICPC Multilingual Collaboratory (IMC) is a World Wide Web-based environment, created to allow the viewing, maintenance, and translation of ICPC content by a dispersed international editorial staff. Based upon open-source software, it represents ICPC content using the Unicode standard for character encoding. The system implements three interfaces to ICPC data: 1) a password-protected editorial interface which instantiates a hierarchical authority model and communication channels for review and control of content, including a means of uploading new candidate translations; 2) an openly accessible read-only interface, with email access to the editors (providing another level of content review); and, 3) a management interface for the system administrator. The completed system powerfully demonstrates the ability of the World Wide Web, open-source software, and Unicode to expedite and simplify international multilingual collaboration, even in a world in which Unicode support is incomplete on existing computing platforms.

Keyword:

Classification, Collaboration, ICPC, Internet, Language, Medical Informatics, Unicode, World Wide Web, ICPC Multilingual Collaboratory: a Web- and Unicode-Based System.

Introduction

The International Classification of Primary Care (ICPC)[1-4] is a collection of 726 clinical concepts and associated information, "providing logically structured classes for ... [international family practice's] common symptoms, complaints, diagnoses/health problems and interventions. [3]" As supplemented by the much more specific (and much larger) ICD-10 vocabulary, it has become the official classification system for family practice in Finland, The Netherlands, and Norway, and has proven helpful for clinical research in several other countries [5-7].

Table 1: Biaxial ICPC structure

Seven Components	
1	Symptoms and complaints
2	Diagnostic, screening, preventive
3	Medication, treatment, procedures
4	Test results
5	Administrative
6	Referrals and other reasons for encounter
7	Diseases
Seventeen Chapters	
A	General and unspecified
B	Blood, blood-forming organs and immune mechanism
D	Digestive
F	Eye
H	Ear
K	Circulatory
L	Musculoskeletal
N	Neurological
P	Psychological
R	Respiratory
S	Skin
T	Endocrine, metabolic and nutritional
U	Urological
W	Pregnancy, child-bearing, family planning
X	Female genital
Y	Male genital
Z	Social problems

ICPC is a biaxial classification system (see table 1). Seventeen chapters (represented by alphabetic letters) referring to a body system/problem area form one axis; seven components (represented by single-digit numeric codes) form the second. Each ICPC record contains the following fields: ICPC code, Long Title, Short Title, Component (referring to one of the two axes of the biaxial classification), ICD10 codes, and five other fields not discussed here. An ICPC code consists of a chapter letter followed by a number (which is not related to the component). For example, for ICPC code A25, both the Long and Short titles are "Fear of death," and the concept is associated with Component 1 ("Symptoms and complaints") and a single ICD-10 code (Z71.1).

The working language of ICPC is English. ICPC's sponsoring body, the World Organization of Family Doctors (WONCA) encourages translations of the classification into as many languages as possible. It currently exists in 20+ languages, with further translations underway (see Table 2).

Table 2: Languages in ICPC

Currently Available			
Afrikaans(**1**)	Basque(**1**)	Chinese(2)	Croatian(2)
Danish(**1**,2)	Dutch(**1**,**2**)	English*(**1**, **2**)	Finnish(**1**)
French(**1**,2)	German(**1**)	Greek(**1**,2)	Hebrew(**1**)
Hungarian(**1**)	Italian(**1**,2)	Japanese(**1**,2)	Norwegian(**1**,2)
Polish(**1**)	Portuguese(**1**,**2**)	Romanian(**2**)	Russian(**1**,2)
Serbian(**2**)	Slowenian(2)	Spanish(**1**,2)	Swedish(**1**)
In Development			
Finnish(2)	German(2)	Swedish(2)	
Notes: *: original version ICPC version(s) in which a language is available are indicated within parentheses; those appearing in IMC are emboldened.			

It is daunting to manage a process based upon volunteer labor that spans languages, cultures, and widely separated geographical locales. With the issuance of ICPC-2 [2], content quality control issues became apparent, which resulted in the release of an electronic version, ICPC-2-E [3]. Newer translations have been offered to the editors, in which the translators have taken liberties with record format, making it difficult or impossible to integrate the new translation into ICPC. ICPC's multilingual content has existed in a wide variety of (often unidentified) character encodings (that is, the schemes by which alphabetic characters are represented by numerical values for purposes of their use in digital computers), making it difficult to display and to share. It was evident that the quality of translations would be improved by standardizing the submission process and character encoding, and exposing ICPC content to broader review. To address these problems, the present authors undertook the challenge of building an online mechanism for viewing, maintaining, and augmenting ICPC content.

Methods

Goals of the Project

The design goals for the ICPC Multilingual Collaboratory (IMC) include:

1. Maximization of platform- and vendor-independence, through the use of open-source software and open standards.
2. Creation of an asynchronous collaborative environment that works well across time zones and between cultures.
3. Creation of an operational structure that ensures disciplined control of content, while maximizing the scrutiny of content for enhanced quality control.
4. Unification of ICPC content through a single character encoding scheme encompassing all of its languages.
5. Promotion of the awareness and use of ICPC.

Program Design & Implementation

The IMC employs the server-client model of the World Wide Web. It runs as a service offered by an Apache web server [8], and is written in PHP [9], a server-side scripting language that operates as part of Apache. IMC data is stored using the MySQL relational database system [10], which is easily accessed from within PHP. The service currently runs on a Sun UltraSPARC 2 computer under the Solaris 2.8 operating system.

Users access IMC services through use of a web client ("browser"). As originally deployed, the web provided a single transaction between the source of information (a "server") and the software requesting information (a "client"). The web communication protocol is *stateless*; there is no memory that carries across multiple transactions. The rich interactions users have come to expect from web-based services arise from *stateful* mechanisms that have been grafted onto the original web environment by several means. The most common of these are: conveying state information within "hidden" fields in web-based forms; embedding information within the Uniform Resource Locators (URLs) that are used as web addresses; and, the use of "cookies," small packets of information that can be sent by a server for storage on the client (and which can be returned to the server upon request). PHP itself provides a server-side state information storage mechanism referred to as "session" support. To maintain state, IMC employs a combination of hidden fields, session cookies, and PHP sessions.

IMC supports four levels of users, in a hierarchy: *editors*, *subeditors*, *commenters* (intentionally *not* called commentators), and *viewers*, working together through a delegated authority model. Access to IMC by the first three (editorial) levels is by password only. Only an editor can make permanent changes to the ICPC database; an editor also controls the accounts of users lower in the hierarchy. A subeditor has the authority to make *proposed* changes to content for specified languages, which can be seen only by an editor, subeditor, or commenter. An editor or subeditor can create an account for a commenter, who can send comments to the editors/subeditors dealing with specified languages. Communication between the editors, subeditors, and commenters occurs via email or through an online forum, created by incorporating a modified version of the open-source w-Agora system [11] within IMC. Viewers are non-editorial users; they can access IMC without a password, but only to search and view official (and, at the editor's option, proposed) ICPC content, and to send email to the editorial staff. They can not read any editorial communications. Viewer comments serve as an added level of content review.

It was envisaged that the membership of each tier of the user hierarchy, moving from editor to viewer, would grow in size, providing a small tightly coordinated editorial staff making changes

to ICPC content, based upon the work of a large group of reviewers.

Unicode

The multilingual content of the IMC is represented using Unicode[12,13]. This character encoding specification arose out of a collaboration between Apple and Xerox in the late 1980s, which blossomed into a large formal consortium. Unicode shares the numerical codes it employs for representing characters with the ISO-10646 standard. Initially based on 16-bit characters, it has since been extended beyond that. Version 4 supports 96,382 assigned characters within its 1,114,112 numerical code points. The assigned character space is divided into numerous zones for *scripts* (alphabets), by far the largest amount being devoted to Asian languages. The standard attempts to provide for backward compatibility with older standards whenever possible; the initial 256 characters are identical to ISO 8859-1, the first 128 characters of which are ASCII (which appears as the first script of Unicode, "Basic Latin"). Unicode characters are commonly managed within a computer using a *character encoding form*, one of the most commonly used being UTF-8, which represents a single Unicode character by using between one and four eight-bit bytes, in a way which can be safely transported and manipulated within nearly all existing computing environments. ASCII is recognized as valid UTF-8 Unicode.

Using Unicode poses interesting technical challenges. There is not a strict one-to-one correspondence between Unicode code points and characters (for example, some accented characters can be represented by a single code point, or composed from the code points for the base character and the added accent mark); this has led to rules for normalization of character strings to allow for more consistent behavior, especially with respect to sorting and searching. Certain scripts need to be presented right-to-left (as for Arabic, Hebrew, Syriac, Thaana, and Yiddish ligatures). Identification of the language in use is sometimes important, as it may effect the way in which characters are treated, though it is important to stress that Unicode does not convey *linguistic* information, and should not be confused with the closely related (but broader) problem known as *internationalization* (or "*i18n*", in programming jargon). Unicode support is now claimed for a number of operating systems (including BeOS, Plan 9, Windows NT/2000, Mac OS X, Solaris 2.X, PalmOS), programming languages (including Java, Javascript, Perl, tcl/tk) and applications (including Netscape Navigator). In Unicode terminology, a *glyph* is a physical representation of a character. Both commercial and open source Unicode glyph sets are available, though all remain incomplete.

Translation of Existing ICPC Data into Unicode

Considerable *ad hoc* technical manipulations were required to translate existing ICPC data into Unicode for storage within a single database. The original data posed three problems:

1. It was preserved on multiple types of storage media.
2. It was encoded using many different (usually unidentified) character encoding schemes.
3. It was formatted using many different types of software applications (text, database, spreadsheet).

The conversions were performed by a combination of using the original software application within which the data was formatted to save the data in a more accessible form, by manual editing, and through the use of various purpose-written *awk* scripts on a UNIX workstation. Other useful tools include the GNU program *iconv* and a tool known as *native2ascii*, which accompanies the Java version 2 Software Developer's Kit. The process was carefully documented and the tools saved, for inclusion within the IMC online manual for use by others.

Multilingual Text Display

Because Unicode is not universally implemented on existing computer platforms, IMC makes minimal assumptions about the local availability of Unicode support for the display of text. This work is confined to the server: Unicode text strings are created using the Bitstream Cyberbit Unicode glyph set, and converted to JPEG graphical image files by a PHP-associated library, and can thus be viewed on any graphically-capable web client.

Multilingual Text Entry

Entry of Unicode text by editors is a more difficult problem, due to incomplete and non-uniform support for Unicode across existing computing platforms. English is the working language of the ICPC, so much of the communication will be in that language; however, the need for occasional entry of other ICPC languages is inescapable. When the editor/subeditor/commenter logs on to ICM for the first time, she is led through a series of tests to ascertain the level of support for Unicode on her computer, and directed to any necessary resources that need to be installed locally, which may include vendor-supplied operating system supplements, or third-party glyph sets and Unicode editors.

During the course of the project, operating system support for Unicode improved remarkably among the three platforms we concentrated upon: Windows, Apple Macintosh, and Solaris. Support remains partial, however, and is often installed only as an optional feature (especially in the case of Asian languages).

Initially, we hoped to provide a single open-source tool to operate on all platforms, to provide a uniform shared interface for Unicode data entry among all IMC participants. This effort focussed upon a Unicode-aware Java applet known as *JMUTT* [14], which draws a multilingual keyboard on the users monitor, which may be used to enter data. Sadly, we had difficulty getting this program to work reliably, and had to abandon its use. As an alternative, we identified and tested Unicode entry tools (all available either as open-source or a free binary) for the major existing computing platforms, finding five for Windows, and two for UNIX (native Unicode support within Mac OS X 10.2 was deemed sufficient for that platform). We worked closely with the author of a UNIX-based open-source Unicode editor, *mined* [15], to help him implement a flexible and extensible input entry method that enables the user to enter Asian characters by using multiple keystrokes on a standard QWERTY or ASERTY keyboard.

Results

In the course of installing ICPC data into the IMC database, we observed a number of errors and inconsistencies in various trans-

lations, an early affirmation of the utility of putting the information into a single encoding.

Of the numerous screens devised to implement the editorial, viewer, and administrator interfaces to IMC, we present just two examples here, from the viewer and editorial interfaces.

The publicly accessible search/view component of IMC, which confines the need for Unicode to the server, is illustrated in Figure 1, which contains the results display for a search for ICPC code A25, "Fear of Death," in multiple languages (the languages to be displayed are specified by the viewer in the course of the search process). In addition to several European languages based on the Latin alphabet, this example contains Greek, Hebrew, Japanese, and Russian, all rendered in a single display.

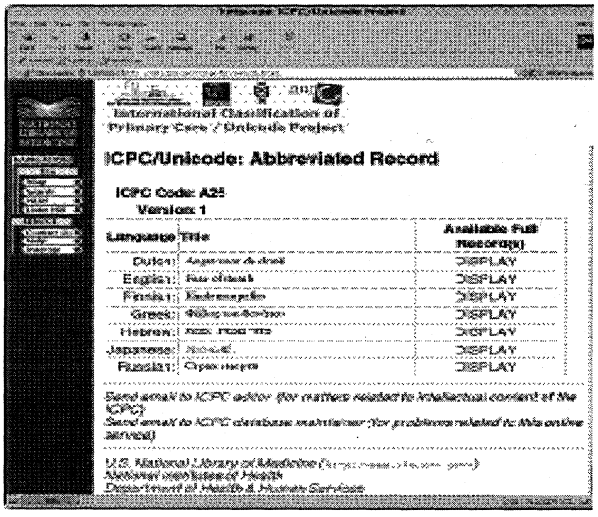

Figure 1 - Portion of the results screen from a search of IMC, showing the title for code A25, "Fear of Death", in several languages (Netscape/Solaris 2.8/Sun workstation).

Figure 2 presents an example of the use of the built-in discussion forum mechanism to send a comment to the editorial participants concerned with Hebrew, with respect to a translation of an ICPC concept. It demonstrates the use of the "Character Palette", one of several natively supported means by which a user can enter Unicode data under Mac OS X version 10.2 and above. In this case, the Mac OS cut-and-paste mechanism has been used to copy Hebrew characters from the palette into the forum message window. The result is a message in English that also contains a Hebrew word.

Each of the Unicode entry tools we explored had limitations. The only entry method that works on all tested platforms (Windows XP, Mac OS X, and SPARC/Solaris 2.8) is the Java-based *simredo* text editor [16]. Unfortunately, it lacks support for a number of languages in ICPC, and has no keyboard entry method applicable to Asian languages.

Discussion

The World Wide Web was invented to expedite scientific collaboration and communication, and collaboratories have been an active area of development from the outset. Computer-based collaboration has been employed in the development of medical vocabularies [17], and as an aid toward convergence among independently-developed local extensions to medical vocabularies [18].

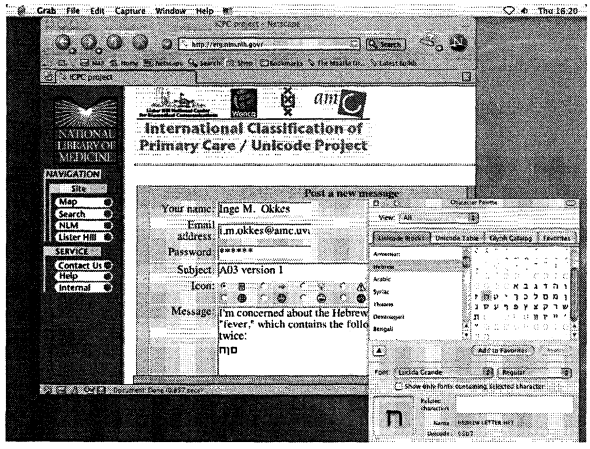

Figure 2 - Use of a mixture of English and Hebrew in the discussion forum (Netscape/iMac G4/ Mac OS X).

A successful collaboratory takes into account both technical requirements and the sociology of human interactions. This instance of using a collaboratory to enable people to "work together apart [19]" is complicated by its multilinguality. We believe we have addressed most of the technical requirements for IMC. Deployment is still at an early stage, and it is too early to assess the acceptability of the system among the various participants. We anticipate the need to make adjustments in system design to better accommodate their interaction and communication needs. We anticipate adding audio/video teleconferencing capability to the system in the future. We are currently adding an extension which will allow a new translation to be either directly entered into IMC from a web client, or uploaded in complete form according to a pre-defined XML Document Type Definition (DTD), currently under development.

Conclusion

Unicode is a standard of major importance, clearly poised to become the universal method for representing written human language within computers. It's incomplete deployment makes it a challenge to develop applications based upon it, but that situation is changing rapidly. During the source of this project, we observed heartening improvements in Unicode support within newly released operating systems. We were able to find a number of excellent open-source Unicode applications as well.

The quality control benefits of the IMC project to date, through sweeping aside a plethora of storage media and character encodings, and placing the languages side-by-side in the same database using the same character encoding, have been considerable. This system has made the ICPC much more widely and easily available. The remaining challenge is to carefully observe how IMC participants work within the established hierarchical edito-

rial framework, and make appropriate adjustments to optimize the efficiency of their interaction, as they labor together to expand and improve this clinical resource.

The publicly accessible portion of IMC, full documentation, and the encoding conversion and Unicode editing tools amassed during the project, are available from the project web site: *http://etg.nlm.nih.gov/project/IMC/.*

Acknowledgement

We thank the following individuals: Mark Leisher (New Mexico State University) for making his JMUTT tool available, and for working with us to improve it; Mark Leisher and Dr. Nelson Beebe (University of Utah) for numerous helpful technical remarks; Dr. Thomas Wolff for his energetic collaboration in improving his *mined* Unicode editor to support a multi-stroke keyboard entry method for Asian languages.

References

[1] Lamberts, H, Wood, M. *ICPC: International Classification of Primary Care.* Oxford: Oxford University Press, 1987.

[2] WONCA Committee IC. *International Classification for Primary Care,* Second Edition (ICPC-2). Oxford: Oxford University Press, 1998.

[3] Okkes, IM, Jamoulle, M, Lamberts, H, Bentzen, N. ICPC-2-E: the electronic version of ICPC-2. Differences from the printed version and the consequences. *Family Practice* 2000; 17: 101- 107. http://www.fampra.oupjournals.org/.

[4] Okkes IM, Lamberts H. Classification and the domain of family practice. In: Jones R, ed. *The Oxford Textbook of Primary Medical Care.* Oxford: Oxford University Pres, 2003. Vol 1, pp 139-52.

[5] Hofmans-Okkes, IM, Lamberts, H. The International Classification of Primary Care (ICPC): new application in research and computer-based patient records in family practice. *Family Practice* 1996; 13: 294- 302.

[6] Lamberts, H, Okkes, I. Patients with Chronic Alcohol Abuse in Dutch Family Practices. *Alcohol & Alcoholism* 1999; 34: 337-345.

[7] Okkes IM, Polderman GO, Fryer GE, et al. The role of family practice in different health care systems. A comparison of reasons for encounter, diagnoses, and interventions in primary care populations in the Netherlands, Japan, Poland, and the United States. *J Fam Pract* 2002; 51: 72-3. Electronically available at:www.jfponline.com

[8] Laurie, B, Laurie, P. *Apache: The Definitive Guide,* Second Edition. Sebastopol: O'Reilly & Associates, 1999.

[9] Choi, W, Kent, A, Lea, C, Prasad, G, Ullman, C. *Beginning PHP4.* Birmingham: Wrox Press Ltd., 2000. (775 pages).

[10] DuBois, P. *MySQL.* Indianapolis: New Riders, 2000. (756 pages).

[11] w-Agora: web publishing and forum software. http://www.w-agora.net/.

[12] Graham, T. *Unicode: A Primer.* Foster City, CA: M&T Books, 2000. (475 pages).

[13] Unicode Consortium. *The Unicode Standard,* Version 4.0. Reading, MA: Addison-Wesley, 2003.

[14] Leisher, M. *JMUTT Java applet.* http://crl.NMSU.Edu/~mleisher/.

[15] Wolff, Thomas. *The mined 2000 text editor.* http://towo.net/mined.

[16] Lendon, C. *Simredo 3.4 - Java Unicode Editor.* http://www4.vc- net.ne.jp/~klivo/sim/simeng.htm.

[17] Levy, DH, Dolin, RH, Mattison, JE, Spackman, KA, Campbell, KE. Computer-Facilitated Collaboration: Experience Building SNOMED-RT. *Proc AMIA Annu Fall Symp.* AMIA, 1998: 870-874.

[18] Campbell, KE, Cohn, SP, Chute, CG, Shortliffe, EH, Rennels, G. Scalable methodologies for distributed development of logic-based convergent medical terminology. *Methods Inf Med* 1998; 37: 426-439.

[19] Kouzes, RT. Collaboratories: Scientists Working Together Apart. http://www.emsl.pnl.gov:2080/docs/collab/presentations/WorkingTogetherApart/world.html.

Address for correspondence

R. P. Channing Rodgers, MD
Office of High Performance Computing & Communications
U.S. National Library of Medicine
Bldg. 38, Room B1N-30D
8600 Rockville Pike
Bethesda MD 20894 USA
(301)435-3267 rodgers@nlm.nih.gov