

## ĐƯA CHỮ HÁN-NÔM VÀO THIẾT BỊ CẦM TAY

### 1. Sơ lược giới thiệu.

Việc bảo tồn, nghiên cứu và phát huy di sản Hán Nôm trong nước đang được nhiều tổ chức và cá nhân quan tâm, hơn nữa trong nước đang có phong trào học tiếng Hoa, tiếng Hoa hiện đang là một ngoại ngữ được phát triển, có lẽ chỉ sau tiếng Anh. Người dùng hiện đang rất quan tâm tới các phần mềm tra cứu và sách điện tử về chữ Hán và chữ Nôm. Không kể vấn đề bảo tồn nền văn hóa Hán-Nôm của dân tộc mà xét ngay trong việc phát triển của tiếng Việt hiện đại có thể thấy lượng từ gốc Hán Việt chiếm tỉ trọng rất lớn và vẫn tiếp tục phát triển, nên nhu cầu tra cứu nghĩa gốc Hán Việt là không thể thiếu cả hiện tại cũng như trong tương lai.

Cùng lúc đó, nhờ những tiến bộ không ngừng của công nghiệp vi điện tử, và để đáp ứng nhu cầu thị hiếu người dùng, thị trường điện thoại di động “thông minh” và PDA Việt Nam gần đây đang phát triển mạnh. Hiện tại các ứng dụng tiếng Việt trên nền công nghệ Java cho ĐTDĐ và PDA đang được khuyến khích phát triển, về phần cứng thì các ĐTDĐ và máy PDA đời mới đều hỗ trợ các ngôn ngữ tiếng Việt, tiếng Hoa, nên việc làm các phần mềm tra cứu Hán, Nôm trên các thiết bị cầm tay không còn là chuyện “phi thực tế” như cách đây mới vài năm.

Từ những lý do trên Trung tâm Công nghệ Thông tin Thừa Thiên Huế đã lên kế hoạch và vừa hoàn thành dự án “Xây dựng phần mềm Từ điển Hán Nôm trên PDA”. Bài viết này xin giới thiệu cụ thể về dự án đó và các kết quả thu được.

### 2. Các thông tin chung.

#### 2.1. Mục tiêu cụ thể : Từ điển Hán Nôm trên PDA, gồm:

1. Phần mềm tra cứu Hán - Nôm trên các thiết bị cầm tay PDA.
2. Bổ sung một số phần Trung văn hiện đại để phục vụ khách du lịch và người học tiếng Hoa hiện đại (sinh viên Trung văn).

#### 2.2. Nhận xét, đánh giá hiện trạng:

Về phần cứng, máy PDA hiện tại đã đạt tới cấu hình phổ thông là 64MB Rom + 64 MB Ram + 265 MB Thẻ nhớ SD, tốc độ 300MHz trở lên. Thường chạy hệ điều hành Windows Mobile 2002/2003, có thể chạy các ứng dụng lớn với hàng chục MB dữ liệu, do đó chạy tốt các ứng dụng tra cứu từ điển vì loại ứng dụng này chỉ lớn về dữ liệu chứ chương trình tương đối không phức tạp và không đòi hỏi cấu hình máy mạnh.

Chức năng tra cứu từ trước đến nay vẫn là chức năng thông dụng nhất trên điện thoại di động và PDA, và cả trong tương lai. Chẳng hạn hầu như máy PDA nào bán ra trong nước hiện nay cũng cài sẵn một bộ từ điển Anh-Việt-Anh.

Về từ điển Hán Việt cho PDA hiện nay đã có bộ “Từ điển Hán Việt Thiều Chửu” do Viện Việt Học và Trung tâm Việt ngữ Văn Lang - San Jose (Hoa kỳ) viết bằng ngôn ngữ JAVA, phổ biến miễn phí tại [www.viethoc.com](http://www.viethoc.com), nhưng chưa có chữ Nôm, và hơi khó dùng, vì phải sử dụng một máy ảo JAVA ít quen thuộc trong nước là SuperWaba.

Trong nước hiện chưa có đơn vị CNTT chính quy nào làm từ điển Hán Việt và từ điển chữ Nôm trên PDA. Trên máy để bàn thực tế chỉ có hai nhóm nghiệp dư có làm là nhóm Hanosoft của Tổng Phước Khải & Lê Anh Minh tại tp H.C.M và nhóm của người viết dự án này là Phan Anh Dũng & Nguyễn Thế ở Huế có làm. Trên điện thoại di động thì đã có bộ từ điển Hán Việt cỡ nhỏ của chúng tôi (Phan Anh Dũng) hiện phổ biến tại <http://www.samsungmobilegames.com.vn/2005/default.asp> được đánh giá khá tốt.

Người sử dụng các thiết bị PDA thường là những người hay đi công tác xa như các nhà báo, phóng viên, doanh nhân, khách du lịch... và thường là những người có học và có điều kiện kinh tế tương đối nên việc làm và phổ biến Từ điển Hán-Việt-Nôm trên PDA cũng có phần thuận lợi.

### **3. Các vấn đề khó khăn về kỹ thuật và giải pháp.**

#### **3.1.1. Vấn đề mã hóa font chữ Hán-Nôm**

Hiện tại chữ Nôm đã được cấp mã chuẩn Unicode, nhưng phần lớn các chữ thuần Nôm đều được mã hóa ở các mặt phẳng mã hóa bổ sung 2, mã số lớn hơn 2 byte, phải dùng kỹ thuật surrogate (thay thế) để biểu diễn nó bằng 2 ký tự 2byte ở mặt phẳng mã hóa cơ sở, thường gọi đó là các chữ „surrogate“. Số lượng các phần mềm hỗ trợ kỹ thuật Surrogate hiện còn khá hạn chế như Windows XP, Office XP, và các phần mềm khác dựa trên công nghệ UniScribe của Microsoft ; Mac OS X, và các phần mềm khác dựa trên công nghệ ATSUI của Apple.

Như vậy việc dùng các chữ Nôm ở mặt phẳng 2 trong điều kiện ở nước ta hiện nay là có nhiều cái khó khăn, nhất là vấn đề đưa chữ Nôm lên mạng internet, nếu theo đúng chuẩn Unicode thì vấp phải một số vấn đề kĩ thuật khá rắc rối. Máy PDA đời mới dùng HĐH Windows Mobile 2002-2003-2005 về nguyên tắc thì đã hỗ trợ surrogate, nhưng khả năng hỗ trợ khá hạn chế, hiển thị không đúng yêu cầu. Ví dụ chúng tôi cho hiển thị một dòng chữ Nôm Surrogate trong một Label của một dự án PDA viết bằng VS.NET C# 2003 thấy các chữ Nôm đó hiển thị được, nhưng bị đè một phần lên nhau !... Hiện tại chúng tôi chưa biết có phần mềm ứng dụng nào sử dụng được chữ surrogate trên PDA.

Giải pháp tạm thời là đưa chữ Nôm vào vị trí các chữ Hán bổ sung CJK ít được sử dụng trong mặt phẳng cơ sở. Tuy chiếm dụng mã CJK cơ sở, nhưng không chiếm dụng mã các chữ Hán có trong BIG5 và GB tức là các chữ phồn thể và giản thể đã được nhà nước TQ tiêu chuẩn hoá. Do đó khi dùng bộ font TTF chữ Hán-Nôm này để xem các trang WEB của TQ cũng như Đài Loan thấy vẫn hiển thị đầy đủ, hầu như không thấy lỗi. Việc đưa các chữ Nôm này lên mạng Internet, hay sao chép giữa các chương trình Windows rất dễ dàng, thông suốt, không có trục trặc gì cả. Đây là giải pháp chúng tôi đã ứng dụng thành công trên thực tế.

### **3.1.2. Vấn đề kích thước font chữ Hán-Nôm**

Font Hán Nôm của nhóm Huế có kích thước khoảng 12 MB, nhỏ hơn font Nom Na Tong của nhóm Nôm Na (Hà nội, 18MB), nhưng vẫn khá lớn đối với bộ nhớ của máy PDA.

Giải pháp: Sử dụng các chương trình vẽ font xóa bớt các chữ thừa hoặc ít khi sử dụng tới, đồng thời kết hợp dùng giải pháp tạo chữ mới từ tổ hợp các glyph đã có để giảm bớt dữ liệu font. Ngoài ra chúng tôi đã nghiên cứu thực hiện được kỹ thuật nạp font chữ từ thẻ nhớ ngoài mà không phải copy vào thư mục FONT của WINDOWS, tức là không chiếm dụng bộ nhớ chính của PDA.

Hiện tại chúng tôi đã hoàn thành bộ FONT Hán Nôm cho máy PDA, kích thước giảm gấp đôi, chỉ còn hơn 5MB, thẻ nhớ ngoài của các loại PDA thông dụng trên thị trường tối thiểu là 256MB, nên kích thước 6MB này hoàn toàn chấp nhận được...

### **3.1.3. Giải pháp hiển thị đồng thời chữ Việt và chữ Hán Nôm**

*Yêu cầu đặt ra:* Hiển thị đồng thời chữ Việt và chữ Hán Nôm trong chương trình với nhiều kích thước, màu sắc phong phú. Chữ Hán-Nôm đầu mục nên hiển thị với cỡ chữ thật lớn để người mới học dễ nhận dạng các nét chữ.

*Vấn đề kỹ thuật:* Bộ VS.NET 2003 cho PDA không cung cấp nhiều Tool điều khiển rất thông thường của Windows chạy trên các máy để bàn (có từ thời VS5.0/6.0) như điều khiển RichTextBox, HTMLBrowser... mà chỉ có TextBox hay Label, khi đã chọn định dạng font là chữ Việt (English) Unicode thì không hiển thị được chữ Hán và ngược lại.

*Giải quyết:* Tìm kiếm các công cụ mã nguồn mở trên mạng. Hiện tại chúng tôi sử dụng thư viện đối tượng HLTMBrowser của OpenNetCF, về cơ bản đã giải quyết được các yêu cầu trên, nhưng còn hạn chế là chưa có thể copy các đoạn văn bản tô đen (selected text) của nó bằng mã lệnh để có thể thực hiện chức năng tra chéo và nhảy tới mục từ khác ngay trong từ điển.

### **3.1.4. Tra từ điển trong ứng dụng khác.**

#### **1. Yêu cầu đặt ra:**

Thông thường các ứng dụng tra cứu (từ điển) ngoài khả năng tra bằng cách đánh từ khóa trong ứng dụng còn chức năng khá quan trọng là tra nhanh từ trong ứng dụng khác. Ví dụ tô đen từ trong WORD rồi kích hoạt một tổ hợp phím nóng là hiển thị ngay một cửa sổ giải nghĩa nhỏ dạng popup, ở mức cao hơn là dùng kỹ thuật HOOK API để nhận dạng chữ nằm dưới trỏ chuột và tra theo kiểu „Click'n'See“.

#### **2. Các vấn đề kỹ thuật:**

Trên máy PDA dùng hệ điều hành Windows khả năng can thiệp hệ thống và tương tác giữa các ứng dụng đã bị hạn chế đến mức tối đa. Ngay khả năng trao đổi dữ liệu qua clipboard bằng mã lệnh cũng không được vì bộ VS.NET 2003 không cung cấp đối tượng clipboard trong các dự án „smart device“.

#### **3. Các kết quả bước đầu:**

- Gọi trực tiếp hàm API để xử lý cho cửa sổ popup luôn luôn nổi lên trên các ứng dụng khác, để hiển thị giải nghĩa cho từ điển khi tra chéo.
- Gọi trực tiếp hàm API để giả lập các phím gõ CTRL+C, CTRL+V nhằm đưa dữ liệu vào và lấy ra từ Clipboard, khắc phục việc không có đối tượng Clipboard của VS.NET 2003.
- Đưa được biểu tượng chương trình lên thanh tool của các ứng dụng để có thể tra chéo bằng cách nhấn lên biểu tượng, tương đương như việc nhấn tổ hợp phím nóng trên máy để bàn.
- Vấn đề còn vướng mắc là đối tượng WebBrowser (thuộc loại mã nguồn mở) của bộ OpenNetCF không cung cấp khả năng copy dữ liệu bằng giả lập phím CTRL+C, nên tra chữ trong WORD tốt nhưng khi cần tra chữ trong chính cửa sổ của chương trình lại gặp khó khăn.

### **3.1.5. Vấn đề định dạng nén và lưu trữ dữ liệu của từ điển.**

#### **1. Yêu cầu đặt ra:**

Các phần mềm từ điển là loại phần mềm có dữ liệu lớn, và dữ liệu đó quyết định giá trị của từ điển. Một bộ từ điển chuyên nghiệp bao giờ cũng phải có chức năng nén và mã hóa dữ liệu, trên PDA do không gian lưu trữ hạn chế nên việc này lại càng cần thiết.

#### **2. Vấn đề kỹ thuật:**

Việc xử lý nén nguyên một file hay chuỗi không khó vì trên mạng có sẵn các giải pháp mã nguồn mở NZlib bằng C#. Cái khó là nếu nén riêng lẻ từng mục từ theo từng stream thì hiệu suất nén rất thấp. Còn nếu nén tất cả các mục từ theo một stream duy nhất, (có theo dõi position của từng mục để lập chỉ mục, giúp đọc lại mục từ đó ra), thì kích thước data giảm rất nhỏ, nhưng giải pháp mã nguồn mở nói trên có những vấn đề rắc rối hiện chúng tôi chưa giải quyết được.

#### **3. Giải pháp:**

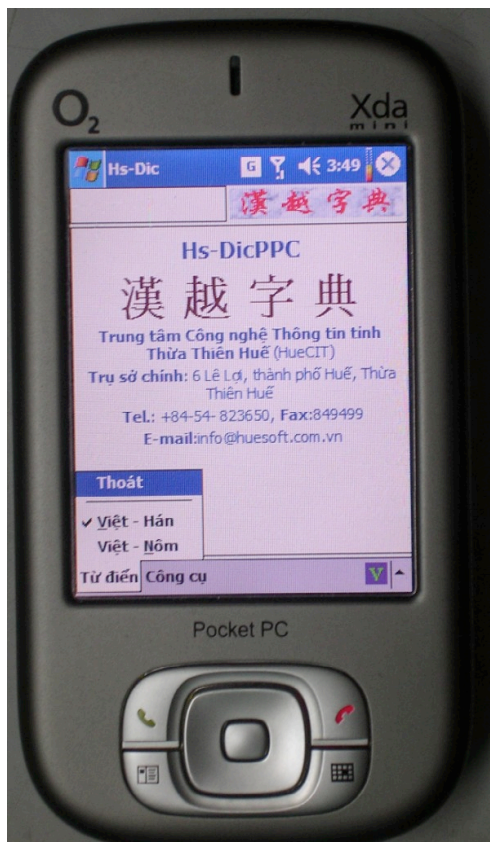
Phiên bản hiện tại do các từ điển không lớn lắm, nên có thể chấp nhận dùng bộ NZlib, nén theo từng mục từ, chấp nhận tỉ số nén nhỏ. Khi phát triển bổ sung các từ điển lớn hơn chúng tôi sẽ đồng thời hoàn thiện kĩ thuật nén để giảm nhỏ kích thước file data.

## **4. KẾT QUẢ**

Đã hoàn chỉnh phần mềm Từ điển Hán Nôm trên máy PDA, gồm các từ điển sau:

1. Từ điển Hán Việt Thiệu Chửu tăng bổ, bổ sung giải nghĩa nhiều chữ mới lấy từ bộ từ điển Hán Việt Trần Văn Chánh và Từ điển Trung Việt của NXB KHXH. Từ điển này thích hợp khi nghiên cứu di sản Hán-Nôm cổ điển, tra cứu các văn bản cổ....
2. Từ điển Việt Hán. Đây là một từ điển cỡ nhỏ, phần giải nghĩa bằng chữ Hán sử dụng chữ giản thể tức là theo tiếng Hoa hiện đại.

3. Từ điển Trung-Việt (tra theo âm PinYin). Nội dung gốc thực tế là ở Từ điển Trung Việt của NXB KHXH rút gọn lại. Rất ích lợi cho các đối tượng sinh viên, học sinh học tiếng Hoa Hiện tại và khách du lịch.
4. Từ điển chữ Nôm: Chủ yếu lấy theo cuốn Bảng tra chữ Nôm của NXB KHXH 1976, có bổ sung vài trăm chữ mà chúng tôi thấy cần thiết.

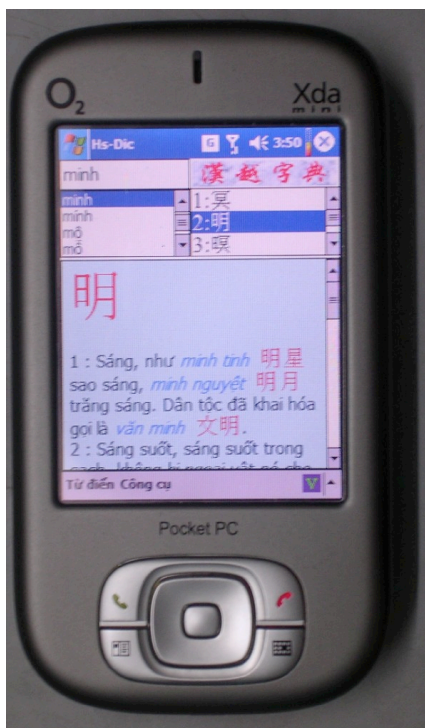


## MỘT SỐ HÌNH ẢNH

### Giao diện chính

(Chỉ có tính cách tham khảo, có thể sẽ thay đổi nhiều trên sản phẩm thật sự )





2. Từ điển Hán Việt  
(Tra chữ Hán theo âm Hán Việt)



2. Từ điển chữ Nôm  
(Tra chữ Nôm theo âm Nôm)

- Bài viết này không trình bày cụ thể cách sử dụng, người dùng có thể xem trong help1 của chương trình.
- Tuy còn một số khuyết điểm và hạn chế về kỹ thuật, về tính hoàn chỉnh và tính chuyên nghiệp, nhưng sản phẩm đã đáp ứng được các yêu cầu cơ bản của một phần mềm tra cứu, và vẫn có khả năng tiếp tục phát triển
- Khả năng sử dụng và phổ biến của sản phẩm trong nước cũng như quốc tế khá cao. Trước mắt chúng tôi sẽ đưa sản phẩm này giới thiệu tại Hội nghị quốc tế về chữ Nôm 2006 tổ chức ở Huế.

## 5. ĐỀ XUẤT PHÁT TRIỂN

- Hoàn chỉnh kỹ thuật nén dữ liệu của phần mềm, cũng như tối ưu hóa mã nguồn và các thủ tục hiện đang sử dụng.
- Tu sửa lại giao diện chương trình cũng như các thủ tục đáp ứng sự kiện phong phú hơn.

- Đưa thêm những giải pháp tra khác như tra theo bộ thủ, tra theo số nét, tra theo mã thương hiệt giản .v.v.
- Bổ sung thêm nhiều từ điển, cũng như thêm dữ liệu cho các từ điển có sẵn, tiến tới một bộ từ điển đa ngữ có giá trị.
- Chuyển cơ sở dữ liệu của các từ điển thu được khi làm phần mềm này vào trang WEB tra cứu Hán Nôm của trung tâm, để trang này thêm phong phú về nội dung và có giá trị sử dụng cao.
- Viết phần mềm tương đương cho Điện thoại di động đời mới (các dòng ĐTDĐ hỗ trợ JAVA MIDP2.0 và nhà sản xuất đã cài sẵn font tiếng Hoa) và chuyển các từ điển đã làm vào ĐTDĐ.

***Nhóm làm phần mềm:***

Phan Anh Dũng

Dương Văn Việt

Hoàng Thị Ngọc Dung

Liên hệ: Trung tâm Công nghệ Thông tin tỉnh Thừa Thiên Huế (HueCIT)

Địa chỉ: 6 Lê Lợi, thành phố Huế, Thừa Thiên Huế

Tel: (054)849499, Fax: (054)849498

Email: [info@huesoft.com.vn](mailto:info@huesoft.com.vn)

Website: <http://www.huesoft.com.vn>

HUẾ, ngày 20-4-2006

Thay mặt nhóm làm phần mềm

*Phan Anh Dũng*