Multi-lingual Label Printing with Unicode

White Paper

Table of Contents

1 Introduction

Businesses moving to new regions in the world are often confronted with the problem of multi-lingual label printing. Many times labels must include more than just one language and combine printed information in English or Western European languages with Eastern European, Asian or Arabic languages.

Printing international characters to thermal printers has always been a challenge for the user. Different regions in the world use different languages that are encoded in different codepages, which are tables that have all the characters available for the selected language. If you must use characters that are available in different codepages and print them on the same label, you have a problem with non-Unicode software applications. Non-Unicode software applications do not allow you to design labels using characters from more than one codepage.

The only universal solution to print characters retrieved from different codepages is to use Unicode standard. Unicode makes it possible to use standardized and straightforward solutions for any combination of input languages on the label. Using Unicode enables you to create multi-lingual labels while reducing implementation and maintenance costs.

The main advantages of Unicode for multi-lingual label printing are:

- **Straight-forward label design**: no need for any code page selection.

- **Portability:** Unicode is independent of any application or platform.

- **Standardization**: all major business software, operating systems and hardware providers are adopting Unicode.

- **High performance printing:** thermal printers support Unicode.

2 Unicode

2.1 Introduction to Unicode

Character encoding and codepages

All characters that are printed on labels are defined in so-called codepages. A codepage is a table of predefined characters that maps each character with a unique numeric code. Computers can only work with this numeric code. Every character (letter, digit, special character) is represented by a number. The character "A" is always associated with number 65 and character "a" with number 97. The number represents the position of the character in the table.

The first computers could only work with 128 different combinations. Then the American National Standard Committee (ANSI) developed American Standard Code for Information Interchange (ASCII), which was a standardized encoding table (a codepage) covering all upper case and lower case English letters, digits, punctuation characters, as well as some special and control characters.

However, developers soon noticed that 128 codes are not sufficient for all characters. The diacritical marks of the Western European languages could not be covered. Standard committees, such as ANSI and International Standards Organization (ISO), and computer companies like IBM, Apple, and Microsoft started extending the ASCII codepage with various character sets. The additional 128 codes had been filled with different characters such as graphical symbols and mathematical signs.

Today most computer applications and operating systems can easily access 256 characters at a time, because they originally use 8 bits (one byte) that make 256 different combinations (2^8=256).

But each of the organizations enforced their own "standard" for the characters and their position in the table of additional 128 characters. ANSI and Microsoft invented the codepage 1252 (ANSI Latin-1), the ISO established the ISO-8859-1 (ISO Latin-1), IBM developed the codepage 850 (IBM Latin-1), and Apple created the Macintosh Roman character set. The standards can have different positions of the characters so that the same numerical code points to different characters. However, the basic 128 characters were the same for all codepages.

In the early 1990s, the political changes in Central and Eastern Europe (Poland, former Czechoslovakia, Hungary, former Yugoslavia), Baltic countries (Lithuania, Latvia, Estonia), and countries of the former Soviet Union enforced the definition of new codepages. Different manufacturers supplied different solutions. For example, Polish, Czech, Slovak, Hungarian, Albanian, Romanian, Croatian and Slovenian languages have been artificially grouped into Central European codepages (also called Eastern European or Latin-2). ANSI developed CP 1250 (ANSI Latin-2), ISO had the ISO 8859-2 (ISO Latin-2), IBM invented CP 852 (IBM Latin-2), and Apple created Macintosh CE codepages.

In addition, alphabets of Asian languages can have up to thousands of letters that would never fit into the 8 bits that were used to encode the 256 characters in the codepage. This was solved with a system called the "double byte character set" (DBCS) that allows mixing of single-byte and double-byte encoded characters. Two bytes provide enough space for addressing the numerous glyphs in the Asian languages. There are enough combinations of numbers to assign each glyph with unique identification.

Common problems with codepages

The solutions with different codepage standards were acceptable if you never used a language from more than one language family in your documents. However, with the rise of the internet there was suddenly a need for true multi-lingual support for documents or

labels. Labels created on one system must be opened anywhere in the world and retain the font formatting and data. Normally, you can use your data written in one character set (codepage) on your labels but you won't be able to read that data with a computer that does not use the same codepage. Different regions in the world use different codepages and different operating systems, applications and databases do not look at the data in the same way.

Following the model of individual codepages that define characters of one specific language is no longer a solution because codepages cannot be used simultaneously on the same operating system. Characters defined from some other codepage are usually converted to question marks.

Unicode as the universal character encoding solution

The Unicode standard is a character encoding system designed to support written texts of languages with different characters. Unicode was developed to solve the challenge of international character support. It is based on double-byte character encoding and not the one byte per character relationship that exists today. The total number of characters defined by current Unicode Standard 4.0 is over 70,000 with the number growing constantly.

Many languages are not written in the Latin alphabet. In fact many languages are not written with alphabets at all. Many languages do not write from left-to-right and top-to-bottom, do not have spaces between words, and do not have an alphabetical order.

Unicode promises to become the "one and only" standard and may solve the issues that come with codepages. Using Unicode, you no longer have to think about different code pages and the location of characters inside the font file. The same Unicode font file provides you with all international characters you will ever need in any written language in Africa, North and South America, Asia, Australia, and Europe.

Each character in Unicode definition gets a name and a number, for example the Latin capital letter A is 65. Unicode includes a table of useful character properties such as "this is lower case" or "this is a number" or "this is a punctuation mark." Unicode numbers are given in four hex digits preceded by U+ so "A" is U+0041.

Unicode stores text on a computer as a series of numbers, using one number per character. There are many different ways how to save these numbers (which are called "encodings") in the computer memory. Unicode itself defines several different encoding schemes, such as UTF-8, UTF-16 and UTF-32.

Unicode is available to users in

- Latest operating systems (Windows 2000, Windows XP, Windows 2003 Server, AIX, Solaris, HP/UX, AS/400, or Mac OS.)

- Many high-end applications (SAP, XML, or AS/500 .)

- Many popular databases (Oracle, MS SQL Server, or MS Access.)

Font encodings

TrueType fonts provided with Windows 95 supported a subset of Unicode that included characters required by Western, Central, and Eastern European languages including characters required by Greek and Turkish. The fonts were available with just over 600 different glyphs, which was enough to cover these languages. To make the use of multi-lingual fonts in non-Unicode applications easier, Microsoft introduced a mechanism that allowed users to choose a font and change codepages as needed (see figure below). There were multiple instances of each multi-lingual font available in the Windows system. For example, the font "Times New Roman" was available as "Times New Roman Baltic" or "Times New Roman Central European." The first 128 characters in the codepage were the same but the upper 128 characters (positions from 128 to 255) differ accordingly to

the selected variant of the font. Alternatively, the font selection dialog box in Microsoft applications provided users with an option to select the desired script in a combo box.

Selection of the scripts and individual codepages did offer a subset of true multi-lingual label printing but did not provide all benefits of Unicode.
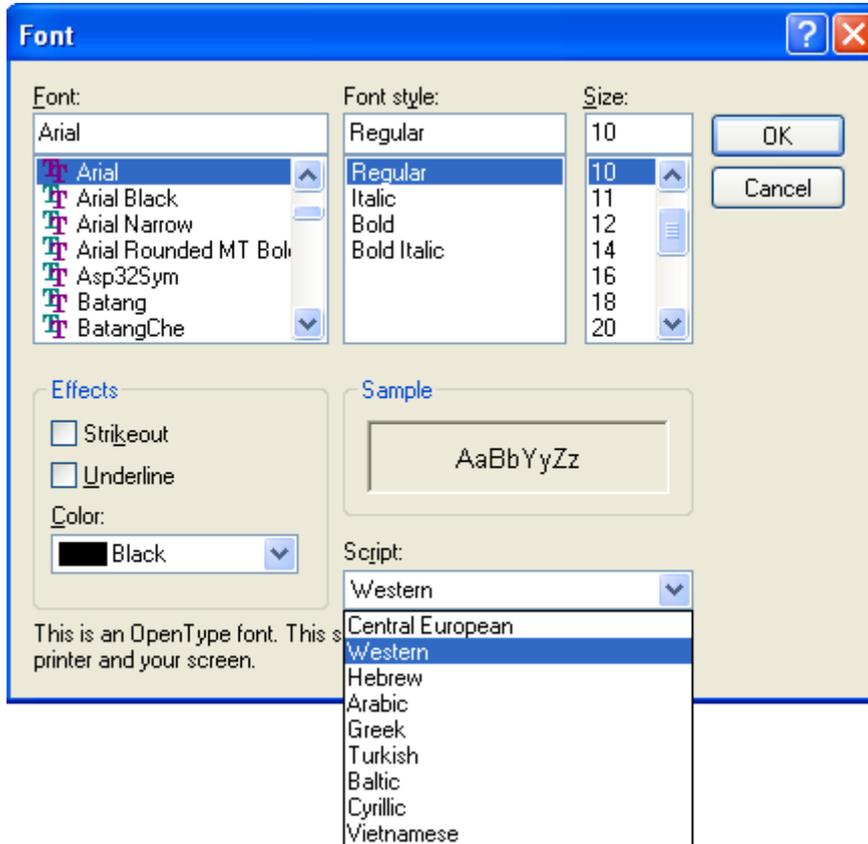
Cumbersome script (codepage) selection prior to use of Unicode

Limitations of codepages when used for multi-lingual labels

- When using multi-lingual text, you need to know which codepage to use.

- Additional font installed for each codepage.

- Different encoding standards for the characters in the codepage (Windows, ANSI, ISO, or other company standards)

- Impossible to use international characters from different codepages within one field on the label (text object, field in the database). For example, you cannot use Hebrew and Greek on one label at the same time.

- Non-Unicode applications can use characters from only one language – the one selected on your operating system (under regional settings in the control panel). Copy/Paste operations work only for characters that are available in the codepage for the selected language. Other characters are not supported and in most cases are displayed as question marks.

- Custom solutions are possible only for individual languages or geographic areas. No universal solution available for the global market.

Unicode fonts

In 1995, Microsoft Windows 95 was introduced. Fonts that shipped with Windows 95 included a subset of Unicode. Later, Microsoft and other vendors released fonts that included a smaller or larger character set. For example, Arial Unicode MS covers most of the Unicode standard with over 51,000 glyphs.

Recent versions of Windows operating systems offer a simplified language selection. The TrueType font is no longer represented in different language variants and users see only one font. The selection of the language is done by switching the keyboard layout to the desired language. For example, if you select American keyboard, the American codepage in the Unicode font is automatically selected and used; if you select Hebrew keyboard layout, the Hebrew codepage in the font is selected.

Selection of the keyboard layout defines the characters from the Unicode font

To take advantage of the new functionality, applications using the TrueType font must be Unicode-compliant. Unicode applications use double-byte text encoding and are able to address the extended characters directly.

Benefits using Unicode

1. Universal support for all characters of any language in the world

Unicode is a worldwide accepted standard and is the only solution for multi-lingual label printing. Unicode is a »universal codepage« for font character mapping and provides a unique number for every international character. Unicode provides uninterrupted data-interchange between Unicode-aware computers and applications without any data corruption or lost characters.

The Unicode standard supports all international characters from different languages (Western, Central and East-European, Cyrillic, Hebrew, Greek, Arabic, Chinese, Thai, Japanese, etc.). To use the international characters, you must have the TrueType font that includes the characters. Not every Unicode TrueType font includes all characters. Windows operating system comes with some TrueType fonts that have a large number of multi-lingual characters. One example of such a font is Arial Unicode MS.

2. Select the characters easily without the need for any additional settings

True Unicode-aware applications provide simple multi-lingual design of any document. For example, the Label Gallery Plus application is a truly Unicode-aware application and allows you to use special characters from different languages on the same label (see figure below). A change of the keyboard layout in the Windows operating system is enough to enable characters in the selected language.

3. Mix languages in the same text object

You can not only use the text elements formatted in different languages on the same label, but you can actually mix the languages in the same text object on the label. You can also obtain the data formatted in different languages from the same field in the database table. Such functionality was not even imaginable using codepages alone because only one particular codepage could be used at a given time.
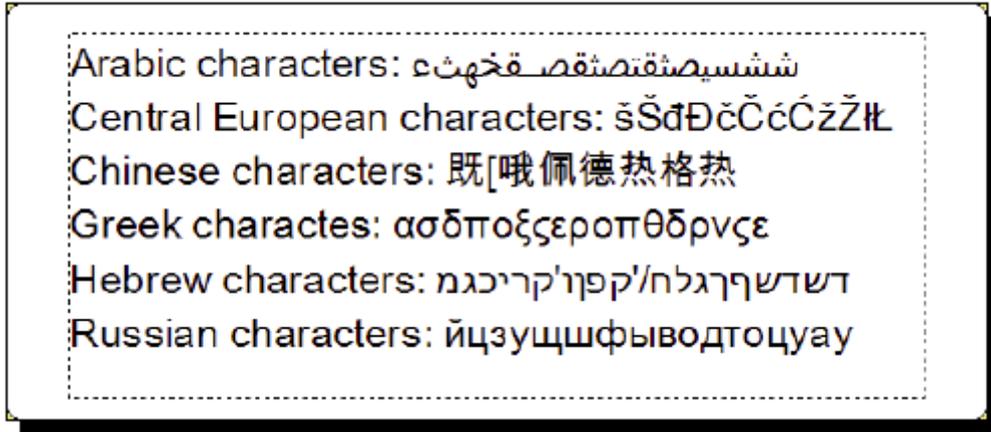
Label created in Label Gallery Plus software uses multi-lingual Unicode data in the same Paragraph element

4. Transfer multi-lingual data between applications

You can exchange data between two Unicode-aware applications with the simple Copy/Paste operation. Multi-lingual text data can be simply imported onto the label designed in Label Gallery Plus software regardless of the regional settings properties on the Windows operating system. You can have defined one regional setting on your operating system but work with data and languages from a region using different language characters. The full support for Unicode standard in the Label Gallery Plus software will enable you to use different language characters simultaneously.

2.2 Unicode Support in Label Gallery Software and Gallery Drivers

Label Gallery Plus and Label Gallery Easy

Label Gallery Plus and Label Gallery Easy are a true Unicode bar code label design environments that provide the following advantages:

### 1. Unlimited use of multi-lingual fixed and variable text on a label
Label Gallery software offers many methods for using text on the label. There are three objects (Text, Paragraph and RTF) that can be used for character manipulation on the label and that are all compliant with Unicode formatted data. The objects can have non-changeable content or the content can change with each label. You can also copy the text formatted in Unicode in some other application and paste it to Label Gallery.

NOTE:
Label Gallery Easy does not have available all text objects as are in the Label Gallery Plus.

2. Mix languages within a single text object

If you want to write in more than one language, simply switch the keyboard layout and Label Gallery will respond to new characters.

![SATO DCS & Labelling Worldwide logo]
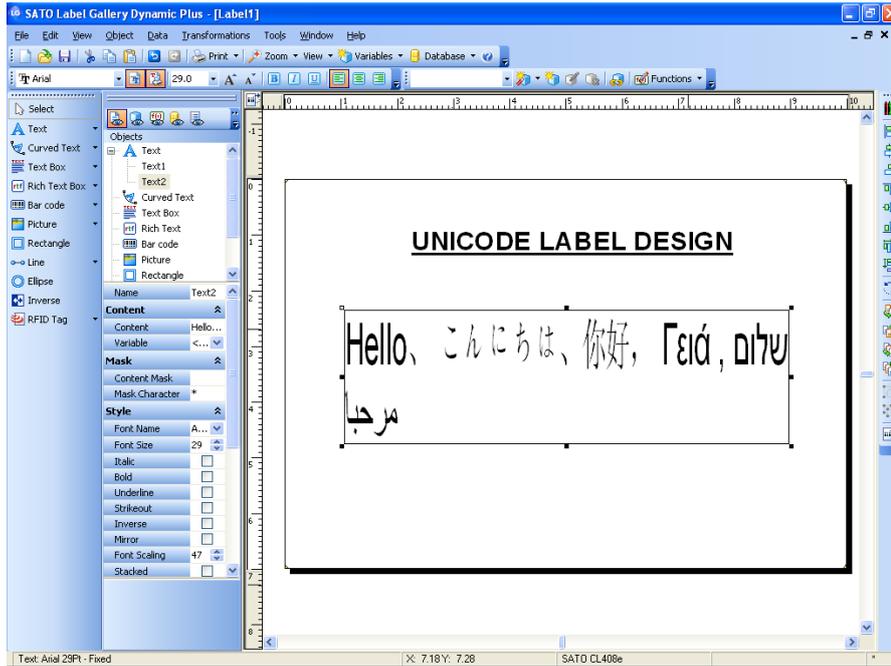
Different languages can be used in the same text element

3. Data retrieval from any Unicode database

Label Gallery can connect to and retrieve data from any kind of Unicode compliant database. The database can be file-based like MS Access or server-based like MS SQL Server. The input data is automatically recognized as Unicode formatted data. The detection is automatic and the user does not have to make any manual modifications. The advanced user can still set the UTF encoding manually. The next figure shows data from three consecutive records from the Unicode database correctly displayed on the labels.
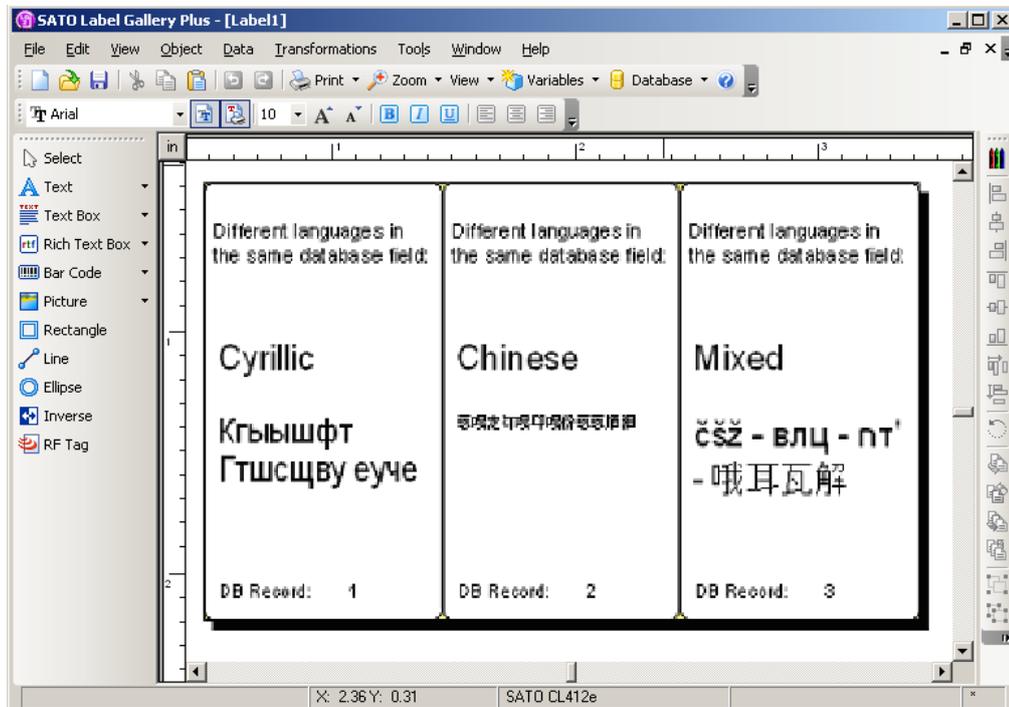
Text in different languages is retrieved from the same database field

4. Optimal performance during label production
Label Gallery and GalleryDrivers are developed to provide the fastest printing label solution. Label Gallery software automatically uses recognized resident printer functionality. Printing optimization ensures that only the necessary data is sent to the printer. There is no data overhead. If the thermal printer has built-in Unicode support, you can use Unicode formatted data with the resident font.

GalleryDrivers

Thermal printers are beginning to support Unicode. Thermal printer manufacturers have noticed the evolving demand for native Unicode label printing and are adding printer resident Unicode fonts to the thermal printers. Unicode fonts are usually available as a built-in in the printer firmware or in an external PC cards.

GalleryDrivers follow the development of Unicode-aware thermal printers and enable users to print Unicode characters from resident printer fonts. GalleryDrivers increase printing speed. Characters are no longer imaged and sent to the printer as graphic but are recalled from a resident font in the printer.

Using Unicode-aware thermal printers with Label Gallery and GalleryDrivers provides you with the following advantages:

1. Built-in Unicode printer resident font
If the thermal printer has built a Unicode font in an internal RAM or external PC card, Unicode can be used in Label Gallery software directly. There is no need to purchase the fonts separately.

2. No need to sacrifice performance to achieve ease-of-use
The Unicode fonts are built in the thermal printers like all other resident printer fonts. When the printer processes the label, the Unicode font does not have any extra impact on the printer's speed and performance.

3. No need for custom programming to achieve maximum performance
Forget about modifying the printer stream data manually. You will never have to tweak the printer command to achieve maximum printing performance. The GalleryDriver will generate and send the optimal printer stream to the printer.

4. Fast first label out
Unicode font is stored directly in the printer's RAM and accessible on demand instantly. When Unicode formatted data is required on the label, the necessary characters from the selected language are recalled from the printer. There is no need to send characters from the PC computer as a series of bitmap images. The optimized print stream contains less data to describe the same label. Less data can be sent to the printer quickly, which results in fast label printing.

5. No pause between label batch printing
Printer drivers usually send each label as a separate print job to the printer. Because each label is sent separately, the printer is constantly receiving data. For example, TrueType Unicode fonts are transferred into images during the label printing process and the transmission can take a lot of time especially for complex labels. The manufacturers are beginning to develop printers with built-in Unicode-aware printer fonts that will minimize the amount of data sent to the printer.

3 Conclusion

3.1 Who Needs Unicode?

The use of Unicode is inevitable for all needs of multi-lingual label design. Whenever there is a need to use characters formatted in different languages on the same label or even the need to mix the languages in the same text object with data coming from a database, there is no more elegant and easy solution than Unicode support. Unicode provides a global solution and solves all previous problems with proprietary codepages that worked only for certain language areas.

Multi-national companies that operate across borders and in different language areas recognize the main advantage of Unicode. Unicode enables companies that sell products in foreign markets to provide identification on the product written in the languages used in the selected market in an easy and cost-efficient manner.

3.2 The Answer for Multi-lingual Label Needs

The flexibility to print new languages without adding fonts and redeveloping label formats provides a significant total cost of ownership advantage for systems that will be deployed to support multiple languages. Unicode support is a great advantage to companies that operate globally. The complexity of solutions and fear of additional development costs used to discourage modifications to labeling systems. Using the Unicode support, companies can quickly respond to the demands on new markets without building up costs.

Native Unicode support in thermal printers and Label Gallery software guarantee that your printing solutions are applicable worldwide. There is no need for label modifications or even to purchase extra equipment or software. Once you have created your label printing solution, any user anywhere in the world can apply it.

As a result of using Label Gallery software that supports Unicode in your label printing system you save time, resources and money. To print Unicode data on multi-lingual labels, you need only to install Label Gallery software and the GalleryDriver for your thermal printer.

4 Glossary

Unicode	Unicode is a standard that defines how characters from different languages are encoded in the font files and describes how the applications can find and retrieve the required characters. Individual Unicode fonts can contain characters from all languages (Western, Central and East-European, Cyrillic, Hebrew, Greek, Arabic, Chinese, Thai, or Japanese.) or only characters from a subset of these languages.
	Unicode is accepted worldwide in all modern operating systems, applications and databases and is the only method for reliable multi-lingual data interchange between computers and applications.
Codepage	A codepage is a table of predefined characters that maps a character with the unique numeric code. Every character (letter, digit, special character) is represented by a number. The originally developed codepages were not really designed for international use, so several incompatible variants of codepages emerged.
Double Byte Character Set (DBCS)	Normally, Latin characters require 8 bits (one byte) to encode the frequently used characters in the codepage. One byte data makes is possible to encode 256 characters. The need for encoding characters for non-Latin based language emerged quickly. One byte was not sufficient to encode characters in languages like Chinese, Japanese, Thai, Arabic and Hebrew.
Font	A collection of glyphs (image representation of characters) used for the visual depiction of character data. A font is often associated with a set of parameters (for example, size, posture, weight, and serifness), which, when set to particular values, generate a collection of imagable glyphs.
Glyph	The distinct visual representation of a character in a form that a screen or printer can display. A glyph may represent one character (the lowercase a), more than one character (the fi ligature), part of a character (the dot over an i), or a nonprinting character (the space character).
Unicode Transformation Format (UTF)	Algorithmic mapping from any Unicode character to a unique byte sequence.
Unicode Transformation Format (UTF-8)	UTF-8 is an encoding scheme that maps all possible double-byte printer encodings to a series of single-byte and multi-byte strings. UTF-8 is compatible with all legacy file systems and other systems that parse for the ASCII byte.
Unicode Transformation Format (UTF-16)	UTF-16 encoding scheme represents each Unicode character as a sequence of two bytes.
Unicode Transformation Format (UTF-32)	UTF-32 encoding scheme represents each Unicode character as a sequence of four bytes.

5 Appendix

Label Gallery Product Overview

Label Gallery is the most advanced professional labeling software for desktop and enterprise users. Label Gallery offers an easy-to-use interface and meets any label design and printing requirement for efficient label printing solutions to users in retail, logistics, health care, chemical, automotive and other industries.

Label Gallery TruePro: Complete software solution for any kind of label design and print requirement. Multiple connectivity options allow users to perform stand-alone printing or integrate label printing into any network environment.

Label Gallery TruePro provides you with interactive label printing capabilities such as integrating label printing to existing applications (ActiveX) or non-programming embedding of label printing to existing systems (GalleryWatch).

Label Gallery Plus: Full-featured software designed for professional label design and printing, including complete database support and ActiveX integration possibilities. A wide range of features and options makes Label Gallery Plus a perfect and easy-to-use tool for any labeling requirement.

Label Gallery Easy: Wizard-based software meeting basic barcode labeling needs. The entry-level software includes many design elements of the Easy edition with the emphasis on simplified user interaction.

Label Gallery Plus Print Only: Label Gallery Plus Print Only offers printing of pre-designed labels but cannot be used to design and alter existing labels. Advanced settings for changing the labels are not available.

Label Gallery TruePro Print Only: Label Gallery TruePro Print Only offers printing of pre-designed labels, using pre-designed forms and automatic printing from pre-designed trigger actions. Label Gallery TruePro Print Only cannot be used to design and alter existing labels, forms and trigger configuration. Advanced settings for changing the labels are not available.

Pocket Label Gallery: Pocket Label Gallery is a program package for Windows CE that brings the power of label printing to portable Windows CE computers (Windows Mobile Device). Pocket Label Gallery is part of the editions Label Gallery TruePro or Label Gallery Pocket PC Designer.

Label Gallery Developer Series:

- **Label Gallery SDK** : Label Gallery SDK is an ActiveX integrator edition of Label Gallery software developed for software publishers who needs label printing capabilities in their software.

- **Pocket Label Gallery SDK**: Pocket Label Gallery SDK is an ActiveX integrator edition of Label Gallery software developed for software publishers who needs label printing capabilities in their software under mobile devices.