Character Encoding

This section describes the character encoding mechanisms for HTTP requests to Tamino and responses from Tamino.

The term "encoding" in this section is used with the semantic defined in the W3C XML specification at http://www.w3.org/TR/REC-xml/. The terms "charset" and "character set" are used with the semantic defined in the HTTP/1.1 description at http://www.ietf.org/rfc/rfc2616.txt.

Character Encoding of Input Documents

Input documents can be supplied for X-Machine commands such as `_process` and `_define`. The encoding of an input document can be specified explicitly in several ways:

- in the `encoding` attribute of the document's XML declaration
- in the `_encoding` parameter passed in the X-Machine command
- in the `charset` value that is defined in conjunction with the document's `Content-Type` parameter in the HTTP request

If the encoding is not specified in one of these ways, the document is assumed to be encoded according to the value of the server XML parameter `XML document default encoding` (for details see the list of server XML properties in the section Creating a Database in the documentation for the Tamino Manager).

All input documents with top-level media type "text" are converted to Unicode. Input data is converted from the client's encoding to Unicode. The original encoding of the input is not remembered. X-Machine uses the internet standards for character set names as defined in the document http://www.iana.org/assignments/character-sets.

Hint for users of Microsoft Windows: Please note that Microsoft code page 1252 is close to but not identical with ISO-8859-1 (latin1).

Example

Database queries specifying character encoding can be sent to the X-Machine using the X-Machine command `_encoding` followed for example by the X-Machine command `_xql` in a single HTTP request, for example

```
http://myhost:80/tamino/mydb/mycollection?_encoding=utf-8&_xql=patient/name[surname="Bloggs"].
```

The value of the `_encoding` parameter will be applied to the values of all commands that are subsequently executed. See also the section *Order of Execution of Commands*.

Character Encoding of Output Documents

Output documents are converted by the X-Machine to the encoding desired by the client. Character references are used to represent characters that do not exist in the desired encoding. The desired encoding of the output can be specified in the HTTP header "Accept-Charset". If "Accept-Charset" is omitted, X-Machine uses the encoding of the client request.

Supported Character Encodings

The Tamino server supports all standard character encodings and their well known aliases, as shown in the following list.

Note:
It is possible that some Tamino product components do not support some of these encodings. Please see the documentation for the individual developer components for a list of their supported encodings.

Encoding Name	Well known aliases
Adobe-Standard-Encoding	csAdobeStandardEncoding
Big5	950, cp950, csBig5, ibm-1370_VSUB_VPUA, x-big5
CESU-8	
cp850	850, csPC850Multilingual, IBM850
cp851	851, csPC851, IBM851
cp856	856, ibm-856
cp857	857, csIBM857
cp858	IBM00858
cp859	
cp860	860, csIBM860, IBM860
cp861	861, cp-is, csIBM861, IBM861
cp862	862, cp867, cspc862latinhebrew
cp863	cp863, csIBM863, IBM863
cp864	csIBM864
cp865	865, csIBM865, IBM865
cp866	866, csIBM866
cp868	868, cp-ar, csIBM868, IBM868
cp869	869, cp-gr, csIBM869
cp921	921
cp922	922

Encoding Name	Well known aliases
EUC-JP	csEUCPkdFmtJapanese, eucjis, Extended_UNIX_Code_Packed_Format_for_Japanese, ibm-33722_VPUA, ibm-eucJP, X-EUC-JP
EUC-KR	csEUCKR, ibm-970_VPUA, ibm-eucKR, X-EUC-KR
gb18030	ibm-1392
GB2312	1383, chinese, cp1383, csGB2312, csISO58GB231280, EUC-CN, gb, gb2312-1980, GB_2312-80, ibm-1383, ibm-1383_VPUA, ibm-eucCN, iso-ir-58, X-EUC-CN
GBK	CP936, ibm-1386_VSUB_VPUA, MS936, zh_cn, windows-936
hp-roman8	csHPRoman8, r8, roman8
HZ-GB-2312	HZ
IBM01140	CCSID01140, CP01140, cpibm1140, ebcdic-us-37+euro
IBM01141	CCSID01141, CP01141, cpibm1141, ebcdic-de-273+euro
IBM01142	CCSID01142, CP01142, cpibm1142, ebcdic-dk-277+euro, ebcdic-no-277+euro
IBM01143	CCSID01143, CP01143, cpibm1143, ebcdic-fi-278+euro, ebcdic-se-278+euro
IBM01144	CCSID01144, CP01144, cpibm1144, ebcdic-it-280+euro
IBM01145	CCSID01145, CP01145, cpibm1145, ebcdic-es-284+euro
IBM01146	CCSID01146, CP01146, cpibm1146, ebcdic-gb-285+euro
IBM01147	CCSID01147, CP01147, cpibm1147, ebcdic-fr-297+euro
IBM01148	CCSID01148, CP01148, cpibm1148, ebcdic-international-500+euro
IBM01149	CCSID01149, CP01149, cpibm1149, ebcdic-is-871+euro
IBM037	cpibm37, ebcdic-cp-us, ebcdic-cp-ca, ebcdic-cp-wt, ebcdic-cp-nl, cp37, cp037, 037
IBM1026	CP1026, csIBM1026, Ibm-1026_STD
IBM273	273, CP273, cpibm273, csIBM273, ebcdic-de
IBM277	277, csIBM277, cpibm277, EBCDIC-CP-DK, EBCDIC-CP-NO, ebcdic-dk
IBM278	278, cp278, cpibm278, csIBM278, ebcdic-cp-fi, ebcdic-cp-se, ebcdic-sv
IBM280	280, CP280, cpibm280, csIBM280, ebcdic-cp-it
IBM284	284, CP284, cpibm284, csIBM284, ebcdic-cp-es
IBM285	285, CP285, cpibm285, csIBM285, ebcdic-cp-gb, ebcdic-gb
IBM290	cp290, csIBM290, EBCDIC-JP-kana
IBM297	297, cp297, cpibm297, csIBM297, ebcdic-cp-fr
IBM367	

Encoding Name	Well known aliases
IBM420	420, cp420, csIBM420, ebcdic-cp-ar1
IBM424	424, cp424, csIBM424, ebcdic-cp-he
IBM500	500, CP500, cpibm500, csIBM500, ebcdic-cp-be, ebcdic-cp-ch
IBM852	
IBM855	
IBM857	
IBM862	
IBM864	
IBM869	
IBM870	CP870, csIBM870, ibm-870, ibm-870_STD, ebcdic-cp-roece, ebcdic-cp-yu
IBM871	871, CP871, cpibm871, csIBM871, ebcdic-cp-is, ebcdic-is
IBM918	CP918, csIBM918, , ebcdic-cp-ar2, ibm-918_STD, ibm-918_VPUA
ISO-2022-CN-EXT	
ISO-2022-CN	
ISO-2022-JP-2	csISO2022JP2
ISO-2022-JP	csISO2022JP
ISO-2022-KR	csISO2022KR
ISO-2022	2022, cp2022
iso-8859-15	
ISO-8859-1	8859-1, cp819, csISOLatin1, IBM819, ISO_8859-1:1987, iso-ir-100, l1, latin1
iso-8859-2	8859-2, 912, cp912, csISOLatin2, ISO_8859-2:1987, iso-ir-101, l2, latin2
iso-8859-3	8859-3, 913, cp913, csISOLatin3, iso-ir-109, l3, latin3
iso-8859-4	8859-4, 914, cp914, csISOLatin4, ISO_8859-4:1988, iso-ir-110, l4, latin4
iso-8859-5	8859-5, 915, cp915, csISOLatinCyrillic, cyrillic, ISO_8859-5:1988, iso-ir-144
iso-8859-6	1089, 8859-6, arabic, asmo-708, cp1089, csISOLatinArabic, ecma-114, ISO_8859-6:1987, iso-ir-127
iso-8859-7	813, 8859-7, cp813, csISOLatinGreek, ecma-118, elot_928, greek, greek8, ISO_8859-7:1987, iso-ir-126
iso-8859-8	916, cp916, csISOLatinHebrew, Hebrew, 8859-8, ISO_8859-8:1988, iso-ir-138
iso-8859-9	8859-9, 920, cp920, latin5, csISOLatin5, ISO_8859-9:1989, iso-ir-148, l5
JIS_Encoding	ISO-2022-JP-1, JIS

Encoding Name	Well known aliases
KOI8-R	cp878, cskoi8r, koi8
KSC_5601	949, csKSC56011987, ibm949, ibm949_VSUB_VPUA, iso-ir-149, johab, Korean, ksc5601_1992, KS_C_5601-1987, KS_C_5601-1989, ks_x_1001:1992
mac	csMacintosh
SCSU	
Shift_JIS	943, cp943, cp932, csShiftJIS, csWindows31J, MS_Kanji, pck, sjis, windows-31j, x-sjis
TIS-620	874, cp874, cp9066, ms874, windows-874
US-ASCII	ANSI_X3.4-1968, ASCII, ANSI_X3.4-1986, cp367, csASCII, ISO_646.irv:1983, ISO_646.irv:1991, ISO646-US, iso-ir-6, us
UTF-16BE	cp1201, UTF16_BigEndian, x-utf-16be
UTF-16LE	cp1200, UTF16_LittleEndian, x-utf-16le
UTF-32BE	UTF32_BigEndian
UTF-32LE	UTF32_LittleEndian
UTF-7	cp65000
UTF-8	cp1208, cp65001
UTF-16	csUnicode, ISO-10646-UCS-2, ucs-2
UTF-32	csUCS4, ISO-10646-UCS-4, ucs-4
windows-1250	cp1250
windows-1251	cp1251
windows-1252	cp1252
windows-1253	cp1253
windows-1254	cp1254
windows-1255	cp1255
windows-1256	cp1256
windows-1257	cp1257
windows-1258	cp1258