

N-014

# Design of Sogd Character Information Processing System

Omerjan Osman†

Katsuko T. Nakahira†

Yoshiki Mikami†

## Abstract

The Sogdians were ancient people of Central Asia, who inhabited the region known to the west as Sogdiana. The Sogdian language was widely spoken in the region from 2<sup>nd</sup> century BC to 8<sup>th</sup> century AD and written with Sogd script. The oldest manuscript written in Sogd script dates back to the 2<sup>nd</sup> century BC. Although the rich cultural heritage was recorded in the script, the script has completely extinct today. In order to make those valuable archives available online, Sogd language processing system is needed. Authors are trying to create a standard character set for the ancient Sogd script as a first step to this objective. In this paper, the first version of standard Sogd character set for inclusion to Unicode is presented.

## Keywords

Sogd, character set, character code, Unicode

## 1 Introduction

### 1.1 Sogd people and language

The Sogdians were ancient people of Central Asia, who inhabited the region known to the west as Sogdiana (Zarafshan River Valley). Sogdiana covered much of the territory of modern-day Tajikistan, southern Uzbekistan, and northern Afghanistan. Chief cities of the region are Samarkand, Panjaket, Fergana. The region name Sogd was mentioned in the Avesta (the primary collection of sacred texts of Zoroastrianism).

The Sogd language is a Middle Iranian language. The language is usually assigned to the Northeastern branch of the Iranian languages. Like all the writing systems employed for Middle Iranian languages, the Sogdian script ultimately derives from the Aramaic script. The oldest Sogdian document dates back to the 2<sup>nd</sup> century BC and had been in use until 8<sup>th</sup> to 9<sup>th</sup> century AD.

### 1.2 Evolution of Sogd Script

The Sogd script is occasionally known as the “sutra script”, because many Buddhist, Manichaean, Nestorian, and Zoroastrian texts as well as all secular material such as letters, legal documents, coin legends, and inscriptions were written in this script through the history.

The Sogd script is derived from the Aramaic script and is the direct ancestor of the Uyghur script, itself the forerunner of the Mongolian script, Manchu Script, Buryat Script and Kalmyk Script (Figure 1).

Sogd script had undergone changes over time. The ancient Sogd script, used in early years looks quite similar to its ancestor, the Aramaic script. A typical sample manuscript of this date is

shown in Figure 2. It does not have vowel letters. It is composed of 23 consonant letters. Most of the letters are distinct and does not change shape when joined [Skjærvø].

In the later years, from 4<sup>th</sup> century AD to 8<sup>th</sup> century AD, vowel letters were added and text gradually became to be written in joint. Typical samples of manuscript written in later years are shown in Figure 2.

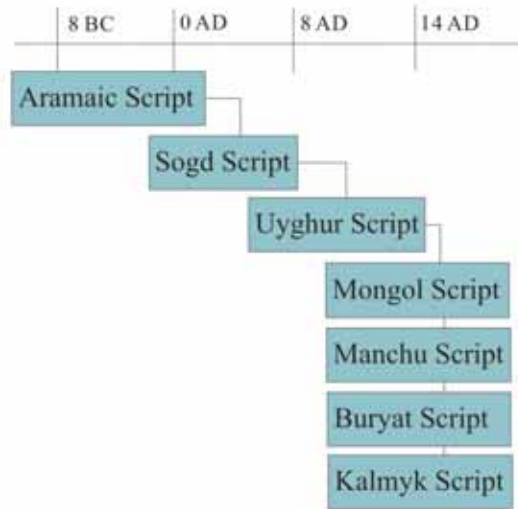


Figure 1. Evolution of Sogd script.

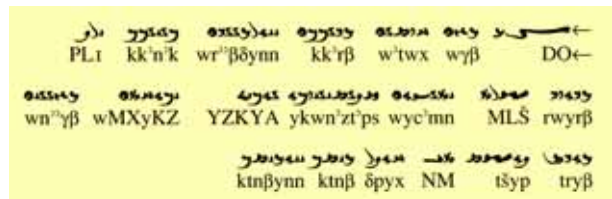


Figure 2. Old Sogd manuscript sample (circa 1st century AD) source: P . Oktor Skjærvø, Aramaic Scripts for Iranian Languages, in [Daniels, p.529]

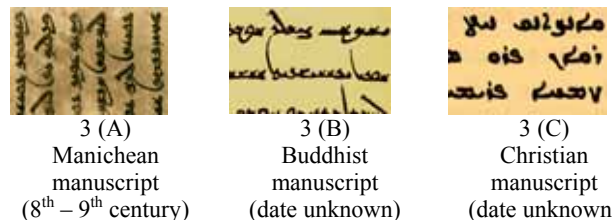


Figure 3. Various Sogd manuscripts.

source: (left) A manuscript kept in Museum fur Indische Kunst, Berlin [Coulmas, p.473]; (center) A Buddhist Sogdian texts kept in British Library (Acta Iranica 10, Brill, Liege) [Sanseido, p.554]; (right) A page taken from [Daniels, p.533]

† Nagaoka University of Technology

### 1.3 Sogd Alphabets

Although Soghd language is an Eastern Iranian language, the Sogd script itself had evolved to those scripts used for many non-Iranian languages, in particular Turkic languages, such as old Uyghur and other eastern Turkic languages. But these were generally superseded by versions of the Arabic alphabet after the conversion of the Turkic peoples to Islam.

The Sogd script is written in horizontal writing from right to left and in vertical writing from top to bottom. The Sogd character set used in the later stage of the script history, around 4<sup>th</sup> to 8<sup>th</sup> century AD, consists of 33 letters, 10 vowel letters plus 23 consonant letters. Most of the letters has 4 kinds of shapes, such as initial, medial, final, and isolated form (Table 1). Sogd character set also contains diacritical marks and a few punctuation symbols. A list of all those categories of letters is given in Table 2.

In Table 1, the Sogd font designed by one of the authors were created based on the shapes which appear in a Sogd Buddhist manuscript shown in Figure 3 (B).

Table 2. Classification of Sogd Script.

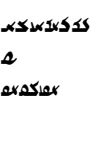
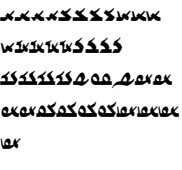
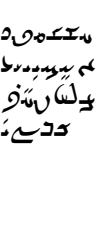
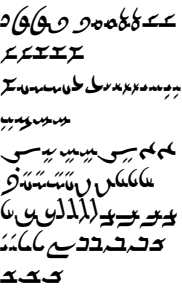



Acronum	Letter Name	Code	Glyph	Code Point
				Glyph Point
V	Sogd Vowels Letter			U+X000~X009 U+X030~X039
				U+0000~0027 U+0070~0097
C	Sogd Consonants Letter			U+X00A~X020 U+X03A~X050
				U+0028~0068 U+0098~00D8
F	Sogd Letter Diacritical Marks			U+X021~X024 U+X051~X054
				Nothing
D	Sogd Digit			U+X02D~X02F U+X05D~X05F
				Nothing
P	Sogd Punctuation			U+X027~X028 U+X057~X058
				Nothing

Table 1. Sogd Alphabet. Horizontally-written

No	Isolated	Final	Medial	Initial	Hebrew name	Uyghur name	Sound value
1	𐰇	𐰇	𐰇	𐰇	ALEPH	ALEPH	/ a /
2	𐰉	𐰉	𐰉	𐰉		AH	/ā/
3	𐰊	𐰊	𐰊	𐰊		E	/e/
4	𐰋	𐰋	𐰋	𐰋		EY	/ē/
5	𐰌	𐰌	𐰌	𐰌		I	/i/
6	𐰍	𐰍	𐰍	𐰍		IY	/ī/
7	𐰎	𐰎	𐰎	𐰎	WAW	O	/o/
8	𐰏	𐰏	𐰏	𐰏		OV	/ō/
9	𐰐	𐰐	𐰐	𐰐		U	/u/
10	𐰑	𐰑	𐰑	𐰑	YODH	UV	/ū/
11	𐰒	𐰒		𐰒		BETH	/b/
12	𐰓			𐰓	PE	PE	/p/
13	𐰔	𐰔	𐰔	𐰔	TAU	TETH	/t/
14	𐰕	𐰕	𐰕	𐰕		ZHE	/ ʈ /
15	𐰖	𐰖	𐰖	𐰖	TSADI	CHE	/ tʃ /
16	𐰗	𐰗	𐰗	𐰗	HETH	CHETH	/x/
17	𐰘			𐰘	LAMED H	DALET H	/d/
18	𐰙	𐰙	𐰙	𐰙	RESH	RESH	/r/
19	𐰚		𐰚	𐰚	ZAIN	ZAIN	/z/
20	𐰛	𐰛	𐰛	𐰛	MARK EDZ	ZHEE	/ /
21	𐰜	𐰜	𐰜	𐰜	SHIN	SAMEC H	/s/
22	𐰝	𐰝	𐰝	𐰝	MARK EDS	SCHIN	/ ʃ /
23	𐰞			𐰞	GIMEL	GIMEL	/ γ /
24	𐰟					VAU	/f/
25	𐰠	𐰠	𐰠	𐰠	<sup>2</sup> - DOTTE D	KOPH	/q/
26	𐰡	𐰡	𐰡	𐰡	KAPH	KAPH	/k/
27	𐰢	𐰢	𐰢	𐰢		GE	/g/
28	𐰣	𐰣	𐰣	𐰣	HOOK DR	LAMED	/l/
29	𐰤	𐰤	𐰤	𐰤	MEM	MEM	/m/
30	𐰥	𐰥	𐰥	𐰥	NUN	NUN	/n/
31	𐰦					HEE	/h/
32	𐰧	𐰧	𐰧	𐰧	BETH	VE	/v/
33	𐰨	𐰨	𐰨	𐰨	YODH	JOD	/j/

## 2 Sogd Language and Script in Standards

The Sogd language and Sogd script do not appear in relevant international standards yet. While some of the ancient scripts, such as Brahmi, Kharoshthi and Orkhon scripts appear in ISO 15924 Codes for the representation of names of scripts, Sogd is not included in the standard. ISO/IEC 10646 and Unicode, of course, does not cover Sogd script so far (Table 3).

Also Sogd language is present only in ISO 639-2 three letter code (Alpha-3) and not presented in ISO 639-1 two letter code (Alpha-1). (see Table 4)

**Table 3. Selected Scripts in Relevant International Standards**

Script Name	ISO 15924 Script code	ISO/IEC 10646 Unicode
Aramaic	not present	not present
Brahmi	Brah	not present
Kharoshthi	Khar	Kharoshthi
Sogdian (Sogdish)	not present	not present
Orkhon	Orkh	not present
Uyghur	not present	not present
Arabic	Arab	Arabic
Mongolian	Mong	Mongolian
Manchu	not present	not present
Buryat	not present	not present
Kalmyk (Oirat)	not present	not present

**Table 4. Selected Languages in ISO 630 Language Code**

Script Name	ISO 639-1	ISO 639-2
Aramaic	---	arc
Sogdian (Sogdish)	---	sog
Uyghur	ug	uig
Arabic	ar	ara
Mongolian	mn	mon
Manchu	---	mnc
Buryat	---	bua
Kalmyk (Oirat)	---	---

## 3 Sogd Character Code Table

Based on above studies, a first version of Sogd character code table is proposed in Table 7.

Table 7 composed of Sogd Characters Horizontally-written (U+X000 ~ X02F), Sogd Characters Vertically-written (U+X030 ~ X05F).

Also Sogd Glyphs Horizontally-written (U+0000 ~ 0068), and Sogd Glyphs Vertically-written (U+0070 ~ 00D8) are included in Table 7. Categories of those code points are shown in Table 8.

Display/rendering processes must select an appropriate glyph form to depict each Sogd letter according to its immediate joining context; furthermore, it must substitute certain ligature glyphs for sequences of Sogd characters.

The appropriate form is determined on the basis of its joining class and the joining class of adjacent characters. Each Sogd character falls into one of the classes shown in Table 5 and Table 6.

**Table 5. Sogd Horizontally-written Joining Glyph Types**

Glyph Types	Description
Xn( Isolated)	Nominal glyph form as it appears in the code charts
Xr( Final)	Right-joining glyph form (both right-joining and dual-joining characters may employ this form)
Xl( Initial)	Left-joining glyph form (both left-joining and dual-joining characters may employ this form)
Xm( Medial)	Dual-foining (medial) glyph form that joins on both left and right (only dualjoining characters employ this form )

**Table 6. Sogd Vertically-written Joining Glyph Types**

Glyph Types	Description
Xn( Isolated)	Nominal glyph form as it appears in the code charts
Xt( top )	Top-joining glyph form (both top-joining and dual-joining characters may employ this form)
Xb( Bottom)	Bottom-joining glyph form (both bottom-joining and dual-joining characters may employ this form)
Xm( Medial)	Dual-foining (medial) glyph form that joins on both left and right (only dualjoining characters employ this form )

## 4 Conclusion

Sogd script used in 4<sup>th</sup> to 8<sup>th</sup> century is studied and the character code for Sogd script is proposed for possible inclusion to ISO/IEC 10646 and Unicode. Due to limitations of time and available materials, only limited number of manuscripts were studied so that further study of other Sogd manuscripts should be done before proposing it to standard developing forum.

## 5 References

1. Peter T. Daniels, William Bright, The World's Writing Systems, New York Oxford (1996).
2. Florian Coulmas, The Blackwell Encyclopedia of Writing Systems (1999).
3. Rokuro Kono, Eiichi, Chino, Tatsuo Nishida, The Sanseido Encyclopedia of Linguistics (2001).
4. G. R. Rachmati, Ergebnisse Der Deutschen Turfan-Forschung (1936).

Table 7. Sogd Character Code and Glyph Table.

	X00	X01	X02	X03	X04	X05	000	001	002	003	004	005	006	007	008	009	00A	00B	00C	00D
0	𐰀	𐰁	𐰂	𐰃	𐰄	𐰅	𐰆	𐰇	𐰈	𐰉	𐰊	𐰋	𐰌	𐰍	𐰎	𐰏	𐰐	𐰑	𐰒	𐰓
X000	X010	X020	X030	X040	X050	0000	0010	0020	0030	0040	0050	0060	0070	0080	0090	00A0	00B0	00C0	00D0	
1	𐰔	𐰕	𐰖	𐰗	𐰘	𐰙	𐰚	𐰛	𐰜	𐰝	𐰞	𐰟	𐰠	𐰡	𐰢	𐰣	𐰤	𐰥	𐰦	𐰧
X001	X011	X021	X031	X041	X051	0001	0011	0021	0031	0041	0051	0061	0071	0081	0091	00A1	00B1	00C1	00D1	
2	𐰨	𐰩	𐰪	𐰫	𐰬	𐰭	𐰮	𐰯	𐰰	𐰱	𐰲	𐰳	𐰴	𐰵	𐰶	𐰷	𐰸	𐰹	𐰺	𐰻
X002	X012	X022	X032	X042	X052	0002	0012	0022	0032	0042	0052	0062	0072	0082	0092	00A2	00B2	00C2	00D2	
3	𐰼	𐰽	𐰾	𐰿	𐱀	𐱁	𐱂	𐱃	𐱄	𐱅	𐱆	𐱇	𐱈	𐱉	𐱊	𐱋	𐱌	𐱍	𐱎	𐱏
X003	X013	X023	X033	X043	X053	0003	0013	0023	0033	0043	0053	0063	0073	0083	0093	00A3	00B3	00C3	00D3	
4	𐱐	𐱑	𐱒	𐱓	𐱔	𐱕	𐱖	𐱗	𐱘	𐱙	𐱚	𐱛	𐱜	𐱝	𐱞	𐱟	𐱠	𐱡	𐱢	𐱣
X004	X014	X024	X034	X044	X054	0004	0014	0024	0034	0044	0054	0064	0074	0084	0094	00A4	00B4	00C4	00D4	
5	𐱤	𐱥	𐱦	𐱧	𐱨	𐱩	𐱪	𐱫	𐱬	𐱭	𐱮	𐱯	𐱰	𐱱	𐱲	𐱳	𐱴	𐱵	𐱶	𐱷
X005	X015	X025	X035	X045	X055	0005	0015	0025	0035	0045	0055	0065	0075	0085	0095	00A5	00B5	00C5	00D5	
6	𐱸	𐱹	𐱺	𐱻	𐱼	𐱽	𐱾	𐱿	𐲀	𐲁	𐲂	𐲃	𐲄	𐲅	𐲆	𐲇	𐲈	𐲉	𐲊	𐲋
X006	X016	X026	X036	X046	X056	0006	0016	0026	0036	0046	0056	0066	0076	0086	0096	00A6	00B6	00C6	00D6	
7	𐲌	𐲍	𐲎	𐲏	𐲐	𐲑	𐲒	𐲓	𐲔	𐲕	𐲖	𐲗	𐲘	𐲙	𐲚	𐲛	𐲜	𐲝	𐲞	𐲟
X007	X017	X027	X037	X047	X057	0007	0017	0027	0037	0047	0057	0067	0077	0087	0097	00A7	00B7	00C7	00D7	
8	𐲠	𐲡	𐲢	𐲣	𐲤	𐲥	𐲦	𐲧	𐲨	𐲩	𐲪	𐲫	𐲬	𐲭	𐲮	𐲯	𐲰	𐲱	𐲲	𐲳
X008	X018	X028	X038	X048	X058	0008	0018	0028	0038	0048	0058	0068	0078	0088	0098	00A8	00B8	00C8	00D8	
9	𐲴	𐲵	𐲶	𐲷	𐲸	𐲹	𐲺	𐲻	𐲼	𐲽	𐲾	𐲿	𐳀	𐳁	𐳂	𐳃	𐳄	𐳅	𐳆	𐳇
X009	X019	X029	X039	X049	X059	0009	0019	0029	0039	0049	0059	0069	0079	0089	0099	00A9	00B9	00C9	00D9	
A	𐳈	𐳉	𐳊	𐳋	𐳌	𐳍	𐳎	𐳏	𐳐	𐳑	𐳒	𐳓	𐳔	𐳕	𐳖	𐳗	𐳘	𐳙	𐳚	𐳛
X00A	X01A	X02A	X03A	X04A	X05A	000A	001A	002A	003A	004A	005A	006A	007A	008A	009A	00AA	00BA	00CA	00DA	
B	𐳜	𐳝	𐳞	𐳟	𐳠	𐳡	𐳢	𐳣	𐳤	𐳥	𐳦	𐳧	𐳨	𐳩	𐳪	𐳫	𐳬	𐳭	𐳮	𐳯
X00B	X01B	X02B	X03B	X04B	X05B	000B	001B	002B	003B	004B	005B	006B	007B	008B	009B	00AB	00BB	00CB	00DB	
C	𐳰	𐳱	𐳲	𐳳	𐳴	𐳵	𐳶	𐳷	𐳸	𐳹	𐳺	𐳻	𐳼	𐳽	𐳾	𐳿	𐴀	𐴁	𐴂	𐴃
X00C	X01C	X02C	X03C	X04C	X05C	000C	001C	002C	003C	004C	005C	006C	007C	008C	009C	00AC	00BC	00CC	00DC	
D	𐴄	𐴅	𐴆	𐴇	𐴈	𐴉	𐴊	𐴋	𐴌	𐴍	𐴎	𐴏	𐴐	𐴑	𐴒	𐴓	𐴔	𐴕	𐴖	𐴗
X00D	X01D	X02D	X03D	X04D	X05D	000D	001D	002D	003D	004D	005D	006D	007D	008D	009D	00AD	00BD	00CD	00DD	
E	𐴘	𐴙	𐴚	𐴛	𐴜	𐴝	𐴞	𐴟	𐴠	𐴡	𐴢	𐴣	𐴤	𐴥	𐴦	𐴧	𐴨	𐴩	𐴪	𐴫
X00E	X01E	X02E	X03E	X04E	X05E	000E	001E	002E	003E	004E	005E	006E	007E	008E	009E	00AE	00BE	00CE	00DE	
F	𐴬	𐴭	𐴮	𐴯	𐴰	𐴱	𐴲	𐴳	𐴴	𐴵	𐴶	𐴷	𐴸	𐴹	𐴺	𐴻	𐴼	𐴽	𐴾	𐴿
X00F	X01F	X02F	X03F	X04F	X05F	000F	001F	002F	003F	004F	005F	006F	007F	008F	009F	00AF	00BF	00CF	00DF	

Table 8. Categories of Code Points in Table 7.

No	Categories	Explanation	Code	total number
1	Horizontally Vowel Letter	Designate 10 Vowel (Horizontally-written from right to left )	X000~X009	10
2	Vertically Vowel Letter	Designate 10 Vowel (Vertically-written from top to bottom)	X030~X039	10
3	Horizontally Consonant Letter	Designate 23 Consonant (Horizontally-written from right to left )	X00A~X020	23
4	Vertically Consonant Letter	Designate 23 Consonant (Vertically-written from top to bottom)	X03A~X050	23
5	Horizontally Diacritical Marks	Diacritical Marks (Horizontally-written from right to left )	X021~X024	4
6	Vertically Diacritical Marks	Diacritical Marks (Vertically-written from top to bottom)	X051~X054	4
7	Horizontally Punctuation Symbol	Punctuation Symbol (Horizontally-written from right to left )	X027~X028	2
8	Vertically Punctuation Symbol	Punctuation Symbol (Vertically-written from top to bottom)	X057~X058	2
9	Horizontally Digit	Digit Horizontally-written from right to left	X02D~X02F	3
10	Vertically Digit	Digit Vertically-written from top to bottom	X05D~X05F	3