

- ứng dụng điều khiển bàn chữ Việt trong môi trường Unix, DOS, MS-Windows 3.1. Ứng dụng này cho phép người dùng đánh chữ Việt trên bàn chữ ASCII Hoa Kỳ nhưng màn ảnh sẽ hiện ra chữ Việt Nam.
- ứng dụng màn ảnh Việt chạy trong khung X (*X windows*) hoặc khung MS-DOS (*MS-DOS windows*).
- ứng dụng thư tín cho phép đọc và viết thư bằng 8-bit VISCII nhưng sẽ tự động biến đổi thư thành dạng 7-bit VIQR khi lưu trữ hoặc gửi đi.
- ứng dụng in dữ liệu 8-bit VISCII trên các máy in Laser, ma-trận-điểm (*dot matrix*), hoặc PostScript.
- ứng dụng viết bài (*editor*) dùng 8-bit VISCII.
- các nhu-liệu dụng-cụ và nhu liệu thư viện (*library*) thường dùng trên khiên hệ Unix, DOS.
- ứng dụng bảng tính (*spreadsheet*) chạy trên Unix.
- ứng dụng sự phạm dùng để ra câu đố hoặc đề thi trắc nghiệm tiếng Việt.
- ứng dụng kiểm soát lỗi chính tả.
- ứng dụng trò chơi.
- các bộ phong chữ ở dạng điểm (*bitmap fonts*), TrueType, PostScript.

Ngoài ra còn nhiều sản phẩm nữa không tiện kể ra đây. Muốn biết chi tiết xin xem thông-báo về việc phát hành nhu-liệu Việt Nam đợt 3 [tr. 41]. Hiện chúng tôi đang phân công thiết kế thêm các bộ phong chữ (*fonts*) VISCII để sử dụng trên các máy in laser. Vì Viet-Std không có tính cách thương mại, tất cả các công trình nghiên cứu và sản phẩm nhu liệu của nhóm đều được phổ biến miễn phí. Một số công ty nhu liệu và tổ chức chuyên gia Việt Nam ở Hoa Kỳ đã tuyên bố ủng hộ bộ tự-mã 8-bit VISCII và quy-định 7-bit VIQR như VNU, TIẾN, Hội Chuyên Gia Việt Nam, v.v...

Bộ tự-mã VISCII và quy định VIQR là công trình nghiên cứu của đông đảo chuyên viên Việt Nam ở hải ngoại thuộc nhiều lãnh vực khác nhau và đã trải qua quá trình thử nghiệm khá lâu dài. Chúng tôi sẵn sàng gửi đến quý vị một số nhu liệu cần thiết để quý vị có thể tự mình chạy và quan sát ưu điểm của VISCII và VIQR một cách cụ thể. Xin quý vị nghiên cứu kỹ lưỡng tài liệu đính kèm và dành cho chúng tôi sự ủng hộ cần thiết để bộ Việt-tự-mã 8-bit VISCII và Quy định 7-bit VIQR trở thành tiêu chuẩn chính thức cho chữ Việt Nam.

Trân trọng kính chào quý vị và mong nhận được ý kiến của quý vị trong thời gian ngắn nhất.

Nhóm Nghiên Cứu Tiêu Chuẩn Tiếng Việt

California, USA  
September 1992

liệu hiện hữu ... Vì những lý do này, phương pháp dấu rời hầu như không còn được sử dụng trong kỹ nghệ điện toán hiện nay.

Để có thể hội nhập vào nền kỹ nghệ điện toán của thế giới, chúng ta phải chấp nhận giải pháp thứ hai, nghĩa là phải mã hóa một số lượng rất lớn mẫu tự Việt. Trừ một số chữ Việt đã có sẵn trong bộ tự-mã ASCII, chúng ta có tất cả 134 mẫu tự cần phải mã hóa trong khi chỉ có 128 mã số còn trống mà thôi. Cách giải quyết thông thường là tìm cách thay thế 6 mã-tự ASCII nào đó bằng 6 mẫu tự Việt. Có rất nhiều cách nhưng cách nào cũng vấp phải một số khuyết điểm riêng.

Sau khi phân tích kỹ càng vấn đề và cứu xét cả chiều hướng phát triển cương liệu và nhu liệu trong tương lai, nhóm Viet-Std đã giải quyết bài toán này dựa trên căn bản triệt để bảo toàn 96 ký tự ASCII trong vùng G0 (mang mã số 32 đến 127). Quyết định này được trình bày kỹ càng trong bài Anh Ngữ in lại trong tập này [xem tr. 6-22], nhưng trong phạm vi lá thư này có thể được tóm tắt như sau: bằng cách bảo toàn 96 ký tự ASCII chúng ta có thể sử dụng hầu hết nhu liệu và cương liệu sản xuất khắp nơi trên thế giới mà không phải đầu tư nhân lực và tài nguyên vào việc điều chỉnh hoặc biến đổi cho thích hợp với chữ Việt Nam, cụ thể như bộ dịch (*compiler*), khiên hệ (*operating system*), khung X (*X windows*), v.v... Chỉ có cách giải quyết này mới giúp chúng ta sử dụng được những thành quả kỹ thuật mới mẻ nhất trên thế giới.

Với phương châm trên, chúng tôi đã thay thế 6 kiễm-tự ASCII trong vùng C0 bằng 6 mẫu tự Việt Nam Ắ, Ằ, Ẵ, Ỡ, Ỡ và Ỡ. Ngoài ra, dựa trên bộ tự-mã tiêu chuẩn 8859/Latin-1 dành cho các nước Tây Âu, chúng tôi cũng quyết định duy trì mã số của tất cả chữ Việt đã có sẵn trong tiêu chuẩn này. Tất cả chi tiết về bộ tự-mã 8-bit VISCII được trình bày rõ ràng trong chương 3 của bài Anh ngữ [xem tr. 8-11] và bản dịch Việt ngữ [xem tr. 26-29].

Sau đây chúng tôi xin sơ lược về Quy-định Đọc-Được-Trong-Ngoặc, VIQR. Đây là quy luật viết chữ Việt Nam bằng bộ tự-mã 7-bit ASCII của Hoa-Kỳ. Đã từ lâu cộng đồng người Việt ở hải ngoại thường hay sử dụng hình thức này để trao đổi điện thư bằng tiếng Việt trên các máy vi tính không có chữ Việt Nam. Theo quy định này, các dấu được đánh sau các nguyên âm; dấu sắc, huyền, hỏi, ngã, nặng được thay thế bằng các ký hiệu ASCII Hoa-Kỳ có dạng tương tự là "'", "`", "?", "~", ". ", dấu trắng ( ) được thay bằng "(", dấu mũ bằng "^", dấu móc (') bằng "+", và chữ đ được thay thế bằng "dd". Chẳng hạn, hai câu thơ Kiều của cụ Nguyễn Du:

Trăm năm trong cõi người ta  
Chữ tài chữ mệnh khéo là ghét nhau

sẽ hiện ra như sau khi viết theo quy định VIQR:

Tra(m na(m trong co~i ngu+o+`i ta  
Chu+` ta~i chu+` me^nh khe`o la` ghe`t nhau

Tuy nhiên một số người lại thích dùng các ký hiệu khác, chẳng hạn như "\*" tượng trưng cho dấu móc (^), "<" thay cho dấu trắng ( ), "\" thay cho dấu huyền, v.v... Nhưng với vai trò là một tiêu chuẩn, VIQR phải ấn định một quy luật duy nhất và tối thiểu để làm cơ sở thống nhất cho việc trao đổi nhu liệu, nghĩa là mỗi dấu Việt Nam phải được tượng trưng bằng một và chỉ một ký tự ASCII mà thôi. Với sự sử dụng rộng rãi bản đánh chữ ASCII trong giới điện toán Việt Nam, chúng tôi quyết định chọn quy định VIQR dựa trên nguyên tắc thực dụng là dễ đọc và dễ nhớ cho tuyệt đại đa số người dùng. Quy luật VIQR được mô tả chi tiết trong chương 4 của tài liệu Anh Ngữ và Việt ngữ đính kèm.

Tiện đây chúng tôi cũng xin nhấn mạnh là bộ Việt-tự-mã 8-bit VISCII và quy định 7-bit VIQR đã được phổ biến rộng rãi trên các mạng thông tin điện toán (*computer network*) ở Hoa Kỳ và các quốc gia tiên tiến khác trên toàn thế giới. Các chuyên viên điện toán Việt Nam ở hải ngoại đã dùng những tiêu chuẩn này để viết nhu liệu ứng-dụng (*software application*) hoặc nhu-liệu dụng-cụ (*software tool*) cho khiên hệ DOS và Unix. Những nhu liệu này đã được cả người viết lẫn người dùng trải nghiệm trên các máy như PC, AT, 386/486, workstation, mainframe, v.v... trong một thời gian khá lâu dài và thực tế cho thấy tất cả đều vận hành tốt đẹp với dữ liệu Việt ngữ. Trong phạm vi lá thư này, chúng tôi xin tóm tắt những sản phẩm sau đây:

- chương trình biến đổi dữ liệu ở dạng 7-bit VIQR sang dạng 8-bit VISCII và ngược lại.

# Thư Ngỏ Gửi Quý-Vị Trong Ngành Điện-Toán

## Cùng Tất-Cả Quý-Vị Quan-Tâm

### Đến Việc Mã-Hóa Chữ Việt-Nam

Kính thưa quý vị:

Chúng tôi là một nhóm chuyên viên Việt Nam ở hải ngoại cộng tác trong tinh thần vô vụ lợi để theo dõi và đóng góp ý kiến chuyên môn về chữ Việt Nam với các Viện Định Chuẩn Tin Học hoặc các công ty điện toán có tầm vóc quốc tế. Trong thời gian qua chúng tôi đã vận động tích cực với tổ chức Unicode và Viện Định Chuẩn ISO để họ tiêu chuẩn hóa bộ Việt-tự-mã (*Vietnamese character encoding*) trong khuôn khổ bộ mã chữ quốc tế 16-bit và 32-bit trong chiều hướng có lợi nhất. Ngoài ra, chúng tôi đã nghiên cứu và đề nghị bộ Việt-tự-mã 7-bit và 8-bit để giải quyết nhu cầu phát triển và trao đổi nhu liệu hiện nay. Bộ tự-mã 8-bit có tên Anh ngữ là *Vietnamese Standard Code for Information Interchange* hay gọi tắt là VISCI để phân biệt với các bộ tự-mã khác. Tiêu chuẩn 7-bit được gọi là Quy-tắc Đọc-Được-Trong-Ngoặc (*Vietnamese Quoted-Readable Specification*) hay gọi tắt là VIQR. Đây là đề tài chính mà chúng tôi muốn thảo luận với quý vị qua lá thư này.

Trước hết chúng tôi xin trình bày về bộ tự-mã 8-bit VISCI. Điềm qua quá trình phát triển ngành điện toán dùng chữ Việt ở hải ngoại, chúng tôi nhận thấy hầu như mỗi công ty hay mỗi nhóm tự đặt ra cho mình một bộ Việt-tự-mã, với hệ quả tất nhiên là nhu liệu không thể trao đổi dễ dàng với nhau. Do đó chúng tôi đã quyết định nghiên cứu một bộ Việt-tự-mã tiêu chuẩn (*Vietnamese character encoding standard*) nhằm giải quyết tình trạng này.

Như quý vị đã rõ, đa số nhu liệu hiện nay được viết dựa trên nền tảng mỗi mẫu tự được mã hóa bằng 8-bit (1 byte). Với 8-bit chúng ta có thể mã hóa được 256 mẫu tự hoặc tín hiệu khác nhau. Vì những lý do có tính cách lịch sử, bộ tự-mã ASCII của Hoa kỳ dùng 128 mã số (*code point or code value*) đầu tiên đã trở thành tiêu chuẩn quốc tế. Bộ tự-mã này bao gồm 32 tín hiệu điều khiển (*control character*) có mã số từ 0 đến 31, và 96 mã số còn lại dành cho một số mẫu tự La tinh, dấu chấm câu hoặc ký hiệu. (Để tiện việc thảo luận, chúng tôi tạm dịch *control character* là kiểm-tự, các chữ còn lại là ký-tự. Đứng trên quan điểm điện toán, chúng ta phải định nghĩa thêm mã-tự là bất cứ cái gì được tượng trưng bằng một mã số; mỗi mã-tự tương ứng với một mã số và ngược lại.) Phần còn lại (128 mã số từ 128 đến 255) thường được quy định và sử dụng tùy theo nhu cầu của mỗi quốc gia hoặc từng bộ nhu liệu. Như vậy chúng ta có thể tiêu chuẩn hóa 128 mã số này cho các mẫu tự Việt Nam không nằm trong danh sách 128 chữ ASCII.

Trên thực tế việc mã hóa chữ Việt là cả một vấn đề phức tạp. Ngoài các phụ âm, chúng ta có 12 nguyên âm chính (A Ă Æ E Ê I O Ô O U U Y) và 60 nguyên âm khác được tạo thành do các nguyên âm chính kết hợp với 5 dấu giọng (sắc, huyền, hỏi, ngã, nặng). Như vậy chữ Việt có tất cả 144 nguyên âm vừa thường vừa hoa. Chúng ta có thể kể ra hai phương pháp chính để mã hóa nguyên âm Việt Nam như sau:

1. Mỗi nguyên âm chính và mỗi dấu giọng được xem như các mã-tự riêng biệt, nghĩa là chỉ cần 29 mã số để mã hóa 12 nguyên âm chữ thường, 12 nguyên âm chữ hoa và 5 dấu giọng. Thí dụ, mẫu tự Æ được xem như gồm có hai mã-tự "A" và mã-tự " `".
2. Mỗi nguyên âm, và dấu giọng nếu có, được xem như một mẫu tự duy nhất (mã-tự), nghĩa là phải cần đến 144 mã số để mã hóa tất cả nguyên âm Việt.

Phương pháp đầu tiên, còn được gọi là phương pháp dấu rời, đã được áp dụng ở nhiều quốc gia dùng chữ La-tinh có dấu phụ (*diacritical mark*). Nhưng quá trình thử nghiệm cho thấy phương pháp này có khuyết điểm lớn lao về nhiều mặt như tốc độ xử lý giảm sút, nhu cầu về sức chứa (*storage*) và bộ nhớ (*memory*) gia tăng, việc thảo luận phức tạp, không thể tích hợp vào môi trường nhu liệu và cương

It should be mentioned that TriChlor Software (a non-profit experimental group) was formed by Cường Tấn Nguyễn, Cương Minh Bùi, and Tín Lê to independently explore and implement any encoding scheme for Vietnamese to gain real experience with the pros and cons in each scheme. It has helped integrate VISCII-compliant Vietnamese into popular public domain software products.

It is our dream one day to be able to read, write, and exchange Vietnamese data of a common format on any machine, any platform, and to take advantage of all the processing tools that have been produced by the computing world. That dream, once a pure exercise in imagination, has today come many steps closer to realization.

Viet-Std Vietnamese Standardization Group

California, USA  
September 1992

## Preface

The Vietnamese Standardization Group (Viet-Std) was formed in the fall of 1989 to promote the standardization of Vietnamese character encoding and to monitor ongoing work of international bodies in this regard. The group has been working on designing and implementing a code table that can be integrated into existing computing environments on many platforms. In addition, the group has contacted the two committees on multilingual character encoding—the Unicode Consortium and the International Standardizations Organization (ISO)—to request that the Vietnamese characters be encoded in a pre-composed form in the same manner as other Latin-based European languages. All these efforts will be reported fully in a special issue to be published in the near future. This special report collects only works concerning the 7- and 8-bit encodings of the Vietnamese alphabet.

The first article is a cover letter in Vietnamese that summarizes the Viet-Std Group’s work in the area of 7- and 8-bit Vietnamese character encoding.

The second article covers the Vietnamese character encoding in 7 and 8 bits. It reviews the pros and cons of current encoding schemes and discusses the vital need to integrate into existing computing environments that a standard must address. It then presents the Viet-Std’s 8-bit proposal for the Vietnamese Standard Code for Information Interchange (known as VISCII) and a Vietnamese Quoted-Readable convention (VIQR) to represent Vietnamese characters in 7-bit ASCII. The article also examines some guidelines and conventions in handling Vietnamese electronic mail over 7-bit channels, Vietnamese keyboarding, and adapting existing Vietnamese applications. It includes two appendices listing both VISCII and VIQR for reference purposes.

The third article is the Vietnamese translation of the above to serve the general Vietnamese community.

The last item included in this report is the announcement of the third release of public-domain Vietnamese software by the TriChlor Group. It summarizes software packages either developed or enhanced by TriChlor members and other independent developers to run on the Unix, X-Windows, DOS, and MS-Windows platforms.

Since the release of version 1.0 in January 1992, VISCII has undergone only one change—swapping the two characters a (a dot-below) and o (o tilde)—to accommodate MS-Windows. To reflect this change, the versions of VISCII and VIQR published in this report are called VISCII 1.1 and VIQR 1.1 although VIQR was unchanged. This report therefore supercedes all previous publications concerning VISCII and VIQR.

A brief history of the Vietnamese Standardization Group is in order. The group was born out of the Viet-Net electronic mailing list, which comprised members at companies and universities throughout the computing world. At the time, Vietnamese software for publishing existed commercially only on the personal computer platform, but each package was limited to its designed function and could not be easily intermixed with others. No standard encoding existed. Developers designed their own schemes to perform the job they needed. Even Viet-Net itself had invented a 7-bit mnemonic writing style for Vietnamese for use in e-mail. For example, “Nu<sup>o</sup>c Vi<sup>e</sup>t Nam” would be typed as “Nu+o+`c Vie`.t Nam.” Everyone agreed that some form of standardization was necessary so as to promote availability of and access to Vietnamese computing, and Viet-Net provided the forum that electronically brought together for the first time a large number of individuals committed to this goal. With this specific interest in mind, Viet-Std was formed by Thành Văn Nguyễn, and standardization discussions moved to the mailing list “Viet-Std@images.Sun.COM” and later to “Viet-Std@Haydn.Stanford.EDU.” Early contributors included Cường Tấn Nguyễn, Tâm Nguyễn, Nhân Trần, Randall Atkinson, Khoa Tôn, Khiêm Hồ, Tước Lương, and many others too numerous to acknowledge properly. Later joining the group were Cương Minh Bùi, Học Đình Ngô, and others. This special report is a testimony to the success of the group. In addition, the group has contributed its expertise to international organizations on matters relevant to Vietnamese encoding; the details will be reported in a separate issue.

<b>TÓM LƯỢC</b>	<b>23</b>
<b>1 LỜI GIỚI THIỆU</b>	<b>23</b>
<b>2 DUYỆT LẠI NHỮNG QUY ƯỚC HIỆN THỜI</b>	<b>24</b>
<b>3 VISCII: QUY ĐỊNH MÃ 8-BIT CHO VIỆT NGỮ</b>	<b>26</b>
3.1 ĐỘNG LỰC . . . . .	26
3.2 CÁC LÝ DO BIẾN MINH VIỆC MÃ HÓA . . . . .	26
<b>4 VIQR: QUY ĐỊNH VIỆT NGỮ ĐỌC-ĐƯỢC-TRONG-NGOẶC</b>	<b>29</b>
4.1 ĐỘNG LỰC . . . . .	29
4.2 QUY ĐỊNH “ĐỌC-ĐƯỢC-TRONG-NGOẶC” (VIQR) . . . . .	30
4.2.1 Phép Tạo Chữ Ngầm . . . . .	30
4.2.2 Phép Tạo Chữ Chỉ Định . . . . .	31
4.2.3 Trạng Thái Nguyên Dạng . . . . .	31
4.2.4 Trạng Thái Anh Ngữ . . . . .	31
4.2.5 Trạng Thái Việt Ngữ . . . . .	31
4.2.6 Nguyên Tụ trong Trạng Thái Anh Ngữ và Việt Ngữ . . . . .	32
4.2.7 Ký Tụ Hoàn Cấu . . . . .	32
<b>5 CÁC ỨNG DỤNG ĐẶC BIỆT</b>	<b>32</b>
5.1 ĐIỆN THƯ CHUYỂN QUA MẠCH 7-BIT . . . . .	32
5.2 ĐÁNH CHỮ VIỆT . . . . .	32
5.2.1 Cách Tạo Hình Lập Túc Trong Phép Tạo Chữ Ngầm . . . . .	33
5.2.2 Cách Tạo Hình Chậm Trong Phép Tạo Chữ Chỉ Định . . . . .	33
5.3 HỢP THỨC HÓA ỨNG DỤNG VIỆT NGỮ HIỆN HÀNH . . . . .	34
<b>6 TÓM TẮT &amp; KẾT LUẬN</b>	<b>34</b>
<b>TÀI LIỆU THAM KHẢO</b>	<b>34</b>
<b>THUẬT-NGỮ ANH VIỆT</b>	<b>34</b>
<b>PHỤ LỤC A: Mẫu Tụ Việt Liệt Kê theo Thứ Tụ Sắp Chữ</b>	<b>38</b>
<b>PHỤ LỤC B: Mẫu Tụ Việt Liệt Kê theo Thứ Tụ Mã Số</b>	<b>39</b>
<b>Announcement of VISCII-Compliant Software Applications</b>	<b>41</b>

## Contents

Preface	1
Thư Ngõ Gửi Quý Vị Trong Ngành Điện Toán	3
A Unified Framework for Vietnamese Information Processing	6
<b>ABSTRACT</b>	<b>6</b>
<b>1 INTRODUCTION</b>	<b>6</b>
<b>2 REVIEW OF CURRENT CONVENTIONS</b>	<b>7</b>
<b>3 VISCII: 8-BIT ENCODING SPECIFICATION FOR VIETNAMESE</b>	<b>8</b>
3.1 MOTIVATION . . . . .	8
3.2 ENCODING RATIONALE . . . . .	9
<b>4 VIQR: MNEMONIC ENCODING SPECIFICATION FOR VIETNAMESE</b>	<b>11</b>
4.1 MOTIVATION . . . . .	11
4.2 QUOTED-READABLE SPECIFICATION (VIQR) . . . . .	13
4.2.1 Implicit Composition . . . . .	14
4.2.2 Explicit Composition . . . . .	14
4.2.3 Literal State . . . . .	14
4.2.4 English State . . . . .	14
4.2.5 Vietnamese State . . . . .	15
4.2.6 Character Literals in English and Vietnamese States . . . . .	15
4.2.7 Closure . . . . .	15
<b>5 SPECIFIC APPLICATIONS</b>	<b>15</b>
5.1 ELECTRONIC MAIL OVER 7-BIT CHANNELS . . . . .	15
5.2 VIETNAMESE KEYBOARDING . . . . .	16
5.2.1 Immediate Echo for Implicit Composition . . . . .	16
5.2.2 Delayed Echo for Explicit Composition . . . . .	16
5.3 ADAPTING EXISTING VIETNAMESE APPLICATIONS . . . . .	17
<b>6 SUMMARY &amp; CONCLUSIONS</b>	<b>17</b>
<b>REFERENCES</b>	<b>17</b>
<b>GLOSSARY OF TERMS</b>	<b>17</b>
<b>APPENDIX A: Vietnamese Characters under VISCII and VIQR by Collating Order</b>	<b>20</b>
<b>APPENDIX B: Vietnamese Characters under VISCII and VIQR by Encoding Order</b>	<b>21</b>

© Copyright 1992, Viet-Std. All rights reserved. Permission is hereby granted for unlimited duplication and dissemination provided no modifications are made. This copyright notice must accompany any duplication, in part or in whole.

For further information, please contact Viet-Std at 1212 Somerset Drive, San Jose, CA 95132, or via electronic mail at [Viet-Std@Haydn.Stanford.EDU](mailto:Viet-Std@Haydn.Stanford.EDU).



# **VIETNAMESE CHARACTER ENCODING STANDARDIZATION REPORT**



## **VISCII and VIQR 1.1 Character Encoding Specifications**

---

**Viet-Std, September 1992**