

**Số 5 ▪ Tháng 7/2005**

---

**Quy trình Nôm Na:  
“Giúp đọc Nôm và Hán Việt”  
và chữ Nôm trên mạng**

**Nhóm Nôm Na** (*Hội Bảo tồn Di sản chữ Nôm*)  
Lê Văn Cường, Tô Trọng Đức, Ngô Thanh Giang, Lương Thị Hạnh  
Ngô Thanh Nhân, Lê Mai Phương, Ngô Trung Việt

**Hội nghị Quốc tế về chữ Nôm**  
Ngày 12-13 tháng 11 năm 2004  
Thư viện Quốc gia Việt Nam – Hà Nội

**Tóm tắt**

Thông tin không những xảy ra tức thời, trực diện truyền miệng qua ngôn ngữ, mà còn có thể xảy ra xuyên thời gian và không gian qua chữ viết và in ấn (cuộc cách mạng thông tin trong quá khứ). Ngày nay, thông tin có thể xảy ra tức thời xuyên không gian nhờ cuộc cách mạng thông tin qua máy tính và mạng web. Máy tính, mạng web, và chuẩn mã đa ngữ quốc tế, là con đường duy nhất để bảo tồn chữ Nôm, vốn văn hoá của dân tộc Việt Nam. Quy trình Nôm Na được đặt ra nhằm tìm ra quy trình tốt nhất và đơn giản nhất cho mục tiêu này.

Trong thời gian từ tháng 6/2002 đến tháng 8/2004, nhóm Nôm Na, bao gồm 4 chuyên viên trẻ, đã triển khai việc nghiên cứu, phân tích, tạo phong và xây dựng chế bản cuốn *Giúp đọc Nôm và Hán Việt* của Linh mục Trần Văn Kiệm để xuất bản thành sách và sử dụng tra cứu trên mạng.

Một khối lượng công việc lớn đã được thực hiện trong thời gian này:

- Tạo phong cho khoảng 4.415 thành tố Hán–Nôm cơ bản; vẽ phong 17.761 chữ Hán–Nôm;
- Xây dựng và quản lý kho chữ Hán–Nôm cho cuốn *Giúp đọc Nôm và Hán Việt*;

- Đối chiếu và thiết lập các mã Unicode cho các chữ đã xây dựng, kể cả việc tạo mã thay thế (*surrogate*) cho những chữ nằm ngoài mặt phẳng đa ngữ cơ bản BMP (*Base Multilingual Plane*) của Unicode và ISO/IEC 10646;
- Tạo định dạng và chuyển đổi quyển *Giúp đọc Nôm và Hán Việt* theo chuẩn đa ngữ HTML để làm chế bản và để sử dụng trên mạng internet cho cuốn *Giúp đọc Nôm và Hán Việt*.

Bài này trình bày ý nghĩa các quy trình mà nhóm Nôm Na đã thực hiện trong thời gian qua. Trên cơ sở đó đúc kết và khái quát hoá để có thể áp dụng vào những hoạt động nghiên cứu và ứng dụng tiếp theo: Quy trình chế tạo các bộ phong theo các thể khác nhau; quy trình đưa các văn bản Hán–Nôm vào máy tính; và quy trình chuẩn hoá các chữ Hán–Nôm và mở rộng kho chữ Hán–Nôm để sử dụng rộng rãi.



Thông tin tự nhiên của con người xảy ra tức thời, trực tiếp qua ngôn ngữ. Mỗi người sử dụng ngôn ngữ theo bản năng, truyền qua không gian từ miệng người này sang tai người khác, đồng thời theo độ nhanh của âm thanh, và quãng cách xa nhất mà độ vang của tiếng có thể chờ được. Thời gian, không gian và ngôn ngữ là ba giới hạn chính của thông tin trong tương tác giữa con người. Con người luôn luôn tìm cách tăng độ xa không gian và tăng độ dài thời gian trong việc trao đổi thông tin nhờ trí nhớ, phiên dịch và tập thể truyền khẩu. Con người cũng luôn luôn tìm công cụ để thay đổi ba biến thiên này. Hai cuộc cách mạng thông tin trong quá khứ là chữ viết và in ấn.

Đặc điểm thứ hai của thông tin giữa con người là khả năng chuyên chở tri thức. Tri thức nhờ ngôn ngữ chuyển từ người này sang người khác. Chữ viết cũng thế. Chữ viết không phải là những hình vẽ đơn thuần, mà nó là công cụ để chuyển giao tri thức xuyên không gian và thời gian. Do đó, người đọc (hay người nghe) phải lấy được thông tin trong các văn bản.

Ngày nay, thông tin có thể xảy ra tức thời xuyên không gian nhờ cuộc cách mạng thông tin qua máy tính và mạng web. Chúng ta sử dụng các công cụ hiện đại, không chỉ chụp ảnh giữ các văn bản cũ mà còn để rút thông tin trong các văn bản ấy. Đây là vấn đề đặt ra trong việc phục hồi và phổ biến sử dụng chữ Nôm. Chữ Nôm đã là chữ quốc ngữ chính của dân tộc Việt Nam trong nhiều thế kỉ trước khi bị chữ la-tinh quốc ngữ thay thế vào đầu những năm 1920. Máy tính, mạng web, và chuẩn mã đa ngữ quốc tế, là con đường duy nhất để bảo tồn chữ Nôm, vốn văn hoá của dân tộc Việt Nam. Quy trình Nôm Na được đặt ra nhằm tìm ra quy trình tốt nhất và đơn giản nhất cho mục tiêu này.

## 1. Giới thiệu khái quát

Nhóm Nôm Na chính thức được tổ chức vào tháng 6 năm 2002, trực thuộc *Hội Bảo tồn Di sản chữ Nôm*. Nhóm bao gồm những chuyên viên trẻ mới ra trường với các chuyên môn về Hán–Nôm và công nghệ thông tin.

Trong thời gian vừa qua, nhóm đã triển khai công việc nghiên cứu, phân tích, tạo phông (*font*), gán mã quốc tế, và xây dựng chế bản cho cuốn *Giúp đọc Nôm và Hán Việt* (viết tắt *Giúp đọc*)<sup>1</sup> để xuất bản thành sách và tra cứu trên mạng.

Cuốn *Giúp đọc* là một công trình nhằm giúp cho độc giả tra cứu và đọc các chữ Nôm, thành quả trong nhiều năm tận tụy làm việc của Linh mục Anthony Trần Văn Kiệm. *Giúp đọc* đã được Nhà xuất bản Thuận Hoá in năm 1998 với phần chữ quốc ngữ và phan âm (*pinyin*) trên phông VNI-Times, cỡ 12, trên trang 8,5” x 11”, lúc in thu nhỏ lại thành 5,25” x 8”. Phần chữ Nôm và Hán-Việt lúc đó viết tay. Do vậy, công việc của nhóm Nôm Na là xây dựng quy trình Nôm Na nhằm mang lại diện mạo mới cho *Giúp đọc*, tạo ra các sách có chữ Nôm trên máy tính và in ra được theo kiểu thông thường. Tổng cộng cuốn sách 920 trang.

Để xây dựng quy trình Nôm Na, *Giúp đọc*, và chữ Nôm trên mạng được thuận lợi, việc sử dụng riêng thông tin trong *Giúp đọc* vẫn chưa đủ. Chúng tôi đã tham khảo thêm một số tự điển và từ điển chữ Nôm như *Bảng tra chữ Nôm*<sup>2</sup>, *Tự điển chữ Nôm*<sup>3</sup>, *Đại tự điển chữ Nôm*<sup>4</sup>, *Lời dẫn Tự điển chữ Nôm tiếng Việt*<sup>5</sup>, *Bảng phiên âm Nôm Việt*<sup>6</sup>.

Ngoài ra chúng tôi còn tham khảo thêm các tự điển như *Nam Việt—Duong Hiệp Tự vị* của J. L. Taberd (1838), *Từ vựng lịch sử chữ Nôm* [Dictionnaire historique des idéogrammes vietnamiens] của Paul Schneider (1992), *Đại Nam Quốc Âm Tự vị* của Paulus Huỳnh Tịnh Của (1895), v.v.

Những vốn quý giá về chữ Nôm này chứng tỏ sức sống mãnh liệt của chữ Nôm trong quá khứ, đã tồn tại trên bình diện quốc tế, nhưng cũng cho thấy những vấn đề còn lại trong việc phổ biến sử dụng chữ Nôm hiện nay. Đó là việc cần phải tiếp tục quốc tế hoá chữ Nôm trên mô thức mới, với sự hỗ trợ của kỹ thuật hiện đại, đặc biệt là việc đưa chữ Nôm vào máy tính và mạng máy tính.

## 2. Nghiên cứu ban đầu và định hướng công việc

Mục tiêu đặt ra khi nhóm bắt tay vào công việc này là tận dụng phần nội dung chữ quốc ngữ đã có sẵn, chuyển sang dạng chuẩn chữ quốc ngữ Unicode, vẽ phông và lập mã quốc tế hoặc mã Việt Nam cho các chữ Nôm có trong sách, tổ chức cơ sở dữ liệu về kho chữ Hán–Nôm trong sách, làm chế bản cho sách để có thể in thành sách, và sử dụng được trên mạng máy tính.

Do đó định hướng công việc cần được thực hiện là: nghiên cứu cấu trúc của cuốn *Giúp đọc* để xác định việc cần làm, và nghiên cứu các công cụ phần mềm hiện có trên thị trường để thực hiện các công việc này.

### 2.1. Cấu trúc quyển *Giúp đọc*

Cấu trúc cuốn sách bao gồm hai dạng chữ: dạng chữ Hán–Nôm và dạng chú giải quốc ngữ. Cuốn sách được chia thành 2 phần, tìm âm và tìm nghĩa. Do đó các xử lý cần tính tới để thực hiện cuốn sách này là:

- Xử lý chuyển đổi phần chữ quốc ngữ từ VNI sang mã chuẩn Unicode.
- Tạo phông chữ Nôm, hình thành kho chữ theo chuẩn Unicode.
- Làm chế bản, ghép chữ Nôm vào văn bản quốc ngữ của *Giúp đọc*.
- Quản lý kho chữ đã tạo ra.

## 2.2. Chọn công cụ phần mềm

Trên thị trường lúc đó các phần mềm cho phép xử lý chữ biểu ý là Twinbridge, FontLab, Fontographer, Microsoft Office, Microsoft Volt. Vấn đề là chọn phần mềm nào thích hợp nhất cho sự phát triển lâu dài của công việc với chữ Nôm.

— Twinbridge hay *Song Kiều* là phần mềm của Công ty Twinbridge Software Corporation, có khả năng thích ứng với một số phần mềm khác, như Microsoft Office hệ điều hành Microsoft Windows, là công cụ đắc lực trong việc in ấn và chế bản chữ Hán của giới Nôm học trong những năm gần đây. Trong Twinbridge còn có một chương trình tạo chữ mới chưa có sẵn trong hệ thống, một tiện ích thường được các nhà Nôm học sử dụng để tạo chữ Nôm hay chữ Hán trong các văn bản cổ của chúng ta mà kho chữ trong phần mềm này không đáp ứng được.

Với chương trình tiện ích tạo chữ này, người dùng bộ các nét chuẩn và dùng chương trình vẽ chữ có trong TwinBridge để vẽ chữ mới. Theo cách này sẽ có một công cụ vẽ rất thuận tiện, nhưng nhược điểm của hệ thống là chỉ cho phép lưu giữ tối đa 1.000 chữ mới. Những chữ này có thể chép ra từ TwinBridge theo tệp riêng dưới dạng *bitmap* 128x128 điểm nhưng không tạo được phong loại TrueType độc lập với TwinBridge, và sau khi chép ra, không thể nhập lại trong hệ thống của TwinBridge.

Nhược điểm chính của Twinbridge là chúng ta không tạo ra được chữ Nôm mang đặc tính riêng của chữ Nôm mà mang đặc tính của chữ Hán do TwinBridge qui định. Và dù chữ có được tạo ra thì cuối cùng vẫn phải phụ thuộc vào TwinBridge, lại có bất tiện là phải dùng cùng một mã để mã hoá cho các chữ khác nhau, nếu người sử dụng có số chữ mới trên giới hạn 1.000. Chương trình này chưa được cập nhật với các thay đổi về hệ điều hành, do đó không có khả năng sử dụng với Windows XP nếu không có các bản nâng cấp mới.

— Fontlab là phần mềm của Công ty FontLab Ltd., phiên bản 3.1, được chúng tôi lựa chọn sau khi đã làm thí điểm cho việc tạo phong trên TwinBridge và thấy không đáp ứng được yêu cầu phát triển lâu dài cho cả kho chữ Nôm. Fontlab có thể thích ứng được với nhiều phần mềm mới hiện nay như Windows 2000 và sau đó: Microsoft Excel, Microsoft Word, Volt... Bản thân Fontlab có thể lưu giữ được hàng chục nghìn kí tự và từ đó có thể tạo phong TrueType độc lập, hiển thị kí tự trên nền Windows và HTML (mạng internet). Chính vì vậy, chúng tôi đã chọn Fontlab làm công cụ chính cho quy trình tạo phong.

— Volt là phần mềm của Microsoft cho phép hỗ trợ mã hóa các chữ theo chuẩn Unicode nếu các chữ đó nằm ngoài mặt phẳng đa ngữ cơ bản (*Base Multilingual Plane*, BMP) sử dụng kĩ thuật mã hoá thay thế (*surrogate*).

Ngoài Volt và Fontlab chúng tôi còn sử dụng một số những phần mềm văn phòng: Microsoft Excel, Microsoft Access và Microsoft Word; các công cụ lập trình mạng như PHP (*PHP: Hypertext Preprocessor*), Perl (*Practical Extraction and Report Language*), v.v. nhằm phục vụ cho việc quản lý hệ thống.

## 3. Quá trình làm việc

### 3.1. Vẽ các thành tố Hán-Nôm cơ bản (6/2002–12/2002)

Thành tố là những bộ phận có nghĩa, có cùng hình dáng, xuất hiện đều đặn, lặp đi lặp lại, ở nhiều chữ Hán-Nôm, có khi là những nét, những bộ thủ. Ta gọi chúng là thành tố

mô tả cơ bản (hay thành tố cấu tạo cơ bản, hay **tiền mẫu tự**) để tạo nên toàn bộ kho chữ Hán–Nôm. Trong thời gian từ 6/2002–12/2002 chúng tôi đã tập trung vẽ 4.415 thành tố Hán–Nôm cơ bản theo các nét chữ Nôm trong *Thiên tông bản hạnh*<sup>7</sup>, một lối chữ theo kiểu Tống thể được khắc in tại Việt Nam vào khoảng 1933. Các thành tố Hán–Nôm này dùng danh sách do Đỗ Quốc Bảo giới thiệu<sup>8</sup>, vẽ theo *thể Thiên tông bản hạnh* với để có thể ráp thành các chữ Hán–Nôm với phong thái Việt Nam.

### 3.2. Xây dựng cơ sở dữ liệu chữ Nôm (csdl) và vẽ thể chữ Nôm cho Phần II của *Giúp đọc* (12/2002–9/2003)

Sau khi công đoạn vẽ các thành tố Hán–Nôm cơ bản đã hoàn tất, công việc xây dựng cơ sở dữ liệu chữ Nôm bắt đầu bằng việc tạo phong cho Phần II của *Giúp đọc*. Chúng tôi đã tiến hành làm song song hai việc chính: lập cơ sở dữ liệu chữ Hán–Nôm và vẽ chữ.

### 3.3. Gắn chữ Nôm vào chế bản: Phần I và Phần II theo trật tự *Giúp đọc* (9/2003–2/2004)

Sau khi hoàn chỉnh bộ phong cho *Giúp đọc*, đặc biệt là cơ sở dữ liệu chữ Hán–Nôm và bộ chữ, chúng tôi tiến hành đưa chữ Hán–Nôm vào làm chế bản cho *Giúp đọc*. Công đoạn này được chia làm hai bước.

Bước 1: đưa chữ Hán–Nôm vào Phần II – phần **Tìm nghĩa**, lần lượt theo từng vần và từng mục từ trong mỗi vần từ A cho đến Y.

Bước 2: đưa chữ Hán–Nôm vào Phần I – phần **Tìm âm**, từ bộ một nét (bộ ất) cho đến bộ 14 nét (bộ ị) và phần các Bộ gốc.

Cả hai bước trên muốn tiến hành được đều phải dựa vào cơ sở dữ liệu chữ Nôm hoàn chỉnh song song.

### 3.4. Làm 6 lần chế bản v.1–v.6 (2/2004–5/2004)

Làm chế bản là công đoạn cuối cùng để chuẩn bị cho việc in ấn và xuất bản. Sau khi đã đưa chữ Hán–Nôm vào *Giúp đọc*, chúng tôi tiến hành chỉnh sửa: Dàn trang, dẫn trang, sửa theo các phiên bản, sửa bản bông để đưa đi xuất bản. Công đoạn này có nhờ các chuyên gia của *Trung tâm Từ điển học* phụ giúp kiểm tra thêm.

### 3.5. Làm phong mã hoá thay thế, đồng bộ hoá và chỉnh sửa csdl (5–8/2004)

Các công việc được tiến hành song song với nhiều tệp dữ liệu phát sinh và sửa đổi, không thể tránh được các lỗi nảy sinh. Do vậy việc đồng bộ hóa dữ liệu là cần thiết giữa cơ sở dữ liệu, bộ phong chữ và cuốn *Giúp đọc*.

Làm phong với cơ chế thay thế (*surrogate*) là bước tiếp theo sau khi đồng bộ hóa để đưa ra một bộ phong hoàn chỉnh tuân thủ chuẩn quốc tế Unicode, và để sử dụng rộng rãi cho các ứng dụng máy tính và mạng.

### 3.6. Chuyển *Giúp đọc* sang html dùng trên mạng (Phần II) (9/2004)

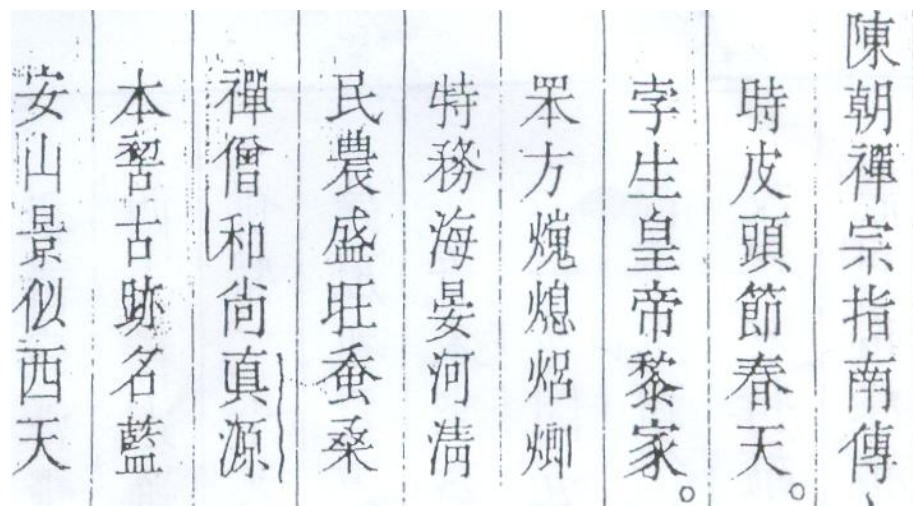
Sau khi *Giúp đọc* đã hoàn chỉnh, và việc in ấn đã hoàn tất. Chúng tôi tiến hành chuyển *Giúp đọc* sang dạng html (*Hypertext Markup Language*), chuẩn mạng của Mạng toàn cầu (*Worldwide Web*) dùng trên mạng để tra cứu rộng rãi.

#### 4. Các sản phẩm đã thực hiện

Trong khoảng thời gian 2 năm, với sự nỗ lực của các thành viên trong nhóm Nôm Na và sự giúp đỡ của một số chuyên gia công nghệ thông tin và chuyên gia Hán-Nôm, chúng tôi đã hoàn tất những sản phẩm sau:

##### 4.1. Bộ các thành tố vẽ chữ

Vẽ toàn bộ 4.415 thành tố Hán-Nôm cơ bản theo phong cách chữ Nôm trong *Thiền Tông Bản Hạnh*. *Thiền tông bản hạnh* có tên đầy đủ là *Trần Triều Thiền tông Chỉ nam Truyền tâm Quốc ngữ Hạnh*, Trúc Lâm Yên Tử, do Hòa Thượng Chân Nguyên tức Tuệ Đăng viết. Đây là tác phẩm chữ Nôm ra đời vào cuối thế kỉ 17. Văn bản mà chúng tôi sử dụng cho việc tạo phong, khắc in lại vào khoảng năm 1933 theo lối Tổng thể, dáng chữ mềm mại thanh thoát cân đối và mang những nét đặc trưng của chữ Nôm Việt Nam.



Bảng 1: Ảnh chữ Nôm trong “*Thiền Tông Bản Hạnh*”.

Kiểu phong Nôm Na dựa vào phong cách chữ Nôm trong “*Thiền Tông Bản Hạnh*” để tạo dáng chữ trong máy tính. Đây là đặc trưng bộ phong của nhóm Nôm Na.



Bảng 2: Chữ Nôm của bộ phong NomNaTongLight.ttf

*Thành tố Hán–Nôm cơ bản dùng vẽ chữ:* Với sự giúp đỡ của Đỗ Quốc Bảo, chúng tôi đã dùng 4.415 thành tố Hán–Nôm cơ bản để vẽ bộ phong Nôm Na Tổng thể, bước đầu dùng cho *Giúp đọc* và sau đó sẽ hoàn thiện dần bộ phong với đầy đủ các thể khác của chữ Nôm. Trước khi bắt tay vào việc vẽ các thành tố, chúng tôi tổ chức esdl thành tố Hán–Nôm cơ bản trong MS Access để tiện cho việc quản lý, tra cứu.

*Đặt tên Quốc ngữ cho thành tố Hán–Nôm cơ bản* để nhận diện các thành tố này. Việc đặt tên có không ít khó khăn vì nhiều nét hay bộ không có tên quốc ngữ nên phải đặt tên quốc ngữ cho mỗi đơn vị mới để sử dụng nội bộ. Bước tiếp theo là tiến hành so sánh, đối chiếu giữa 4.415 thành tố Hán–Nôm cơ bản với các thành tố Hán–Nôm cơ bản trong *Thiên Tông Bản Hạnh* để tạo dáng đặc thù cho chữ Nôm.

Phương pháp vẽ các thành tố Hán–Nôm cơ bản được dựa trên những tiêu chí chung của nhóm tạo phong và ý kiến của các chuyên gia Hán–Nôm. Tiêu chí bao gồm:

- *Kích thước chữ:* Phong cách chữ, độ cao (rộng) của chữ, vị trí giữa các bộ, các nét trong khung chữ.
- *Kích thước nét chữ:* độ dày của nét ngang, bề rộng của nét sổ, kiểu nét ngang gấp, nét ngang móc, nét sổ móc, sổ hất, nét móc, nét phẩy...

Mục tiêu chính của nhóm là cố gắng vẽ giống với tự dạng của chữ trong nguyên tác *Thiên tông bản hạnh*, từ các thành tố Hán–Nôm cơ bản đến công đoạn ghép chữ.

#### 4.2. Cơ sở dữ liệu Hán–Nôm quản lý kho chữ

Xây dựng tệp esdl Hán–Nôm Access quản lý kho chữ, phiên bản 1.02, có 31.325 mục ghi theo trật tự *Giúp đọc* và bảng đề nghị CJK Extension C1 phần Việt Nam cho bộ chuẩn quốc tế Unicode.

Tệp quản lý kho chữ Hán–Nôm sử dụng Microsoft Access để tổ chức kho tri thức về các chữ Hán–Nôm tích lũy từ các chuyên gia. Chúng tôi coi đó vừa là kho dữ liệu kết quả làm việc vừa là một công cụ quản lý số lượng chữ rất lớn, có thể lên tới hàng chục nghìn đơn vị. Hiện nay, trên thực tế esdl quản lý kho chữ được xây dựng cho sách *Giúp đọc*, có tổng cộng khoảng 31.325 mục, được sắp xếp, phân tích theo vần A, B, C và theo từng mục từ trong mỗi vần.

Tệp quản lý kho chữ được lấy tên là NomnaTongLight\_db, mỗi mục gồm 10 trường cơ bản (cột), mỗi trường quản lý một loại thông tin riêng biệt về chữ. Các chữ được đưa vào tệp cơ sở dữ liệu này được quản lý trên các thông tin như sau:

- a. **Mã Unicode:** đây là trường mã lưu giữ mã chuẩn Unicode của các chữ trong mặt phẳng BMP và mặt phẳng 2. Trường này cho ta biết các chữ đã có mặt trong kho mã quốc tế.
- b. **Mã Việt Nam** (mã nội bộ) bắt đầu từ 60000, đánh theo hệ cơ số 16 (*hexadecimal*), trường này mang thông tin về mã quản lý nội bộ đối với những chữ Nôm chưa có mã trong kho quốc tế Unicode.
- c. **Hình chữ:** trường này chúng tôi cho hiển thị hình chữ. Đây là trường lưu trữ thông tin trực quan về hình chữ.
- d. **Tên gọi quốc ngữ:** đây là trường chứa âm đọc của các hình chữ ở mục c. theo cách viết quốc ngữ.
- e. **Mẫu ghép và các thành phần cấu tạo:**

– **12 mẫu ghép mô tả chuẩn (description characters):** Có 12 loại mẫu ghép điển hình theo chuẩn Unicode và ISO/IEC 10646 cho biết loại hình kết hợp nội tại của các thành tố để mô tả một chữ.

2ff0:	☐ (left to right):	字 + 字 → 孛 ( <i>chữ</i> )
2ff1:	☐ (above to below):	都 + 赤 → 黠 ( <i>đỏ</i> )
2ff2:	☐ (left to middle and right):	彳 + 重 + 丁 → 衝 ( <i>xung</i> )
2ff3:	☐ (above to middle and below):	𠂇 + 大 + 出 → 罌 ( <i>ra</i> )
2ff4:	☐ (full surround):	口 + 或 → 國 ( <i>quốc</i> )
2ff5:	☐ (surround from above):	門 + 月 → 閒 ( <i>nhàn</i> )
2ff6:	☐ (surround from below):	凵 + 丨 → 山 ( <i>son</i> )
2ff7:	☐ (surround from left):	匚 + 品 → 區 ( <i>khu</i> )
2ff8:	☐ (surround from upper left):	广 + 郎 → 廊 ( <i>làng</i> )
2ff9:	☐ (surround from upper right):	米 + 气 → 氣 ( <i>khi</i> )
2ffa:	☐ (surround from lower left):	乙 + 少 → 𠂇 ( <i>ít</i> ) và
2ffb:	☐ (overlaid):	口 + 丨 → 中 ( <i>trung, trong...</i> ).

– **Các thành phần cấu tạo:** Thông tin về các thành phần cấu tạo (các tên gọi hình chữ của thành tố 1 và thành tố 2).

- f. **Tên của bộ gốc:** Cho biết thông tin về tên gọi của các bộ gốc.
- g. **Số Unicode của bộ gốc:** viết tắt là URN (*Unicode Radical Number*).
- h. **Nét:** cho biết thông tin về số nét còn lại của chữ ngoài bộ gốc. Ví dụ: chữ 蛭 *giời* bộ 虫 *trùng* và 7 nét (chữ 丩 *trời*).
- i. **Thông tin dị thể, tục thể, giản thể:** các chữ dị thể (cùng chữ, viết khác dạng), tục thể (cùng chữ, viết theo dân gian), và giản thể (cùng chữ, viết theo kiểu Trung Quốc mới) có cột chính thể chỉ vào chữ phồn thể, thường là chữ theo Khang Hy.

Sau đây, xin đơn cử một vài ví dụ tổ chức nội bộ của csdl chữ Hán–Nôm (giản lược), nhằm mục đích giao diện với các phần mềm khác để trữ tri thức của mỗi chữ.

Mã	Nôm	Âm Qn	Mẫu	TT1	TT2	URN	Bộ	Nét	Âm bộ
272f9	蛭	<i>giời</i>	☐	虫	丩	142	虫	7	<i>trùng</i>
20cd2	啞	<i>lời</i>	☐	口	丩	030	口	7	<i>khẩu</i>
215f6	丩	<i>trời</i>	☐	天	上	037	大	4	<i>đại</i>
866b	虫	<i>trùng</i>		虫		142	虫	0	<i>trùng</i>
53e3	口	<i>khẩu</i>		口		030	口	0	<i>khẩu</i>
5929	天	<i>thiên</i>	☐	一	大	037	大	1	<i>đại</i>
4e0a	上	<i>thượng</i>	☐	丨	一	001	一	2	<i>nhất</i>
5927	大	<i>đại</i>		大		037	大	0	<i>đại</i>
535c	丨	<i>bốc</i>		丨		025	丨	0	<i>bốc</i>
4e00	一	<i>nhất</i>		一		001	一	0	<i>nhất</i>

Bảng 3: Vài ví dụ của cơ sở dữ liệu tri thức Nôm na.

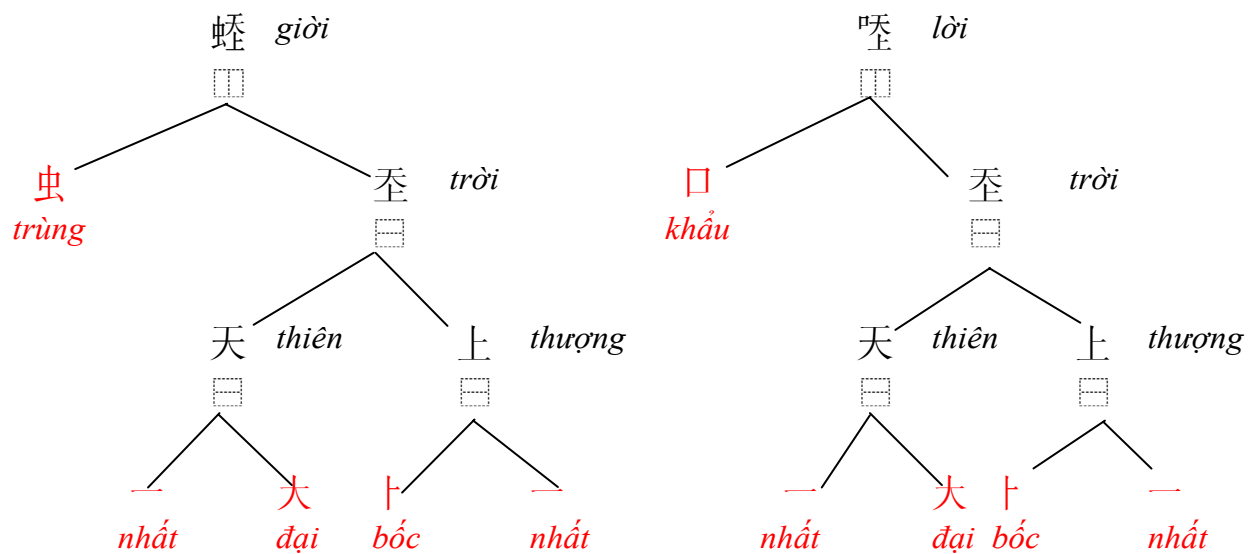


### 4.3. Tập phong

Mặc dù số lượng quản lý mục chữ trong csdl Hán–Nôm là 31.325, nhưng vì có những hình chữ Hán–Nôm trùng nhau, vì cách đọc và nghĩa khác nhau, nên số chữ Hán–Nôm thực sự khác nhau về hình dáng cần tạo phong là 17.761 chữ.

Tập phong được chúng tôi thực hiện trên phần mềm Fontlab 3.1 để xây dựng. Phần mềm Fontlab 3.1 có thể chứa được vài chục nghìn chữ ( $2^{16} = 65.536$ ). Mỗi một chữ Hán–Nôm trong đó được gắn với 2 thông số mã tương ứng với csdl quản lý kho chữ. Đó là: mã nguồn và mã Unicode. Bộ phong có nhiệm vụ đặt cho mỗi hình chữ một mã duy nhất, và csdl Nôm có nhiệm vụ đặt các thông tin khác của một hình chữ Nôm như cách đọc quốc ngữ, các bộ phận hình chữ có nghĩa tạo ra hình chữ Nôm, bộ, số nét, v.v.

Cơ sở dữ liệu tri thức chữ Nôm (xem ví dụ ở Bảng 3) cho phép chúng ta vẽ ra hình cây cấu tạo (hình cây mô tả) của các chữ, 蛭 *giời*, 啞 *lời*, 𠂇 *trời*, 天 *thiên*, 上 *thượng*, và lá cây là các thành tố cơ bản hay bộ (màu đỏ): 虫 *trùng*, 𠂇 *khẩu*, 大 *đại*, 卜 *bốc*, 一 *nhất*.



Bảng 4: Hình cây cấu tạo rút từ cơ sở dữ liệu tri thức Nôm na ở Bảng 3.

Bộ phong Nôm Na Tổng thể, phiên bản 1.02 có tổng cộng 17.761 chữ, trong đó:

- Chữ Hán Việt ở mặt phẳng cơ bản BMP: 10.159
- Chữ Hán Nôm ở phần Extension A: 388
- Chữ Hán Nôm ở phần Extension B: 3.100
- Chữ Nôm ở mặt phẳng 6: 4.014.

### 4.4. Giúp đọc Nôm và Hán Việt

*Giúp đọc* được xuất bản tháng 9/2004, có chỉnh sửa bổ sung với bản thảo tối ưu nhất. Đặc biệt phần chữ Nôm được sử dụng bộ phong NômNaTongLight thể chữ mảnh thay cho chữ Hán–Nôm viết tay. Xin xem Phần 2 của *Giúp đọc Nôm và Hán Việt* (Nxb Đà Nẵng 2004), các mục *giời* tr. 431; *lời* tr. 553.

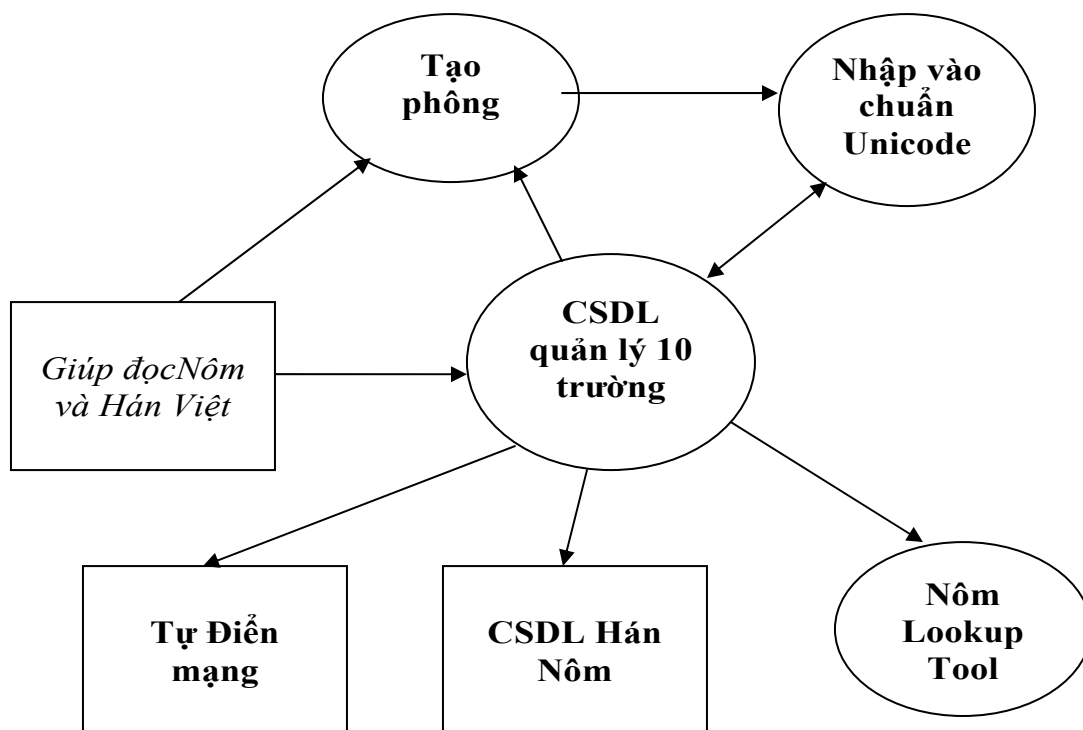
#### 4.5. Tra cứu *Giúp đọc* trên mạng

*Giúp đọc* sẽ được chuyển lên mạng, phổ biến rộng rãi trong nước cũng như quốc tế, nhằm phục vụ cho việc tra cứu và học tập.

### 5. Các công đoạn làm việc

#### 5.1. Công đoạn tạo phông

Công đoạn tạo phông là một quy trình mới nằm trong quy trình Nôm na vừa tạo dáng chữ Hán-Nôm theo phong cách chữ Hán-Nôm Việt Nam vừa đặt chúng ngay vào chuẩn đa ngữ quốc tế và trong các phần mềm thông dụng hiện nay: Bộ kí tự Unicode cho chữ quốc ngữ và chữ biểu ý CJKV (C: Trung, J: Nhật, K: Triều, V: Việt Nam).



Bảng 5: Sơ đồ quy trình Nôm na

Fontlab 3.1 không hỗ trợ biểu diễn các mã trên mặt phẳng (Plane) 2 và mặt phẳng (Plane) 6 của Unicode nên việc mã hóa trở nên khó khăn hơn, phải sử dụng Microsoft Volt Software thực hiện mã hóa thay thế. Dựa trên kết quả của việc đồng bộ hóa dữ liệu, việc mã hóa thay thế được hoàn thành trong thời gian 1 tháng.

#### 5.2. Công đoạn tạo chế bản *Giúp đọc*

*Giúp đọc Nôm và Hán Việt* được chia làm 2 phần chính: Phần Thứ nhất – Tìm âm và Phần Thứ hai – Tìm nghĩa đã được Linh mục Anthony Trần Văn Kiệm hoàn thành bản thảo và đã được xuất bản 3 lần. Phần chữ Hán-Nôm được viết tay. Công việc chính của nhóm Nôm Na là đưa chữ Nôm từ một tệp phông duy nhất vào máy tính thay thế cho phần chữ Nôm viết tay trước đây.

Công việc tạo chế bản được thực hiện trên nhiều công đoạn và sử dụng tệp NomNaTongLight.ttf để hiện thị chữ Nôm:

- Chia *Giúp đọc* ra từng tệp nhỏ.
  - Phần thứ nhất - Tìm âm: được chia thành bộ gốc và các bộ thủ.
  - Phần thứ hai - Tìm nghĩa: được chia theo vần quốc ngữ.
- Đưa chữ Hán–Nôm vào *Giúp đọc*
  - Đưa chữ Hán–Nôm vào từng mục từ trong mỗi tệp theo trật tự *Giúp đọc*.
  - Đưa chữ Hán–Nôm vào các tệp bộ gốc và các bộ thủ.
- Tiến hành biên tập bao gồm các vấn đề:
  - So sánh với bản gốc của *Giúp đọc*.
  - Sửa lỗi chính tả trong *Giúp đọc*.
  - Sửa định dạng chữ.
  - Dàn trang.

Chế bản cuốn *Giúp đọc* là một phần nằm trong tổng thể quy trình Nôm Na. Chính sửa theo *Giúp đọc* luôn gắn liền với tính chính xác của các chữ Hán–Nôm trong csdl quản lý kho chữ và bộ phong NomNaTongLight.

### 5.3. Công đoạn đồng bộ hoá, thống nhất hoá và kiểm sửa các tệp dữ liệu

Kết quả thu được của quy trình Nôm Na là csdl quản lý kho chữ, bộ phong và *Giúp đọc*. Đồng bộ hóa có nghĩa là các tệp dữ liệu được tạo ra trong quá trình làm việc đều có sự liên hệ mật thiết với nhau, và cần được điều chỉnh ăn khớp nhau. Thống nhất hoá ở đây nghĩa là tất cả các chữ trong kho chữ Nôm của chúng ta phải được xác định vị trí tương ứng trong bộ kí tự Unicode và dùng mã Unicode đã được cấp. Những chữ nào chưa có mã Unicode mới được cấp mã Việt Nam riêng.

Với khối lượng 17.761 chữ Hán–Nôm, việc đồng bộ hoá dữ liệu giữa ba sản phẩm là cần thiết để dữ liệu tương ứng với nhau. Trong thời gian 2 tuần công việc này đã được hoàn thành, nhờ sử dụng một số công cụ hỗ trợ như csdl Access, chương trình để truy tìm các chỗ bất đồng.

Csdl quản lý kho chữ dùng để quản lý các chữ trong bộ phong nên giữa chúng phải thống nhất, bộ phong có bao nhiêu chữ thì csdl phải tương ứng. Từ việc thống nhất đó quay trở lại sửa *Giúp đọc*. Lấy bộ phong làm gốc để tìm ra các lỗi, sự chênh lệch giữa csdl quản lý và *Giúp đọc*, bằng cách sử dụng lập trình xuất ra từng chữ trên nền mạng html để so sánh với csdl quản lý với *Giúp đọc*.

Trong Unicode, mặt phẳng 0, còn gọi là mặt phẳng đa ngữ cơ bản BMP, chứa khoảng 27.000 chữ biểu ý. Mặt phẳng 2 chứa khoảng trên 40.000 chữ biểu ý. Khi nói tới chữ biểu ý trong Unicode, điều đó hiện nay có nghĩa là tất cả các chữ đã được cấp mã trong các mặt phẳng 0, 1 và 2. Nhóm Nôm na dùng mặt phẳng 5 (chưa ai dùng) làm vùng nháp để đưa tất cả các chữ của văn bản vào. Mặt phẳng này được dự định để dùng làm vùng nháp cho việc đưa vào các loại văn bản khác. Mặt phẳng 6 được dự định để mã hoá cho các chữ Nôm thuần Việt vốn chưa được cấp mã trong Unicode.

## 6. Các quy trình làm việc được rút ra

Trên cơ sở những công đoạn đã được thực hiện và đúc kết kinh nghiệm để tổ chức làm việc tốt hơn, chúng tôi rút ra một số quy trình nhỏ sau đây. Vì việc phát hiện chữ Nôm mới vẫn còn tiếp diễn khi chúng ta đi vào nghiên cứu các văn bản cổ, nên việc hoàn chỉnh quy trình tối ưu hỗ trợ cho công việc lưu giữ và thể hiện chữ Nôm là cần thiết.

### 6.1. Quy trình tạo phông cho chữ mới nhận được

Đây là một quy trình nhỏ nhằm đưa thêm chữ mới vào bộ phông, đương nhiên xảy ra trong nhiều năm tới. Mục đích và yêu cầu của kho chữ là tổng hợp đầy đủ các thể loại chữ Nôm của Việt Nam. Do vậy để tiến hành tốt cho việc tạo phông chữ mới, quy trình cần tuân thủ là:

1. *Xác định các thành tố của chữ*: Khi phát hiện một chữ mới trong bất kì một tài liệu Nôm nào mà trong bộ phông vẫn chưa có, việc đầu tiên là xác định các thành tố, phân tích cấu trúc nội tại của chữ đó.

2. *Vẽ các thành tố*: Nếu các thành tố trong một chữ mới bất kì nào đó đã có hình chữ thì sử dụng mẫu chung để vẽ. Nếu chưa có thì tiến hành vẽ thêm thành tố mẫu mới cho phù hợp với bộ phông.

3. *Lắp ghép các thành tố*: Sau khi vẽ các thành tố xong, tiến hành lắp ghép tạo và điều chỉnh chữ mới cho hợp phong cách của bộ phông.

4. *Đưa chữ vào phông chữ Nôm Na Tổng*: Khi đã hoàn thành công việc vẽ chữ, bước tiếp theo là đưa chữ Nôm mới nhận được vào csdl quản lý bộ phông. Những chữ Nôm đã có mã Unicode thì gán cho nó mã nguồn và mã Unicode, những chữ chưa có mã Unicode thì gán cho nó mã nguồn và mã tạm.

### 6.2. Quy trình đưa văn bản vào kho và quản lý chữ trong văn bản

Việc quản lý kho các văn bản cổ chữ Nôm cần một dự án nghiên cứu khác. Tuy nhiên song song với việc đưa các văn bản cổ này vào máy tính, thì cũng xuất hiện nhu cầu đưa các chữ Nôm mới phát hiện vào kho chữ Nôm, theo chuẩn Việt Nam và chuẩn quốc tế. Vì vậy cần xử lý đồng bộ việc xây dựng hai kho cơ bản: kho chữ Nôm và kho tài liệu chữ Nôm. Ở đây chúng tôi chỉ xin nêu phác họa vắn tắt quy trình cần có cho việc đưa vào một văn bản chữ Hán-Nôm.

Với mỗi văn bản chữ Nôm mới được đưa vào kho tài liệu, nhằm phục vụ nghiên cứu và tìm trên mạng internet:

1. Cất giữ thông tin thư viện, xuất bản và bảo tàng... liên quan tới văn bản (tên văn bản, năm xuất hiện, người phát hiện,...) và ảnh chụp các trang nguyên bản.
2. Xác định mã Unicode cho mọi chữ Nôm trong văn bản.
3. Nếu là chữ mới chưa có mã, khởi động quy trình tạo phông cho chữ này.
4. Đưa các thông tin liên quan tới chữ này vào kho thông tin chữ.
5. Ghi nguồn của chữ này (thông tin xuất hiện) vào kho thông tin chữ.
6. Đưa chữ vào văn bản trên máy.

### 6.3 Quy trình chuẩn hoá kho chữ Nôm

#### Đây là quy trình nhận diện chữ Nôm trong kho chữ biểu ý quốc tế.

1. *Phân tích cấu trúc nội tại*: Việc đầu tiên phải xác định chữ đó thuộc loại chữ Nôm mượn nguyên chữ Hán hay chữ thuần Nôm. Sau đó tiến hành phân tích cấu trúc nội tại, xác định bộ thủ, số nét theo chuẩn Unicode.

2. *Xác định mã chuẩn theo Unicode*: Cần phải tiến hành kiểm tra xem chữ đó đã có trong Unicode chưa, nếu có rồi, xác định điểm mã. Nếu chưa có, xác định xem có trùng với chữ trong kho chữ Nôm của Việt Nam không. Nếu có rồi, cho chữ đó mã Việt Nam. Nếu chưa có, cấp mã Việt Nam mới cho nó. Đồng thời xác định các thông số Khang Hy của chữ, xác định bộ thủ và số nét chung của chữ, số nét phần thứ nhất của chữ ngoài bộ thủ, để có thể nộp chữ này cho kho chữ IRG khi có điều kiện.

3. *Gắn âm đọc và thông tin xuất xứ của chữ*: Ngoài việc xác định mã Unicode cần phải gắn âm đọc cho nó và những thông tin xuất xứ của chữ để biết được chữ Nôm đó được lấy từ văn bản nào, thời nào.

4. *Đưa tất cả các thông tin này vào kho quản lý chữ Nôm*: Chúng tôi dùng phương pháp quản lý tri thức trong hệ quản lý cơ sở dữ liệu để trữ các tri thức về một chữ Hán–Nôm của các chuyên gia Hán–Nôm. Tri thức này được dùng ngay trên mạng, khi một người tìm chữ Hán–Nôm theo cách đọc quốc ngữ, hay tìm các cách đọc quốc ngữ của một chữ Hán–Nôm. Khi gắn chúng với xuất xứ, chúng tôi hy vọng tạo điều kiện cơ bản cho việc nghiên cứu sâu về sau của các chuyên gia Hán–Nôm.

## 7. Kết luận

Việc bảo tồn và phát huy những giá trị văn hóa truyền thống càng ngày càng được coi trọng, không những ở trong nước mà cả trong những chuyên gia ở nước ngoài. Chữ Hán–Nôm, một loại hình văn tự được người Việt sử dụng ngót một nghìn năm hiện đang có nguy cơ bị mai một. Quy trình Nôm na: “*Giúp đọc Nôm và Hán Việt*” và chữ Nôm trên mạng là một trong những dự án được Hội Bảo tồn Di sản chữ Nôm tiến hành tổ chức xây dựng, nhằm mang lại diện mạo mới, cách nhìn mới về chữ Nôm Việt Nam.

Tại Việt Nam nhóm Nôm Na đã tiến hành xây dựng quy trình tạo phong chữ Nôm. Bước đầu nhóm đã thực hiện thành công việc tạo phong chữ Hán–Nôm cho *Giúp đọc và CJK Extension C1* mang phong cách chữ Nôm trong *Thiền Tông Bản Hạnh*. Những bước tiếp theo dự định sẽ chế tạo các phong cho chữ Hán–Nôm theo các thể loại khác nhau, có nguồn gốc xuất xứ từ những văn bản Hán–Nôm tiêu biểu cổ nhất cho đến những văn bản mới nhất, đáp ứng được những nhu cầu nghiên cứu khoa học, làm công tác chế bản cũng như in ấn các văn bản Nôm khác nhau, nhằm khôi phục lại nguyên bản những tác phẩm Hán–Nôm bằng công nghệ thông tin hiện đại.

Hiện nay, cuốn *Giúp đọc Nôm và Hán Việt* của Anthoy Trần Văn Kiệm đã được xuất bản. Bản quyền bộ phong chữ Nôm thuộc nhóm Nôm Na – Hội Bảo tồn Di sản chữ Nôm trực tiếp quản lý. Cuốn *Giúp đọc* hiện đang được chuyển lên mạng phục vụ nhu cầu học tập và tra cứu rộng rãi trong và ngoài nước. Sắp tới chúng tôi dự định triển khai quy trình đưa các văn bản Hán–Nôm tiêu biểu lên mạng máy tính.

Để có thể thực hiện được tốt những công việc trên, quy trình *chuẩn hoá các chữ Hán-Nôm* là cần thiết, nó vừa thể hiện tính truyền thống, vừa mang tính hiện đại.



### Tài liệu tham khảo

1. Các tập mã chữ Nôm do Việt Nam cung cấp cho nhóm ISO/IEC 10646 JTC1/IRG từ năm 1994 đến nay: NPCT 2.1, TCVN 5712: 1993, TCVN 5773: 1993, TCVN 6056: 1995, VHN1: 1998, VHN2: 1998. Đề nghị CJK Extension C1 của Việt Nam.
2. ISO/IEC 10646 JTC1/IRG từ năm 1994 đến nay, Unihan 3.1 Radical-Stroke Index.
3. Lê Mai Phương. *Học chữ Nôm theo Tam thiên tự*. Trình bày tại Hội thảo Hè 2002, Đại học Maine.
4. Ngô Thanh Nhân, Ngô Trung Việt và Nhóm Nôm Na. *Quy trình Nôm Na*, trình bày tại Hội thảo Hè 2002, Đại học Maine.

Copyright © 2004 by The Institute of Hán Nôm Studies  
& The Vietnamese Nôm Preservation Foundation.

- 
- <sup>1</sup> Linh mục Trần Văn Kiệm. *Giúp đọc Nôm và Hán Việt*. Nhà xuất bản Thuận Hóa, 1999.
  - <sup>2</sup> Viện Ngôn ngữ học. *Bảng tra chữ Nôm*. Nhà xuất bản Khoa học Xã hội Hà Nội 1976.
  - <sup>3</sup> Vũ Văn Kính & Nguyễn Quang Xỷ. *Tự Điển Chữ Nôm*. Trung tâm Học liệu. Sài Gòn 1971.
  - <sup>4</sup> Vũ Văn Kính. *Đại tự điển chữ Nôm*. Nhà xuất bản Văn Nghệ Tp Hồ Chí Minh & Trung tâm Nghiên cứu Quốc học Huế. 1999.
  - <sup>5</sup> Nguyễn Quang Hồng (chủ biên). Lời dẫn *Tự Điển Chữ Nôm Tiếng Việt*. Viện nghiên cứu Hán Nôm. Đang in.
  - <sup>6</sup> Trương Đình Tín. *Bảng Phiên âm Nôm Việt*. Nhà xuất bản Thuận Hóa, 2003.
  - <sup>7</sup> *Thiên tông bản hạnh*. Bản khắc in năm Bảo Đại thứ 7 (1933).
  - <sup>8</sup> Đỗ Quốc Bảo, 4.415 nét cơ bản.