

NHÁP

XÂY DỰNG CƠ SỞ TRI THỨC CHỮ NHIỀU BẬC ĐỘ QUY VÀ KHO THÀNH TỔ CƠ BẢN CỦA CHỮ NÔM

Ngô Thanh Giang & Tô Trọng Đức (Nhóm Nôm Na, Hà Nội)
Ngô Thanh Nhân (ĐH Temple) & Ngô Trung Việt (Viện Công nghệ Thông tin Việt Nam)

Tóm tắt

Chữ Nôm hiện nay có thể được sử dụng rộng rãi trong vùng Đông Á và trên thế giới nhờ sự phát triển của ngành tin học, mạng thông tin và nhất là chuẩn mã chữ quốc tế Unicode và ISO/IEC 10646. Tuy thế, thông lệ quốc tế về chữ Hán-Nôm còn nhiều điểm cơ bản chưa chính xác: cụ thể, mỗi chữ Hán-Nôm bị coi là một chữ cái (*character*), và từ đó cách phân tích nội tại của chữ Hán-Nôm còn phải dùng phương pháp bộ và số nét của *Tự điển Khang Hi* làm năm 1710-1716. Từ đó, kho chữ “cái” Hán-Nôm trong bộ chuẩn quốc tế lên đến hơn 71.000, là một điều kỳ lạ. Ai cũng biết chữ Hán-Nôm là một âm tiết, được tạo thành bằng những bộ phận giống nhau, *Tự điển Khang Hi* bắt đầu công tác phân tích và tìm ra 214 bộ (*element, group, bucket*, mà phương Tây dịch thành *radical*). Có thể nói *Tự điển Khang Hi* là một bước tiến cách mạng về mặt phân tích chữ Hán theo các bộ phận tự dạng nội tại của chữ. Nhưng việc dùng cách đếm số nét (không phải là bộ phận tự dạng nội tại) làm phức tạp cho việc tìm chữ trong văn bản hay tự điển—không một người thành thạo chữ Hán Nôm khi nhìn mặt chữ lại nghĩ đến số nét.

Quy trình Nôm Na thiết lập cơ sở tri thức chữ Nôm Việt Nam, ngoài việc tuân thủ các bộ chuẩn Việt Nam và chuẩn quốc tế hiện hành, nó còn giữ thông tin phân tích thành tố của mỗi chữ. Thành tố là một bộ phận của chữ Nôm có nghĩa, là một chữ (hay một bộ) tạo thành chữ mới. Trong bài này, chúng tôi trình bày quy trình Nôm Na, phân tích kho chữ Nôm có sẵn thành một bảng thành tố cơ bản nhất tạo chữ, nhất quán trong toàn bộ kho Nôm Na 21.000 chữ. Phân tích này dùng các mẫu ghép nhị phân của Unicode.

Liên lạc:

Nhóm Nôm Na, Hà Nội

Hội Bảo tồn Di sản chữ Nôm Việt Nam

Số 2 ngõ Hàng Bún, Phố Hàng Bún

Quận Ba Đình, Hà Nội

Điện thoại: +84 4 927 4200

Email: nomna@fpt.vn

DRAFT

BUILDING

AN IDEOGRAPHIC RECURSIVE KNOWLEDGE BASE

AND THE IDEOGRAPHIC REPertoire

FOR THE VIETNAMESE NÔM SCRIPT

Ngô Thanh Giang & Tô Trọng Đức (Nôm Na Group, Hà Nội)
Ngô Thanh Nhân (Temple University) & Ngô Trung Việt (Institute of Informatics)

Abstract

The Vietnamese Nôm script can be accessed widely thanks to the new multilingual Unicode and ISO/IEC 10646 computer character standard, specifically, the UniHan. Nevertheless, the standard has not overcome one basic problem: each ideogram in the standard is a character, which overtly contradicts with the basic principle of 康熙字典 *Kangxi Dictionary*, compiled in 1710-1716, on which the standard is based. The *KangXi Dictionary* describes ideograms in terms of radicals and number of strokes. *KangXi* began with collecting a Chinese ideographic repertoire of 47,035 unique ideograms and sorted them into 214 buckets (called 部 *bộ*, “radical”). 部 *bộ* is a graphic regularity that appears in each ideogram of the bucket. *KangXi* further sorted each bucket into smaller buckets by their number of strokes. *Kangxi* is revolutionary in its time for several reasons: it analyses ideograms *purely graphically*, it considers each ideogram not a character—not as the basic unit of the Chinese script, rather each ideogram can be identified and sorted according to 部 *bộ* plus the number of strokes in the remaining graph. However, UniHan continues to encode over 71,000 ideograms and called them characters. One problem of *KangXi* method is psychologically, no Chinese reader remembers how many strokes an ideogram should have and to which 部 *bộ* “radical” it belongs.

The Nôm Na process establishes an ideographic knowledge base (ikB) for all ideograms encountered that were created and used in Vietnam. The ikB contains, among other information, *KangXi* information, and the binary decomposition of each ideogram into two ideograms. The Nôm Na process inherits the *KangXi* process without using the stroke count. Each element is called an ideographeme. An ideographeme is the smallest graphic element regularly appearing in ideograms. It is thus a radical or another ideogram. In this paper, we describe the Nôm Na process applied uniformly to the Nôm Na repertoire of 21,700 Nôm ideograms, using the Unicode description characters as decomposition order map.