Unicode

Best Practices

Version: Foundation 24.1

Written by: Documentation Team, R&D
Date: March 2024

Hyland

Documentation Notice

Table of Contents

Unicode implementation for Perceptive Content

The Perceptive Content Foundation EP2 (7.5.x) release for Perceptive Content Server, Perceptive Content Client, Interact Desktop, and several Perceptive Content agents enables storage, processing and display of Unicode characters falling along the Basic Multilingual Plane (BMP), as well as supplemental characters, throughout the core components of database, server, clients, and several core agents. For information on converting your existing Percpetive Content database to Unicode using ImageNow iConvert, see *ImageNow iConvert Product Guide Foundation EP4*.

About Unicode character sets

Unicode is a standardized Universal Character Set (UCS) that includes almost all of the text characters in use throughout the world. It enables computer programs to display text and handle the input and output of data in a user's local language regardless of the language. Operating systems also utilize Unicode as a method of processing and storing text in different languages.

There are several methods for implementing a Unicode character set, but two of the most common methods are with UTF-8 (UCS Transformation Format 8-bit) and UTF-16 (UCS Transformation Format 16-bit) character encoding.

About UTF-8 and UTF-16 encoding

Content utilizes UTF-8 and UTF-16 character encodings, which maintains system efficiency and ASCII compatibility. These encoding methods also enable the use of non-English languages without having to change an operating system's base language, convert code during execution, or exponentially increase the file size of programs in order to support multiple languages.

UTF-8 encoding

With UTF-8 encoding, a unique sequence of one to four bytes, where one byte represents ASCII characters, and up to four bytes for non-ASCII characters represents each character in a written language. UTF-8 is primarily for file input and output and for inter-process communication between the client and server.

UTF-16 encoding

UTF-16 utilizes one or two 16-bit elements to encode each of the Unicode code points. Other characters are comprised of four 8-bit bytes. At the database level, Perceptive Content uses UTF-16 encoding for internal data but uses UTF-8 encoding for file data.

Perceptive Content Unicode functionality

There are several things to keep in mind when using the Unicode implementation of Perceptive Content, including supported platforms and languages, the type of text editor to use when viewing configuration files, the structure of some files, and so on. The following sections describe Perceptive Content and Unicode functionality.

About Supported Perceptive Content Server and database configuration

The only supported Perceptive Content Server and database configuration for Unicode support is a SQL Server database using a Windows x64 Server. Perceptive Content does not support Unicode for an Oracle database in a Windows or UNIX environment. You can only implement Unicode for Perceptive Content with a new installation of Perceptive Content. For more information about installing Percpetive Content with Unicode support, contact your Perceptive Software representative.

For instructions on installing and configuring Perceptive Content, see the *Perceptive Content Server Installation Guide Foundation EP2, Perceptive Content Client Installation Guide Foundation EP2 and the Perceptive Content Database Installation Guide Foundation EP2*.

Core components

All Perceptive Content core components are Unicode compatible. The supported core components include:

- Forms Server
- Business Insight (BI) Suite
- Perceptive Content Client (32-bit only, with the capability of running on Windows 7 64-bit)
- Integration Server
- Interact Desktop
- Interact for Microsoft Outlook (32-bit only, with the capability of running on Windows 7 64-bit)

Perceptive Content agents

Perceptive Content includes functionality for most Perceptive Content embedded and external agents. The following sections describe agent functionality in Perceptive Content.

External agents

The following Perceptive Content external agents are Unicode compliant:

- Mail Agent
- Output Agent
- User Replication Agent

Embedded agents

All Perceptive Content embedded agents installed with the Perceptive Content Server installation are Unicode compliant with the exception of Monitor agent, which has limited functionality. The agents are 32 and 64-bit compatible. The supported agents include:

- Alarm
- Batch
- External Message
- File System

- Import
- Job
- Monitor
  - ArchiveLogs
  - AbnormalTermination
  - RestartProcess
  - TimeOfDay

    Note The [Ignore] and [Email] setting groups in the inserverMonitor.ini file have been nullified.
- Notification
- OSM
- Retention
- Task
- Workflow

About Byte order marks

Some Perceptive Content files, including INI files, are UTF-8 encoded and contain a byte order mark (BOM). A BOM indicates the byte ordering and is only found in UTF-8 files on Windows.

When viewing log files using a text editor, ensure that the text editor supports UTF-8 encoding. Otherwise, the information in the log files may not display correctly. However, if the data in the log is all ASCII, then it will display correctly even without a UTF-8 enabled editor.

Important If you change settings in any Perceptive Content files or if you write external files, ensure that you are using a text editor that supports UTF-8 encoding, and that you do not alter the BOM if it is present in the file.

Integration server configuration

The following procedures enable UTF-8 encoding for Integration Server. Perform the procedure specific to your environment prior to installing Integration Server.

Enable UTF-8 encoding for Integration Server usingTomcat

To enable UTF-8 encoding for Integration Server using Tomcat, complete the following steps.

1. In the server location where Tomcat is installed, navigate to the **conf** subdirectory, and then open the **server.xml** file with a text editor.
2. In the **server.xml** file, in the **[Server]** group, under the **[Service]** element, locate the **[Connector]** element that contains the HTTP/1.1 protocol.
3. Add the following parameter to the **[Connector]** element:

   ```
   URI Encoding=”UTF-8”
   ```
4. Save the changes, and then close the text editor.

Enable UTF-8 encoding for Integration Server using WebSphere

To enable UTF-8 encoding for Integration Server using WebSphere, complete the following steps.

5. In the IBM Integrated Solutions Console, on the left pane, click **Servers** > **Server Types** > **WebSphere application servers**.
6. In the **Application servers** pane, click the desired server link.
7. On the **Configuration** tab, under **Server Infrastructure**, click **Java and Process Management**>**Process definition**.
8. In the new screen, under **Additional Properties**, click **Java Virtual Machine**.
9. Under **Generic JVM arguments**, enter the following text:

```
-Dclient.encoding.override=UTF-8
```

10. Click **OK**, and then save your changes.

Enable UTF-8 encoding for Integration Server using WebLogic

To enable UTF-8 encoding for Integration Server using WebLogic, complete the following steps.

11. In the directory where Integration Server is installed, navigate to the **WEB-INF** subdirectory, and then open the **weblogic.xml** file with a text editor.
12. In the weblogic.xml file, locate the **[charset-params]** element. Insert the following elements in the **[charset-params]** element:

```
<charset-params>
<input-charset>
       <resource-path>/*</resource-path>
       <java-charset-name>UTF-8</java-charset-name>
</input-charset>
</charset-params>
```

13. Save the changes, and then close the text editor.

Note If you enable the native-email-support setting in the [Application server settings] group, users cannot email Unicode data from Integration Server.

Settings for iScript, agents, and XML files

For iScript, Import agent, and Output agent, you must encode text-based data files, index files, associated text files, and combo files in the ANSI file format if you use Perceptive Content on an ANSI Server, and in UTF-8 file format if you use a Unicode Server. If you use Perceptive Content on a Unicode Server, but run legacy ANSI files, you can use the following settings in the inow.ini file to direct Perceptive Content Server to treat these files as ANSI or UTF-8 files:

- **ANSI**. This allows ANSI files to run after upgrading Perceptive Content Server to Content Foundation EP2 (7.5.x) with Unicode compatibility.
- **UTF-8.** This setting assumes all files are UTF-8 compatible.

The following sections contain the specific settings for iScript, Import agent, and Output agents. The default value is UTF-8 for Unicode servers and ANSI for ANSI build servers.

iScript

For iScript, you must configure the [iscript.encoding] INI setting in the inow.ini file:

```
iscript.encoding=ANSI | UTF-8
```

For example, if iScript files contain UTF-8 encoding, then the INI [iscript] setting must be:

```
iscript.encoding=UTF-8
```

Import agent

For Import agent files, you must configure the [file.encoding] INI setting in the inow.ini file:

```
file.encoding=ANSI | UTF-8
```

For example, if Import agent files contain UTF-8 encoding, then the INI [file] setting must be:

```
file.encoding=UTF-8
```

Output agent

For Output agent files, you must configure the [keyfile.encoding] INI setting in the inow.ini file:

```
keyfile.encoding=ANSI | UTF-8
```

For example, if a key file is contains UTF-8 encoding, then the INI [keyfile.encoding] setting must be:

```
keyfile.encoding= UTF-8
```

Optional parameters for XML files

When creating iScripts that modify data that contains Unicode characters, you can include parameters that enable XML files to be read or written in ANSI or UTF-8 formats. The optional parameters are:

- For UTF-8 encoding:
  - XML.toFile(xml, 1, xmlFile,"UTF-8");
  - XML.fromFile(xmlFile,1,"UTF-8")
- For ANSI encoding:
  - XML.toFile(xml, 1, xmlFile,"UTF-8");
  - XML.fromFile(xmlFile,1,"UTF-8")

INTool commands

If you want to pipe a file into a standard input of a Unicode compatible INTool command, the file must be encoded as UTF-8 with a BOM.

There are some INTool commands that are not compatible with Unicode. If you are using any non-Unicode INTool commands, Perceptive Software recommends that you change the command console font to a True Type font.

Note All audit commands, db-export, and db-impor are no longer in INTool.

Changing the font to a True Type font

To change your console font to a True Type font, complete the following steps.

14. Click **Start** > **Run**, and then type **cmd**.
15. Click **OK**. The console window appears.

16. Left-click on the icon at the top left of the command prompt window, and then click **Properties**.
17. On the **Properties** dialog box, select the **Font** tab.
18. In the **Font** area, select a True Type font. True Type fonts have a black T in front of a gray T before the font name.

    Note If there are not any true type fonts displayed in the Font area, or you see font messages before the properties page loads, ensure you are on a valid code page. To verify that you are on a valid code page, run the **chcp** command.
19. Optional. Update the font size.
20. Click **OK**.

About Management Console

You can no longer modify user log levels, or retrieve user logs from Management Console. However, log file functionality is available using INTool.

Error messages for INTool reason commands

If you run an INTool reason command with an INI file that contains an incorrect encoding setting, any resulting error messages do not appear in a log file, but they do appear in a message that INTool displays at a command prompt. For example, in the following message, the line *An error occurred while reading the input file*, indicates that the INI file might be incorrectly encoded:

```
[drive:]\inserver6\bin64>intool --cmd add-digsig-reasons --file c:\inserver6\install_t
emp\[filename]
An error occurred while reading the input file:
[drive:]\inserver6\install_temp\[filename]
```

Note This message can also indicate that there is an error reading a setting from the INI file that is not a result of using the wrong type of encoding. If you receive this error, first ensure that the INI file is encoded as UTF-8, and if it is, check other settings in the file to determine if all of the appropriate settings are correct.

AP eForm

The following UTF-8 configurations apply to AP eForm.

Virtual tables

To import data containing double-byte characters into a virtual table, the associated CSV file must use UTF-8 encoding without a BOM.

Brainware Distiller for Invoices

If you require Brainware Distiller for Invoices to export double-byte characters, in the BW Invoices.ini file, configure the `[EXP_VL_XMLEncodingHeader]` element and set encoding to UTF-8.

For example:

```
EXP_VL_XMLEncodingHeader=<?xml version="1.0" encoding="UTF-8"?>
```

Integration Server X-IntegrationServer-Encode-Headers

The values for this custom header are UTF-8 and ISO-8859-1. Use the X-IntegrationServer-Encode-Headers header in calls if the request/response header values can contain Unicode characters. It indicates whether the request/response headers should be base-64 decoded/encoded.

The value of this header indicates the charset to be used when encoding/decoding custom header values. If this header is omitted, the request and response headers will not be encoded/decoded.

For more information, refer to “Request Headers” in the Integration Server help.

About ZIP file functionality

You should not open UTF-8 and UTF-16 encoded files stored in a ZIP archive in Microsoft Windows Explorer, because the encoded file names do not properly display. Use a third-party product to extract and open ZIP files to ensure that file names and other information displays properly.

Collation and environments using Microsoft SQL Server

Collation in Microsoft SQL Server databases provides sorting rules, language case, and accent sensitivity properties for data in the language you are using. By default, Perceptive Content uses the Western European based collation. The Unicode SQL Server script does not have a default collation. Before installing Content and configuring the database, you must modify the `COLLATE` property in the script to use the appropriate language.

For more information about setting the collation property, refer to the *Unicode Best Practices Guide Foundations EP2*.

Japanese collation and supplementary characters

If you are using Unicode for Japanese, keep in mind that the Japanese_ and Japanese_Unicode_ collations do not support supplementary characters correctly, but instead interpret them as replacement characters. However, the following Japanese collations recognize supplementary characters correctly:

- Japanese_90_CI_AS
- Japanese_Bushu_Kakusu_100_ CI_AS
- Japanese_XJIS_100_ CI_AS

When modifying the `COLLATE` property, ensure that you use a Japanese collation that supports supplementary characters.

About Displaying and inputting Unicode characters

A character must be present in the font set of an operating system in order for it to correctly display as a Unicode character. For example, in order for a computer program or operating system to display Chinese characters, the operating system must contain a Chinese font set. Before installing Content, ensure that your operating system contains the appropriate fonts.

Character input method

There are two primary methods for entering non-English text in a non-English software program. The first method involves using a phonetic keyboard that is specifically designed for the foreign language that is in use, such as a Chinese New Phonetic Keyboard. The second method is with the use of an Input Method Editor (IME), such as the Microsoft Pinyin IME or another third-party IME. The input system you use is dependent on your operating environment and does not affect the ability of Perceptive Content to display non-English characters.

Numerical functionality

All numerical data stored in the database, and numerical data entered by users in Perceptive Content and Interact Desktop are represented with Arabic numerals.

Calendar functionality

For calendar input and display, Content only supports the Gregorian calendar, regardless of your local language.

About Business Insight Unicode constraints

Business Insight does not support computer names that contain Unicode characters. Verify that the name of your computer system does not contain Unicode characters before initiating your installation of Business Insight. If your computer name has Unicode characters, substitute your IP address for the Host Name field in Cognos Config.

Business Insight also does not support report names containing supplementary Unicode characters or a combination of apostrophes and quotation marks. In some situations, certain supplementary characters will not display in prompt text or report parameter text.

Non-Unicode compatible Perceptive Content products, agents, and features

Some Perceptive Content products, agents, and features are not Unicode compatible with the Content Foundation EP2 (7.5.x) release. The following list contains the products and agents that are not Unicode compatible.

- Barcode support
- Datacapture
- DICOM support (Output agent only)

  Note The Unicode version of Content Foundation EP2 (7.5.x) supports DICOM formatting but only with ANSI data. DICOM is not compatible with Output Agent when installed in a Unicode file path, and DICOM will not work if the Output Agent has a Unicode instance name.

- ImageNow Enterprise Report Management (ERM) suite
- Fax Agent
- Forms Server

  Note Transform Parameters do not support Unicode data.
- Mail
- MAPI email support (Client only)
- Recognition Agent
- User Experience Index

About Host names and non-Unicode Perceptive Content products

If you are using any Perceptive Content products that are non-Unicode compatible, then the hostnames of the systems where Content Server is running are restricted to ASCII characters.

About ImageNow Printer

ImageNow Printer does not support UTF-8 encoding. If you change settings in the inowprint.ini file using Unicode characters, ensure that you select the Unicode (UTF-16 or UCS 2/Little Endian) option from the Encoding list when you save the changes.

About Interact Desktop localization

The `[appSettings]` section in the InteractDesktop.exe.config file contains the setting for Interact Desktop localization. Interact Desktop can be localized into any of the languages listed in the following table, and supports the display of any Unicode character that falls along the BMP, as well as supplemental characters, in the user interface.

For more information about configuring this setting, refer to the *Interact Desktop Installation and Setup Guide for 6.8.x*.

Region and Abbreviation columns

The Region column indicates the location-specific language. The Abbreviation column contains an abbreviation for the language and an abbreviation for the locale. For example, Portuguese is abbreviated pt-BR, where pt identifies the language as being Portuguese, and BR identifies the location the language is specific to as being Brazil.

Language	Region	Abbreviation
Chinese (Simplified)	China	zh-CN
Dutch	Netherlands	nl-NL
French	France	fr-FR
German	Germany	de-DE
Italian	Italy	it-IT

Language	Region	Abbreviation
Japanese	Japan	ja-JP
Portuguese	Brazil	pt-BR
Spanish	Spain	es-ES