Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Preliminary Review of Proposal on Encoding Khitan Large Script in UCS (N4631)**
Source: **Andrew West, Viacheslav Zaytsev**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2014-10-14**

1. Background

The Khitan Large Script (KLS) is an important historic script, attested in over twenty monumental inscriptions, in a manuscript codex, as well as on coins, seals and portable artefacts. However, the script is largely undeciphered, and the meanings and/or reconstructured pronunciations are only known for a small proportion of characters, with the result that the proposed repertoire in N4631 is largely a list of glyphs of uncertain status.

Encoding KLS is made especially difficult by the fact that there are no surviving contemporary dictionaries, vocabulary lists or any secondary linguistic materials of any kind for either the Khitan language or its two scripts. Moreover, there are no modern dictionaries of Khitan or comprehensive catalogues of KLS characters, so there is no existing character list to base an encoding on.

This means that it is problematic to consider KLS for character encoding at the present time, and much more work on understanding the script may be required before it can be encoded.

2. Relationship with CJK unified ideographs

N4631 stresses that KLS is completely different to CJK characters. Whilst KLS is a separate script, distinct from the Han script, it nevertheless has a very high proportion of characters that are either borrowed from the Han script or which are coincidentally the same as CJK characters, and this may impact the encoding model for the script.

A preliminary review of N4631 suggests that nearly 400 out of the 2,218 proposed characters (about 18% of the repertoire) are identical or unifiably similar in shape to existing encoded CJK ideographs (see Appendix). These graphic correspondences are of three kinds:

1. Chinese characters are borrowed into the KLS script with the same meaning as in Chinese. For example #1244 = U+7687 皇 (huáng) and #1834 = U+5E1D 帝 (dì) are used together in KLS to mean "Emperor", which is the same meaning as the corresponding Chinese word 皇帝 (huángdì); and #0888 = U+56EF 国 (guó) which means "country" in Chinese and KLS (see 3rd line of N4631 Fig. VI for both of these examples).
2. Chinese characters are borrowed into the KLS script, but with a different meaning or used phonetically. For example, #0448 = U+674F 杏 (xìng) means "country"; #0459 = U+7259 牙 (yá) which phonetically represents "ka"; and #0461= U+81F3 至 (zhì) which phonetically represents "an".

3. KLS characters have the same shape as an encoded CJK character, which may be coincidental. For example, #1368 = U+4F1E 伞 ; #1387 = U+91C6 釆 ; and #1796 = U+3CCA 汝 .

The ad hoc meeting on Tangut and Khitan at WG2 M63 (see WG N4642) discussed the possibility of unifying KLS ideographs with CJK unified ideographs:

- do not encode KLS clones of CJK unified ideographs, but use existing encoded CJK unified ideographs for KLS where the glyph forms are unifiable;
- encode KLS ideographs that are not unifiable with CJK unified ideographs in a separate CJK block named something like "CJK Unified Ideographs Khitan Supplement";
- do not define a separate script property for KLS.

However, there was no consensus amongst experts on whether it would be appropriate to unify KLS with CJK unified ideographs or not. The main argument in favour of unification is that encoding hundreds of KLS clones of CJK unified ideographs would be a security risk, but on the other hand unifying KLS with CJK unified ideographs may cause problems for both the CJK and KLS user communities. It would be useful to have guidance from the UTC on this issue.

Note that we do not believe that it would be appropriate to unify the Khitan Small Script (see WG2 N3918) with CJK unified ideographs as Khitan Small Script characters combine into blocks of 2-7 characters, similar to the way in which Hangul jamo combine to form syllables.

3. Unification of glyph variants

A large proportion of the 2,218 characters proposed in N4631 appear to be minor glyph variants which should be unified in a character encoding. For example, #0013 = #0025, #0023 = #0046, #0043 = #0087, #0077 = #0078, #0106 = #0111, #0149 = #0154, #0119 = #1634, #162 = #1734, #0351 = #0366 = #0369, etc. In addition, there are some pairs of proposed characters where there is a slight difference between them which is probably due to an error or damage in the epigraphic source or rubbing, or a mistake in transcription, for example #0070 ≈ #0150, #0148 ≈ #0151, and #0162 ≈ #0163. Indicating the total number of occurences of each proposed character in the KLS corpus would help determine whether a character form was a one-off error or whether it was consistently across multiple inscriptions.

In many cases the fact that pairs of proposed characters are glyph variants can be demonstrated by textual analysis, for example the common Khitan word meaning "preface" or "beginning" is written either as <#0026 #1093 #1791> 未卅没 or <#0026 #1093 #1797> 未卅没 in epigraphic inscriptions, indicating that #1791 没 and #1797 没 are glyph variants of the same character.

In many more cases we may strongly suspect that graphically similar proposed characters are glyph variants of the same character, but it is difficult or impossible to prove one way or the other as we do not know what the characters mean, and the textual evidence is limited.

Without a better understanding of the script it is difficult to define unification principles for KLS, but it is nevertheless clear that unification of glyph variants has not been sufficiently considered with regard to the list of charaters proposed in N4631.

4. Source images

For many proposed characters N4631 shows a cutout of the character from a photograph or rubbing of an inscription it occurs in, which is very useful, and helps determine the correct glyph shape of the proposed character. However, in a few cases no source image is provided: #0396, #0728*, #1047*, #1230, #1272*, #1332, #1345, #1358, #1459*, #1489*, #1538*, #1703, #1921, #2202* (star indicates that no source reference is provided as well as no source image).

We note that in a few cases that the proposed glyph differs significantly from the glyph shape shown in the inscription image, for example #0610, #0635, #1078.

In many cases the source image is a modern hand-drawn glyph, which is probematic as interpretations of glyph shapes in inscriptions by modern scholars are often flawed. In order to ensure that proposed glyph shape of the character is correct it is essential to include rubbings or photographs of the character in an original inscription (rubbings are preferable to photographs for review purposes).

We note that in at least the following cases different proposed characters use the same source image, which must be a mistake:

0005-1 = 0006-1
0077-1 = 0078-1
0112-1 = 0115-1
0168-1 = 0169-1
0220-1 = 0221-1
0529-1 = 0530-1
0532-1 = 0533-1
0575-1 = 1614-1
0639-1 = 0640-1
0704-1 = 0705-1
0739-1 = 0740-1
0739-2 = 0740-2
0763-1 = 0764-1
0766-1 = 0767-1
0849-1 = 0850-1
0883-1 = 0884-1
0938-1 = 1051-1
0951-1 = 0952-1
0963-2 = 0964-1
0995-1 = 1524-1
1023-1 = 1024-1 = 1025-1 = 1026-1
1169-1 = 1170-1
1226-1 = 1227-1
1266-1 = 1267-1
1273-1 = 1274-1
1279-1 = 1280-1
1807-1 = 1808-1
2034-1 = 2035-1
2121-1 = 2122-1
2214-1 = 2214-2

5. Manuscript characters

The proposed character repertoire in N4631 is almost exclusively based on monumental inscriptions, but it should be noted that the single largest KLS text, a manuscript codex in 127 leaves and approximately 15,000 characters in length, which is held at the Institute of Oriental Manuscripts in St Petersburg (https://en.wikipedia.org/wiki/Nova_N_176) is not a source for the proposed KLS encoding. Unfortunately it is difficult to use this manuscript as a source for encoding as it has not yet been published, and the manuscript is written in a cursive hand, so that it is difficult to identify the cursive characters in the manuscript with the standard forms of characters in monumental inscriptions. Nevertheless, this important source should not be overlooked when encoding KLS, and it would be useful to be able to show in the table of proposed characters the cursive forms from this manuscript when known.

Appendix

Mappings between proposed KLS characters in N4631 and CJK unified ideographs are shown below.

0001	U+04E00	一	0138	U+08012	耒	0309	U+05171	共
0002	U+04E8C	二	0145	U+2B1F2	荃	0311	U+053EF	可
0014	U+04E93	亓	0149	U+0592A	太	0312	U+20B9B	𠘨
0015	U+05E72	干	0152	U+04E08	丈	0322	U+05BFA	寺
0017	U+05929	天	0155	U+053BA	厷	0328	U+05734	均
0024	U+07396	玖	0160	U+05935	夵	0329	U+06756	杖
0026	U+0672A	未	0184	U+053CD	反	0336	U+054E5	哥
0027	U+0592B	夫	0187	U+233B4	不	0338	U+04E0B	下
0034	U+04E3C	井	0190	U+053CB	友	0339	U+081E3	臣
0035	U+04E91	云	0194	U+070C8	烈	0344	U+085E5	藥
0038	U+079C3	秃	0210	U+0538B	压	0348	U+04E9A	亚
0047	U+073C0	珀	0223	U+0767E	百	0354	U+2232C	𢌬
0050	U+03EB3	玟	0224	U+04E01	丁	0376	U+08D70	走
0052	U+06709	有	0228	U+05388	厍	0381	U+06B63	正
0053	U+05187	冇	0232	U+077F3	石	0386	U+05B5D	孝
0058	U+20547	冇	0234	U+05426	否	0388	U+0571F	土
0068	U+0593E	夾	0248	U+096E8	雨	0389	U+0572B	圯
0073	U+2124A	巫	0254	U+05357	南	0390	U+2155F	麦
0089	U+0738B	王	0260	U+0897F	西	0391	U+2123D	𠂤
0097	U+0672B	末	0267	U+079B9	禹	0393	U+05730	地
0113	U+04E09	三	0268	U+06771	東	0396	U+090FD	都
0115	U+04E30	丰	0282	U+05DE5	工	0401	U+26AF8	艾
0118	U+05143	元	0299	U+06275	扵	0403	U+05345	卅
0128	U+04E47	乇	0302	U+05349	卉	0409	U+04E94	五

0412	U+081E3	臣	0569	U+26B11	𡈼	0721	U+06C37	氷
0414	U+065E1	旡	0570	U+0624D	才	0730	U+05149	光
0432	U+06B63	正	0574	U+0652F	支	0731	U+05C1A	尚
0436	U+0536D	卬	0579	U+06B79	歹	0734	U+053E3	口
0438	U+05341	十	0580	U+05939	夹	0744	U+05439	吹
0439	U+06728	木	0583	U+03693	𣏌	0757	U+053E9	叩
0440	U+05DEB	巫	0601	U+0673F	朿	0759	U+05144	兄
0442	U+0674E	李	0610	U+0706D	灭	0760	U+053E6	另
0445	U+233B6	杰	0611	U+05927	大	0778	U+0542E	吮
0448	U+0674F	杏	0613	U+08D64	赤	0781	U+05446	呆
0452	U+221B0	玄	0616	U+06C5E	汞	0784	U+053F7	号
0454	U+20AD4	云	0627	U+09EC4	黄	0786	U+20BC8	吠
0459	U+07259	牙	0633	U+04E05	丁	0787	U+0573C	星
0461	U+081F3	至	0635	U+0706D	灭	0792	U+065F2	昊
0469	U+07CFB	系	0637	U+05720	圢	0796	U+06607	昇
0477	U+21B55	丕	0650	U+06B64	此	0798	U+20BC2	孛
0482	U+05BF8	寸	0653	U+05353	卓	0799	U+053F1	叱
0488	U+04F86	來	0669	U+06534	攴	0800	U+065E6	旦
0489	U+06259	扙	0671	U+04E0A	上	0805	U+06772	杲
0493	U+0672D	札	0672	U+053D4	叔	0806	U+065E5	日
0508	U+04E94	五	0684	U+26B11	𡈼	0808	U+065E9	早
0514	U+04E07	万	0701	U+08336	茶	0819	U+07530	田
0550	U+05B89	安	0707	U+0535C	卜	0821	U+07531	由
0554	U+06765	来	0710	U+08096	肖	0823	U+0592E	央
0556	U+05B88	守	0713	U+05763	坐	0828	U+2BA4F	皿
0560	U+0828F	芏	0717	U+04EE5	以	0834	U+053F2	史
0566	U+053CD	反	0719	U+05C16	尖	0836	U+04E32	串

0838	U+20010	虫	1009	U+054C8	哈	1139	U+201AC	伟
0843	U+04E2D	中	1016	U+082AD	芭	1140	U+05316	化
0878	U+20579	冥	1019	U+065F5	昆	1142	U+04EFF	仿
0880	U+0519A	冚	1021	U+07085	炅	1149	U+04F15	伕
0886	U+05198	冘	1038	U+06C34	水	1155	U+04F4F	住
0888	U+056EF	国	1039	U+06C37	氷	1161	U+04F55	何
0896	U+051F9	凹	1040	U+05C0F	小	1162	U+04EC9	仉
0900	U+056E0	因	1045	U+05348	午	1169	U+05316	化
0902	U+20540	冈	1055	U+0624B	手	1178	U+04FE1	信
0905	U+2626B	冊	1057	U+20092	生	1181	U+04F6E	佮
0906	U+2626A	冂	1059	U+06731	朱	1183	U+04F4F	住
0918	U+05185	内	1067	U+25AEB	竿	1186	U+04EC1	仁
0925	U+09580	門	1069	U+0340C	缶	1187	U+05316	化
0930	U+2627C	罚	1072	U+0751F	生	1188	U+04EFF	仿
0934	U+076BF	皿	1073	U+0820C	舌	1190	U+04F0B	伋
0937	U+05DFE	巾	1094	U+05E8A	床	1192	U+04F0E	伎
0945	U+037A2	屮	1097	U+0592D	夭	1195	U+04F88	侈
0953	U+05C71	山	1098	U+079C3	禿	1198	U+04F30	估
0960	U+2BD73	岕	1099	U+079BE	禾	1203	U+04F2F	伯
0976	U+037A4	尖	1102	U+0540C	同	1209	U+03432	伏
0978	U+21D6C	岽	1103	U+0541E	吞	1218	U+04EC3	仃
0979	U+216B4	岌	1108	U+05E01	币	1221	U+04EDB	托
0980	U+05C95	岝	1128	U+0738D	玍	1224	U+04ED8	付
0988	U+076EE	目	1131	U+04E47	毛	1226	U+04F5C	作
0989	U+0660E	明	1132	U+07F36	缶	1228	U+04F4D	位
1007	U+05415	吕	1134	U+05DDD	川	1244	U+07687	皇
1008	U+05CBA	岑	1136	U+04EF2	仲	1250	U+2009D	乑

1252	U+081EA	自	1376	U+0820E	舍	1545	U+2820F	舟
1258	U+09577	長	1383	U+0516C	公	1550	U+05905	夆
1262	U+04E18	丘	1384	U+04EDA	佘	1556	U+0514E	兎
1274	U+0592D	夭	1387	U+091C6	釆	1557	U+04ECC	亽
1281	U+065A4	斤	1392	U+053D7	受	1561	U+04F78	佸
1282	U+0540E	后	1398	U+052FF	勿	1564	U+28478	迩
1283	U+08FD1	近	1405	U+052FA	勺	1565	U+04E1B	丛
1284	U+0821F	舟	1415	U+04E45	久	1569	U+05E01	帀
1285	U+2BE4D	彷	1416	U+04E43	乃	1570	U+04F4F	住
1306	U+223A6	夸	1435	U+05404	各	1575	U+04EDC	仜
1309	U+201CD	尒	1436	U+051AC	冬	1593	U+04F8D	侍
1313	U+05168	全	1442	U+2B762	外	1597	U+04F11	休
1316	U+05750	坐	1443	U+03688	外	1602	U+09AD9	髙
1319	U+03405	乂	1445	U+05905	夆	1603	U+09AD8	高
1321	U+0547D	命	1460	U+2007E	叵	1610	U+04E3B	主
1323	U+05408	合	1463	U+08EAB	身	1615	U+05E8A	床
1324	U+21240	尘	1469	U+080E6	胦	1617	U+07ACB	立
1327	U+05206	分	1472	U+06708	月	1622	U+0342C	亢
1334	U+04EBD	亽	1485	U+051E0	几	1623	U+04EA1	亡
1343	U+04F58	余	1491	U+0670D	服	1627	U+0516D	六
1344	U+04F59	余	1512	U+21240	尘	1634	U+05F03	弃
1352	U+05C12	尒	1519	U+091D1	金	1636	U+221FB	庆
1357	U+08C37	谷	1522	U+06015	怕	1638	U+0653E	放
1360	U+04ECE	从	1531	U+06535	攵	1644	U+04EB0	亰
1366	U+0516B	八	1533	U+052F9	勹	1657	U+07597	疔
1368	U+04F1E	伞	1537	U+0767D	白	1659	U+06587	文
1374	U+04EBA	人	1539	U+079CB	秋	1661	U+05C06	将

1662	U+05317	北
1668	U+0706B	火
1669	U+233E6	柒
1674	U+2083D	劣
1691	U+22189	羊
1699	U+09996	首
1704	U+05DDE	州
1709	U+07C73	米
1716	U+09053	道
1724	U+05F1F	弟
1746	U+0534A	半
1761	U+20BA6	吂
1768	U+05B88	守
1775	U+25927	空
1777	U+05B82	冗
1781	U+05B89	安
1784	U+05B8B	宋
1787	U+03CC9	汨
1791	U+23CDA	沒
1795	U+06C73	汳
1796	U+03CCA	汝
1797	U+06CA1	没
1798	U+06C7A	決
1805	U+06C38	永
1815	U+22A26	岸
1821	U+0623F	房
1825	U+0623B	戾
1826	U+0826F	良
1827	U+04E4B	之
1829	U+06239	戹
1834	U+05E1D	帝
1836	U+04EAC	京
1838	U+25927	空
1858	U+076CA	益
1860	U+20504	竺
1874	U+05FCC	忌
1875	U+0592C	夬
1879	U+28453	迖
1885	U+21C2F	尾
1886	U+05F13	弓
1913	U+0541B	君
1931	U+06BBF	殿
1938	U+05DF1	己
1941	U+05DF3	巳
1949	U+05C38	尸
1953	U+05C45	居
1972	U+04E5F	也
1979	U+04E11	丑
1991	U+05426	否
2009	U+04E43	乃
2011	U+0529B	力
2017	U+08FB9	边
2026	U+05723	圣
2043	U+0543E	吾
2062	U+04E86	了
2066	U+05B50	子
2067	U+05B56	孖
2068	U+05B54	孔
2072	U+04E5C	乜
2076	U+2007E	吗
2104	U+053C9	叉
2106	U+28D17	閫
2120	U+04E5A	乚
2144	U+05148	先
2154	U+077E2	矢
2157	U+05973	女
2165	U+05983	妃
2177	U+05F01	弁
2183	U+053F0	台
2186	U+053C3	參
2192	U+053BC	尐
2193	U+06BCB	毋
2197	U+07D02	紂
2198	U+07D71	統
2206	U+05E7C	幼
2213	U+05C14	尔