L2/17-328

To: Ad Hoc Meeting on Mongolian
From: Ken Whistler, Peter Constable, Roozbeh Pournader, Greg Eck, Liang Hai, Andrew Glass, Lisa Moore, and Debbie Anderson
Date: 9 August 2017
Subject: Script Ad Hoc Group Recommendations on Mongolian Text Model

收件人： 蒙古文特别会议
发件人： Ken Whistler、Peter Constable、Roozbeh Pournader、Greg Eck、梁海、Andrew Glass、Lisa Moore、Debbie Anderson
日期： 2017年8月9日
主题： 文字特别小组关于蒙古文文本模型的建议

An ad hoc group of script experts met several times over the last few months to discuss details of the text model for the Mongolian script.

The high-level goal of that discussion is to find a way to make Mongolian text and documents interchangeable, the script implementable, and platforms supporting Mongolian interoperable. Current implementers such as Microsoft and Google are holding off on fixes for Mongolian until a stable encoding model can be formalized and agreed upon.

In particular, members of the ad hoc group have analyzed different alternatives to fix Hudum, the modern writing system of the modern Mongolian language using the Mongolian script. The side effects on other writing systems are still being studied.

The following summarizes some of the key features of the two main approaches which can be taken for the representation of Hudum text.

文字专家的一个特别小组在最近的几个月中会面了数次以讨论蒙古文的文本模型细节。

该讨论的高级目标是找到办法让蒙古文的文本和文档可交换，文字可实现，而且支持蒙古文的平台可互通。微软和谷歌这样当前的实现方在一个稳定的编码模型能正规化并得到认可之前都会一直推迟针对蒙古文的修正。

特别小组成员尤其分析了各种修正胡都木文（现代蒙古语使用蒙古文的现代书写系统）的替代方案。对其他书写系统的副作用还在研究中。

下面对于两大可用于表记胡都木文本的方法总结了一些关键特性。

1 Graphetic Approach 字形法

This new approach identifies the letters of Mongolian simply based on their shapes and shaping behavior, rather than based on their readings and phonetic values in the modern Mongolian language (or other languages). This is the typical encoding approach used in the Unicode Standard for all cursive joining scripts descended from Aramaic, including the Syriac, Sogdian and Old Uyghur historical precursors to Mongolian. With this approach, a simple adaptation of the Arabic cursive-joining model

这个新方法仅基于外形和成形行为来识别出蒙古文的字母，而不是基于它们在现代蒙古语（或其他语言）中的读法和音值。这个典型的编码方法在Unicode标准中用于包括叙利亚文、粟特文、回鹘文等蒙古文的历史前身在内所有源自阿拉姆文的连写文字。用这个方法，简单适配的阿拉伯文连写模型可以只用相对小的一

can unambiguously handle all of the required shaping behavior for Hudum, with only a relatively small set of contextual rules.

This strategy requires more character additions for encoded representation of Hudum Mongolian: 15 new characters. It would not change any rendering rules or variation sequences for existing characters, so in that sense it is not a breaking change for existing data and fonts.

This strategy would reuse a significant number of the existing consonant letters which currently do not cause problems for text representation.

组上下文规则就无歧义地处理胡都木文要求的所有成形行为。

这个策略要求为胡都木蒙古文的编码表记增加更多字符：15 个新字符。现有字符的任何渲染规则和变体序列都不会改变，所以在此意义上它对现有的数据和字体不是破坏性的改变。

这个策略会复用目前没有给文本表记带来问题的现有辅音字母，数量可观。

Pluses 优点

- Cleaner, unambiguous representation of text.
- Vastly simpler font implementation, with only local contextual rules.
- No variation sequences required for modern Mongolian (except to support old model for backwards compatibility).
- More straightforward user experience; type what you see.
- Much easier searching capability.
- Would be supportable in internationalized domain names.
- Has significantly less security issues than the status quo.

- 更干净、无歧义地表记文本。
- 极大简化的字体实现，只用到局域的上下文规则。
- 对于现代蒙古文不要求变体序列（为向后兼容而支持旧模型除外）。
- 更直接的用户体验，看到什么就键入什么。
- 搜索会容易得多。
- 在国际化域名中会是能支持的。
- 安全问题大大少于现状。

Minuses 缺点

- Radical reorganization of the existing Mongolian text model.
- Departure from Chinese standard for Mongolian.
- Long-term migration issue for data and documents.
- Would require encoding new Mongolian letters (up to 15 characters).
- Phonetic information would need to be carried as metadata.

- 彻底重整现有的蒙古文文本模型。
- 背离中国的蒙古文标准。
- 长期的数据和文档迁移问题。
- 会要求编码新的蒙古文字母（多至 15 个字符）。
- 语音信息会需要携带为元数据。

2 Phonetic Approach 语音法

The current model of the Hudum writing system for Mongolian distinguishes letters based on their readings (current or historic), rather than their shapes per se. In some cases, as for vowels, and notoriously complex pairs, such as Q/G and T/D, this results in very complex shaping rules that sometimes depend on morphological analysis, on long-distance contextual rules (to account for Mongolian vowel harmony), and which also require the definition of numerous variation sequences for cases where the contextual rules fail to produce the expected form, or to deal with historical variant forms.

当前蒙古文的胡都木书写系统模型，在辨别字母时基于（当前或历史上的）读法而非外形本身。对于元音字母以及 Q/G 和 T/D 这样众所周知的成对字母，该模型在一些情况下会导致非常复杂的有时依赖形态学分析的成形规则，以及长程的上下文规则（由于蒙古文元音和谐），而且还要求定义众多的变体序列，用于上下文规则无法产生期望的形式或处理历史变体形式时。

The current model for the Hudum writing system in Unicode employs the phonetic approach. This approach is inherently complex, ambiguous, and difficult to implement successfully. However, in principle, it is possible to make some changes which could improve the phonetic approach incrementally, without changing the overall model. In particular, it could be improved by addressing the theoretical mismatches in interpretation of Mongolian variation sequences, and by making the full set of required contextual rules more explicit for implementers.

胡都木书写系统当前在 Unicode 中的模型使用语音法。这个方法根本上就是复杂、有歧义、难以成功实现的。不过，原则上还是有可能在不改变整体模型的情况下做一些改变来渐进地改善语音法。尤其，如果蒙古文变体序列解读中的理论失配得到解决，并且全套必需的上下文规则都能明确给实现方，就可以改善语音法。

The pros and cons for the phonetic text models are listed below:

下列为语音法文本模型的优点和缺点：

2.1 Status Quo Phonetic 现状语音法

Pluses 优点

- No change to the existing model.
- Basically consistent with the current Chinese standard for Mongolian.
- Phonetic information in the underlying representation may make collation and linguistic analyses easier than a graphetic approach.

- 不改变现有模型。
- 基本与当前中国的蒙古文标准一致。
- 底层表记中的语音信息或许让排序和语言学分析比字形法简单。

Minuses 缺点

- Model is very complex, is underspecified, and has resulted in ambiguous representation of text.
- Assumes very complex contextual rules that have to

- 模型非常复杂，详细规范不足，而且导致有歧义的文本表记。
- 认为非常复杂的上下文规则必须在字

be implemented in fonts and rendering engines, but without clear specification of those rules.

- Requires large set of variation sequences to force particular outcomes for gaps in the rules.
- Very difficult for implementers and font providers.
- Results in de facto non-interoperability of different implementations and fonts.
- Is subject to attempts to “fix” interoperability by an unstable stream of patches with unstable rationales.
- Will not be supported in internationalized domain names.
- For the average user, searching can be very difficult, because input text for matches is ambiguous.
- Has significant security concerns.
- Phonetic information in the text representation is not actually sufficient for text-to-speech or linguistic analysis applications.

体和渲染引擎中实现，但对于那些规则又没有清楚的规范。

- 在规则的空白处要求很大一组变体序列来强行获得特定的结果。
- 对于实现方和字体提供方非常困难。
- 导致不同的实现和字体之间事实上不可互通。
- 会遭受“修正”互通性的尝试，不稳定的连串修补基于不稳定的理由。
- 在国际化域名中不会得到支持。
- 对于一般用户，搜索会很困难，因为用于匹配的输入文本是有歧义的。
- 显著的安全问题令人担忧。
- 对于文本至语音转换和语言学分析用途，文本表记中的语音信息其实并不充分。

2.2 Improved Phonetic 改良语音法

Pluses 优点

- No radical change to the existing model.
- Could be consistent with an updated Chinese standard for Mongolian.
- Requires no new characters and only relatively small list of variation sequence additions.
- Fixes for variation sequences lead to a chance for marginally better interoperability.
- Has a chance to be seen as the “consensus” fix, reducing the need for the dribbling of patches.
- Retains other pluses of the phonetic approach.

- 不彻底改变现有模型。
- 有可能和更新后的中国蒙古文标准一致。
- 不要求新字符，只要求增加相对小的一组变体序列。
- 对变体序列的修正给稍好一些的互通性带来可能。
- 有可能被视为修正的“共识”，减轻了对断断续续修补的需求。
- 保留语音法的其他优点。

Minuses 缺点

- Basically retains the existing phonetic model, with most of its problems.
- To be useful would still require extended and detailed

- 基本上保留了现有的语音模型，及其多数问题。
- 还是要求对隐含的上下文规则提供充

documentation of the implied contextual rules.

- Has the chance of just becoming another “flavor” of Mongolian, incompatible with the other incompatible implementations, without true convergence.
- Would still change rendering of existing text because of differing interpretation of the use of VSes.
- Will not be supported in internationalized domain names.
- For the average user, searching will still be very difficult.
- Still has significant security concerns.

分扩展且详细的说明文档才会有用。

- 有可能只是成为蒙古文编码的又一个“风格”，不兼容于其他互不兼容的实现，达不到真正的融合。
- 还是会改变现有文本的渲染，因为对变体选择符的用途有不同解读。
- 在国际化域名中不会得到支持。
- 对于一般用户，搜索还是会很困难。
- 还是有令人担忧的显著安全问题。

3 Summary 总结

The assessments of the Script Ad Hoc resulting from this review of Mongolian models are:

1. We have unanimous agreement that the status quo is unsustainable. Even with no change to the phonetic text model, existing text is currently not interoperable. Furthermore, the requirement for continued patches to contextual rules, fonts, and sequences, will make text unstable for the foreseeable future.
2. The Graphetic approach is the preferred model; it is far simpler, is internally coherent, and allows unambiguous representation of Mongolian for the Hudum writing system without use of variation sequences. With a simpler model, we expect more developers would be able to create and maintain software and fonts for Mongolian, which is not currently the case. There are migration problems for this approach, but a one-step migration of legacy data could be done to convert the data.
3. The Improved Phonetic strategy retains most of the complexity of the existing Mongolian text model, but would improve its interoperability if widely adopted. For successful interchange and convergence of implementations, it would require clearly specified contextual rules for fonts and rendering engines to be defined and enforced. This may be the easiest alternative to gain support for in China.
4. Regardless of the strategy taken for improving the

根据对蒙古文模型的此次检查，文字特别小组的评估是：

1. 我们一致认为现状不可持续。即使不改变语音法文本模型，现有的文本目前也无法互通。而且因为需要持续修补上下文规则、字体、变体序列，文本在可预见的未来都不会稳定。
2. 字形法是首选的模型。它远比语音法简单，内部一致，而且不用变体序列就能为胡都木书写系统无歧义地表记蒙古文。有了更简单的模型，我们预计更多开发者会有能力为蒙古文创造并维护软件和字体，而当前情况并非如此。这个方法有迁移上的问题，但在转换数据时可以做到对遗留数据的一步迁移。
3. 改良语音法的策略保留了现有蒙古文文本模型的大部分复杂性，但如果得到广泛采用就会改善互通性。为了各实现之间成功的交换与融合，需要为字体和渲染引擎定义清楚的上下文规则规范并强制实施。这可能是最容易在中国获得支持的替代方案。
4. 不管选定哪个策略来改善现代蒙古文

representation of modern Mongolian for the Hudum writing system, some variation sequences may still be needed to represent historic variant glyphs in historic materials. These variation sequences, however, would represent appropriate use of the mechanism to pick out specific historic variants, rather than a complicated set of contextually driven glyph choices for modern Hudum text forced by the phonetic approach of the current Mongolian text model.

5. In any case, a review of the impact of any model changes should be undertaken to verify their impact on the writing of other languages using the Mongolian script (Todo, Sibe, Manchu). Our initial assessment, however, is that the use of the script for those languages is far simpler than for the Hudum writing system, and it is unlikely that model changes for Hudum will impact them very much.

Recommendations: In summary, we recommend that the graphetic text model be adopted for the Mongolian script. Detailed examples of this model (versus the current approach) will be provided in a separate document. A separate document discussing migration issues for the graphitic approach will also be forthcoming.

胡都木书写系统的表记，需要可能还是一些变体序列来表记历史材料中的历史变体图形。不过，这些变体序列会体现如何恰当地使用该机制选出具体的历史变体，而不会是在当前蒙古文文本模型的语音法强迫之下为现代胡都木文本提供复杂的一组上下文驱动的图形选择。

5. 不论如何，应当保证对任何模型更改的影响进行检查，以核对其他使用蒙古文的语言书写（托忒文、锡伯文、满文）受到的影响。不过，我们的初步评估是，那些语言对蒙古文的使用远比胡都木书写系统简单，为胡都木文而做的模型更改不太可能对他们有很大影响。

建议：总的来说，我们建议为蒙古文采用字形法的文本模型。这个模型的详细示例（对比当前方法）会在单独的一篇文档中提供。一篇讨论字形法迁移问题的单独文档也即将到来。