## LISTSERV Maestro Admin Tech Doc 19

# International Character Sets

September 30, 2021 | © L-Soft Sweden AB
**lsoft.com**

This document is a LISTSERV Maestro Admin Tech Doc. Each admin tech doc documents a certain facet of the LISTERV Maestro administration on a technical level. This document is number 19 of the collection of admin tech docs and explains the topic “International Character Sets”.

Last updated for LISTSERV Maestro 10.0-1 on September 30, 2021. The information in this document also applies to later LISTSERV Maestro versions, unless a newer version of the document supersedes it.



All of L-Soft's manuals are also available at: **http://www.lsoft.com/manuals.html**

L-Soft invites comment on its manuals. Please feel free to send your comments by e-mail to: MANUALS@LSOFT.COM

# Table of Contents

# 1 International Character Sets from a LISTSERV Maestro User's Perspective

## *1.1 Introduction to International Character Sets*

Since computers store all information in form of "bits", those bits (or their 8-bit conglomerations "bytes") are also the entities that are transferred from the sender to the recipient whenever an e-mail is sent.

However, the recipient is obviously not very interested in a sequence of bits or bytes – she wants to read text. Therefore, the bit sequences need to be interpreted as characters, for display to the reader.

For this mapping of bit sequences to characters from an alphabet, several different so called "character sets" (short: "charsets") have been defined and standardized by the international community.

In the English speaking world, probably the most widely used charset is **ASCII** (sometimes also called US-ASCII), which is a charset that maps **7-bit** sequences to the 26 characters from the Latin alphabet. Actually, since 7 bits have enough room for 128 characters (0-127), there are more than only the 26 Latin characters in the ASCII charset: First, each character appears twice (as upper case and lower case), then there are the ten digits 0-9, various punctuation marks like comma, dot, semi-colon, colon, dash, slash, backslash, exclamation and question mark, etc. plus a range of other more "dubious" control characters.

Almost equally often used, at least in the western world, are the charsets from the **ISO-8859** family. These charsets map **8-bit** sequences to letters, digits and characters from various European languages, including Hebrew and Arabic. Since the ISO-8859 charsets use 8 bits, they have twice the range as ASCII, enough room for 256 characters (0-255). For convenience, all ISO-8859 charsets contain the full range of ASCII in their "lower" 128 characters, e.g. the bytes 0-127 from any ISO-8859 charset map directly to the corresponding ASCII character: ISO-8859 is a superset of ASCII. The differences of each ISO charset are in the "upper" 128 characters, e.g. the bytes 128-255.

For example, ISO-8859-1, with an alphabet suitable for West-European languages, has the umlauts Ä, Ö and Ü at the positions 196, 214 and 220.

In comparison, ISO-8859-7, with the Greek alphabet, has the Greek letters Δ, Φ and ά at the same positions.

In addition to the ISO-8859 charsets, there are of course a multitude of other charsets, including the "Unicode" charset (which aims to include all characters from all languages) and for example charsets for the east Asian languages, like Chinese, Japanese and Korean.

See section "2 Defining the Default Mail Charset" for a list of all charsets that are supported by LISTSERV Maestro.

Obviously, the 8-bit range of 0-255 is not enough to accommodate all letters from even the European languages at once (therefore the need for the more than a dozen different members of the ISO-8859 family), not to speak of the other languages of the world, like Asian languages.

To avoid this problem, in recent times the **16-bit** charset **Unicode** with a range for 65536 characters has become more and more widespread. This charset contains more or less all letters and characters from all languages, as well as a good set of symbols and other helpful characters.

Again, for convenience, the first 128 characters of Unicode (0-127) are exactly the same as in the ASCII charset, while the first 256 characters (0-255) are the same as in ISO-8859-1 (West European). A quite large percentage of all the other letters of the languages of the world are assigned values from 256 to 65535. Maestro offers Unicode in form of its UTF-8 variant: UTF-8 is a transfer encoding for the 16-bit Unicode charset, which maps Unicode characters to one, two or more bytes, in a way that more common characters (like ASCII characters) need less bytes than uncommon characters.

## *1.2 LISTSERV Maestro and International Character Sets*

So what happens when you use international characters in a mail you write and send in LISTSERV Maestro?

Internally, LISTSERV Maestro uses pure Unicode, e.g. you can use and mix any characters you like in your mails (including the subject line and even in the merge data that you upload or select from a database) – as long as you have a way of inputting them (for some languages, this simply requires that you install a special keyboard and display driver for that language, for other languages, like Asian languages, you might even require a special keyboard – this depends on the language and on your operating system).

But for sending LISTSERV Maestro needs to decide on a charset that it can use to encode the message with. You either specify the charset to use while defining the content (there is a special option for this on the content definition page), or you tell LISTSERV Maestro that it should try to determine automatically which charset is the optimal one for the text you have written.

In the latter case, LISTSERV Maestro scans the text you have written to determine the optimal charset: If you have only used characters that can be displayed with the ASCII charset (as is the case with most English language texts), LISTSERV Maestro will choose the ASCII charset. If you have used characters outside of the ASCII range, but which can still be displayed with one of the supported ISO-8859 charsets, then LISTSERV Maestro will choose the corresponding ISO-8859 charset. Similarly, if you have used Chinese, Japanese or Korean characters which can be displayed with one of the supported Asian charsets, then LISTSERV Maestro will choose such a charset. And, optionally (only if you have told LISTSERV Maestro that using Unicode is OK), if you have used characters that cannot be displayed with one of the supported ISO-8859 or Asian charsets, or if you have mixed characters from several ISO-8859 charsets and/or from other languages, then LISTSERV Maestro will choose Unicode as the charset.

Once a charset is chosen, LISTSERV Maestro encodes each character as a bit sequence according to that charset. The mail that is sent is then augmented by the information of which charset was used to encode it.

This information is then used by the receiving mail client to decode the bit sequence into characters that can be displayed to the user.

For example, with ASCII charset (where each 7-bit sequence denotes one character), the sequence “1000001” would mean the character with the decimal value 65, which is the Latin ‘A’. With the ISO-8859-1 charset (where each 8-bit sequence denotes one character), the sequence “11000100” would mean the character with the decimal value 196, which is the umlaut ‘Ä’. But with the ISO-8859-7 charset (also 8-bit), the same value 196 would mean the Greek letter ‘Δ’ instead.

Thus, the information of how the bit sequence is to be decoded into displayable characters is very important, shall the mail be shown properly. LISTSERV Maestro takes care to include this information in the mail, so that it is not lost during the transfer.

## *1.3 Text-Merging with International Character Sets*

The issue of international character sets in combination with text merging needs to be considered very carefully, to make sure that the results of the merging appear at the recipient as desired.

The main problem when text merging in texts that use international charsets is, to decide which charset to use. Potentially, the characters in the body of the mail require a certain charset, while some of the merge values may require a different charset. For example, you may have an English text mail but a recipient list with recipients from all over the world, with names that contain letters from various languages. You need to consider what happens when you merge these names into the English body text.

The effect that the chosen charset has on the merge values depends on the kind of recipients definition selected for a certain job:

If recipients are **uploaded as a text file**, or are **selected from a database by the Maestro User Interface,** or come from a **subscriber dataset** or **subscriber list** in the subscriber warehouse, then all recipients and their merge values are known to the Maestro User Interface already before the job is submitted to LISTSERV for sending. LISTSERV Maestro can therefore encode each merge value with the same charset that is used for the mail text. Consequently, if the values are later merged into the text, their charset will match that of the text. However, if a merge value contains a character that cannot be displayed in the charset chosen for the text, then this character will be replaced with a question mark ”?” during the encoding, and this question mark will appear in the mail that reaches the recipient to which the merge value belongs.

This could be a problem for example in the sample described above: If the mail text itself is plain English, then using ASCII as the charset seems an obvious choice. However, if then the names of the international recipients are encoded as ASCII, all non-ASCII international characters will be replaced with question marks. You should therefore take care to use the same charset for the mail as was used for the merge data: If you uploaded the recipients information as a text file, then simply use the same charset for sending as you used during the initial upload. And if you selected the recipients information from a database, use the same charset as is used by the database (you might have to ask your database administrator for this information).

Summary: For the these two recipient types, merge value characters that have no representation in the charset that was chosen for the mail text will be displayed as ”?”.

If recipients are defined by sending to a **classic LISTSERV list**, or as a **classic LISTSERV list posting** to an advanced subscriber list, or by letting **LISTSERV select from a database**, then the Maestro User Interface will not see the actual recipients or their merge values, and can therefore not do any special charset encoding on them either.

Instead, LISTSERV will simply merge the bytes from the recipients source (e.g. from the LISTSERV list or from the database LISTSERV connects to) into the mail text.

Consequently, you need to make sure that the merge values in the original recipients source (LISTSERV list or LISTSERV DBMS) already have the correct charset for the mail they are merged into. For example, if you are sending a mail with ISO-8859-1 (West-European), then all appearances of the byte 196 in the merge values will be interpreted as the umlaut ‘Ä’. Even if the merge value is for example actually a Greek word where the byte 196 should have been interpreted as a ‘Δ’.

While mixing characters from different ISO-8859 charsets will simply display the wrong character to the recipient, mixing ASCII and ISO-8859 or ISO-8859 and Unicode may even result in characters that cannot be displayed at all. Most importantly, if the mail text uses the Unicode encoding UTF-8, then you must make sure that the merge value texts in your recipients source are also UTF-8 encoded

(e.g. the byte sequence that stands for each merge value must be a valid UTF-8 encoded sequence representing a string of characters from the Unicode charset).

Of course, it is usually not possible to define a charset for the mail and then in some way make sure that the merge values in the list or in the LISTSERV database match this charset, since those merge values have usually been stored long before the mail was created. Therefore, the best way to proceed is to check which encoding was used when the data was stored in the list or LISTSERV database (again, you might need to ask your administrator for that information) and then use the same charset for the mail.

Summary: For these two recipient types, merge value characters that have no representation in the charset that was chosen for the mail text will be displayed as a different character, as whichever character from the actual charset that has the same byte value (like 'Ä' from ISO-8859-1 and 'Δ' from ISO-8859-7), or may not be displayed at all, if there is no corresponding byte value in the charset.

## *1.4 International Character Set Recipient Names in the Mail-TO-Header*

The previous subchapter has explained the problems of mixing a mail text in one language with merge values from a different language.

As an example, an English text mail was described, with an international recipient list where the recipient names contain characters from many languages, with the languages possibly differing between recipients from different countries.

The recipient's name as a merge value is probably one of the most common uses for text merging, to be able to merge the name into the text of the mail, to personalize the mail. If this is done, the problems as described above need to be considered.

However, the recipient's name is also often used in the "To:"-header field of the mail, so that the mail appears at the recipient with the recipient's own name visible in the "To:" field (which is usually displayed by the e-mail client in some nice fashion), personalizing the mail one step further.

When using recipients **uploaded as text files** or **selected from the database by the Maestro User Interface,** or from a **subscriber dataset** or **subscriber list** in the subscriber warehouse, then this usage of the name in the "To:"-header field does **not** fall under the constraints regarding charsets and text-merging as described earlier:

The name in the "To:"-header field will **always** be encoded with the charset that is **optimal** for exactly this name. E.g. you may safely write a mail in English and send it to your international recipients. Each recipient will see his or her name with the correct characters in the "To:"-header, e.g. the German recipient will correctly see her umlauts, the Russian will see his name in Cyrillic and the Greek will see his name with Greek letters (under the condition of course, that the original recipient list was in Unicode format and contained the names of the recipients with their respective international characters). Just remember that with such a mixed-language list of recipients merge values you should not also merge the name into the text body itself, unless the text is encoded as Unicode (UTF-8) too (because of the problems described in the previous subchapter).

However, when using recipients that are defined by sending to a **classic LISTSERV list**, or as a **classic LISTSERV list posting** to an advanced subscriber list, or by letting **LISTSERV select from a database**, then again the bytes from the name-merge value will be merged into the "To:"-header right by LISTSERV, without the Maestro User Interface having a chance to encode them. And since it is very improbable that the names (e.g. the byte sequences representing them) already contain the special MIME-header encoding necessary for non-ASCII "To:"-header fields, you should make sure that only ASCII characters are allowed in recipient names when creating your list or database data for these recipient types.

### 1.5 LISTSERV Maestro and Bi-Directional Character Sets

Of the ISO-8859 charset family, there are two charsets which contain letters from languages that have a standard reading direction of right-to-left. This are the charsets ISO-8859-6 (Arabic) and ISO-8859-8 (Hebrew), both of which are supported by LISTSERV Maestro.

Actually, Maestro will not use the charsets with the names ISO-8859-6 and ISO-8859-8, but will instead use the special bi-directional versions **ISO-8859-6-i** and **ISO-8859-8-i**. These charsets contain the same characters as their non-i-suffix counterparts, but the ”-i” suffix tells the receiving mail client that the text should be displayed with right-to-left reading direction.

Without this ”-i” suffix in the charset name, many e-mail clients would probably display the correct characters, but in the (for that language) incorrect left-to-right reading direction.

However, even with the ”-i” suffix, the recipient might need to install a special version of his mail client (or even a special mail client) that is prepared to display text with right-to-left reading direction properly and is also able to properly display bi-directional text (e.g. text that mixes characters with left-to-right and characters with right-to-left reading direction, as is for example the case in a Hebrew text that contains English names). Some clients may only display the characters with the right direction, but still left-align each line of text, instead of the correct right-alignment. This is however up to the mail client itself, and is therefore out of the scope of LISTSERV Maestro.

# 2 Defining the Default Mail Charset

Each mail job that is created in LISTSERV Maestro has a charset associated with its content. This charset is used to encode the content for sending.

When a job is first created as a new job (e.g. not as a copy of an existing job) then this job is initially created with the default charset. Which charset is chosen for this default can be defined with a setting in the Maestro User Interface INI-file.

To do so, you need to edit the following file:

```
[maestro_install_folder]/lui/lui.ini
```

Edit or add the key “`DefaultMailCharset`” and set it to the name of one of the charsets supported by LISTSERV Maestro:

| **Charset Name:** | **Description:** |
|---|---|
| `US-ASCII` | US ASCII |
| `ISO-8859-1` | West European, Latin 1 |
| `ISO-8859-2` | East European, Latin 2 |
| `ISO-8859-3` | South European, Latin 3 |
| `ISO-8859-4` | North European, Latin 4 |
| `ISO-8859-5` | Cyrillic |
| `ISO-8859-6` | Arabic |
| `ISO-8859-7` | Greek |
| `ISO-8859-8` | Hebrew |
| `ISO-8859-9` | Turkish, Latin 5 |
| `ISO-8859-15` | Similar as ISO-8859-1 but with Euro currency symbol |
| `GB2312` | Simplified Chinese (GB2312) |

| | |
|---|---|
| BIG5 | Traditional Chinese (BIG5) |
| ISO-2022-JP | Japanese (ISO-2022-JP) |
| X-EUC-JP | Japanese (X-EUC-JP) |
| X-SJIS | Japanese (X-SJIS) |
| KSC_5601 | Korean (KSC_5601) |
| EUC-KR | Korean (EUC-KR) |
| UTF-8 | International Unicode, encoded in UTF-8 format |
| AUTO-NO-UTF-8 | LISTSERV Maestro will choose either US-ASCII or any of the ISO-8859 charsets (but **not** UTF-8), depending on which characters are actually used in the content. If possible, ASCII is favored over any ISO-8859, so an ISO-8859 set is only chosen if ASCII is not able to display all characters in the content. Of the ISO-8859 sets, the one where the number of non-displayable characters is minimized is chosen. If two sets have an equal number of non-displayable characters, then lower ISO-8859 sets are favored over higher sets (e.g. ISO-8859-1 over ISO-8859-2, over ISO-8859-3, etc.). |
| AUTO-YES-UTF-8 | LISTSERV Maestro will choose either US-ASCII or any of the ISO-8859 or even UTF-8, depending on which characters are actually used in the content. If possible, ASCII is favored over any ISO-8859 and the ISO-8859 sets are favored over UTF-8. The step to the next "higher" set is only made if the "lower" set is not able to display all characters in the content. If several ISO-8859 sets are able to display all characters, then lower ISO-8859 sets are favored over higher sets (e.g. ISO-8859-1 over ISO-8859-2, over ISO-8859-3, etc.). |

The default charset is only initially assigned to the mail job, e.g. it may be changed by the user on the content definition page.

If the administrator wants to stop the users from changing the default charset (e.g. force the users to always accept the default charset that the administrator has chosen), he can do so by editing another entry in the same INI-file: Edit or add the key "`AllowCharsetChoice`". Set to "`true`" if you want to allow the users to change the charset of a job (e.g. be able to assign different charsets to each job) or to "`false`" if you want to disallow changing of the charset. (The default if the key is not present in the INI-file is "`true`").

# 3 Allowing or Disallowing Bi-Directional Character Sets

As described above, LISTSERV Maestro uses the special bi-directional charsets ISO-8859-6-i and ISO-8859-8-i when the Arabic or Hebrew charset ISO-8859-6 or ISO-8859-8 are used. There are a number of advantages to using these charsets. However, if the administrator would for some reason prefer LISTSERV Maestro **not to use** these bi-directional charsets, but use their "normal" counterparts ISO-8859-6 and ISO-8859-8 instead, he can do so with a special setting.

To do so, you need to edit the following file:

`[maestro_install_folder]/lui/lui.ini`

Edit or add the key "`AllowISO-i-Mails=false`" to **disallow** the bi-directional charsets. (If you remove the key from the INI-file, or comment it out or set it to "…`=true`", then the bi-directional charsets will be allowed as is the default).

This INI-file setting will affect **all** mails sent, with any user account. Please note, that changing this setting requires a restart of the Maestro User Interface component to take effect.