Doc Type: ISO/IEC JTC1/SC2/WG2/IRG
Title: IRGN2673 Vietnamese Hán-Nôm Normalization Guidelines Version 2
Source: Vietnam
Status: National Body Proposal
Authors: Lee Collins, Ngô Trung Việt
Action: For consideration by the IRG
Date: 2024-03-06

Introduction
For encoding purposes, it is useful to have normalization guidelines to assist in the identification of each character. This is particularly important in encoding the traditional Hán-Nôm script used in Vietnam. A key feature of this script is that it was never officially standardized during the period in which it was in most active use, from the 12th to early 20th centuries. Hán characters used in writing literary Chinese tended to follow the more traditional norms from China established prior to the Chinese script simplification carried out at the beginning of the People's Republic of China. The Nôm characters used to write Vietnamese, though, often varied according to the individual author. And, while we can discern some common patterns over time and throughout the source documents, a single character might have multiple forms even within one document. This is especially true of many examples which exist only in manuscript form.

This document attempts to provide such guidelines as can be gleaned from the standard reference works, grammars, and primary texts of Vietnamese literature. These guidelines are embodied in the Hán-Nôm reference font, NomNaTong-regular, which represents the work of specialists working on the encoding of Nôm since the early 1990s.

Detailed Examples
The table below describes the primary normalization rules employed for encoding of the Vietnamese characters in the ISO 10646 / Unicode standard. The primary organization is according to the components used to form Hán-Nôm characters. For each component, we show variant forms, indicate which form is preferred, and provide examples.

Rule	Component	Forms	Preferred	Examples
1-0	Rad. 94	犭, 犭	犭	狂
1.1	Rad. 120	糹, 糸	糹	結, 絲
1-2	Rad. 130	月, 月	月	[illegible], [illegible]
1-3	Rad. 140	䒑, 艹, 艹	艹	茄, [illegible]
1-4	Rad. 162	辶, 辶	辶	迡, 通, [illegible]
1-5	Rad. 182.2	凢, 凡, 凡	凢	舤, 芃
2-0	對 (simp.)	对, 对	对	㨃, [illegible]
2-1	疑 (simp.)	亐, 亏, 亐	亐	[illegible]. 汚, 芌
2-3	難 (simp.)	难, 难	难	[illegible], [illegible]

Rule	Component	Forms	Preferred	Examples
2-4	**羅 (simp.)**	𦉪, 罘, 罗, 罤, 𦊛	𦉪	𪐴, 𨓡
2-5	**鬧 (simp.)**	巿, 苿, 巿	巿	抪, 晰
3-0	**别**	別, 别	別	埛, 捌
3-1	**呂**	呂, 吕	呂	宮. 侣
3-2	**彔**	彔, 录	彔	祿, 綠
3-3	灰	灰, 灰	灰	浗, 炩
3-4	**爭**	爭, 争	爭	崢, 筝, 埩
3-5	**盜**	盗, 盜	盗	瞌, 蹓
3-6	**真**	眞, 真	真	滇, 顛
3-7	**羽**	羽, 羽	羽	翎, 習
3-8	**者**	者, 者	者	箸, 猪
3-9	青	青, 靑	青	情, 精, 清
3-10	角	角, 角	角	觜, 觔
3-11	舟	舟, 舟	舟	般, 艟
4-0	**Stroke position** ㇄	⿺毛, ⿰毛, …	⿺毛	毬, 褐, 触, 毪

Exceptions

1. Multiple forms encoded in Unicode
To avoid confusion, distinctions are maintained as in the case of VN-03B3F (U+3B3F 㬿) and V2-8166 (U+2B736 𫜶).

2. Etymological preservation
The apparent Rad. 140 in VN-F0CB9 寺 is actually the top element of 尊, of which VN-F0CB9 is a simplification. This is not normalized to 艹. A simliar example is the top element of 巿, which is a simplification of 鬧.