

Technical Bulletin #21

This technical bulletin describes how to configure Astoria to work properly with UTF-8/UTF-16 encoded documents.

Issue

By default, Astoria only supports XML documents encoded in and explicitly declared as ISO-8859-1 (Latin 1); however, you may want to use Astoria with UTF-8/UTF-16 encoded documents.

Resolution

In order to configure Astoria to properly handle UTF-8/UTF-16 encoded documents, you must add or change the following settings in your **astoria.ini** file, customized as required for your Astoria installation:

[SGML Import]

MaintainCharRefs=1

ReplaceInvalidCharRef=?

DLL0=c:\astoria\bin\xmlfilter.dll

ForceXMLDeclaration=<?xml version="1.0" encoding="ISO-8859-1"?>

[SGML Export]

DLL0=c:\astoria\bin\xmlfilter.dll

Map8BitToEntity=0

[Differencing]

DLL0=c:\astoria\bin\xmlfilter.dll

[XMLFILTER]

Encoding=

For additional details on these entries, please see Chapter 10, Understanding Astoria .ini files in the *Astoria Installation, Configuration, and Administration Guide*.

With these settings, import and save will convert documents to ISO-8859-1 based on the encoding specified in the document's own declaration. An XML document with no encoding specified in its declaration is always assumed to have UTF-8 encoding, which will corrupt any Latin 1 accented characters in the document; therefore, a Latin 1 XML document must specify an ISO-8859-1 encoding in its declaration (e.g., `<?xml version="1.0" encoding="ISO-8859-1"?>`) to properly import the Latin 1 accented characters.

Note: The ForceXMLDeclaration entry in the [SGML Import] section of the **astoria.ini** file specifies the declaration for the imported document once the document is in Astoria. This entry does not affect the document's original file, which must have its encoding specified in its own declaration.

Export and edit will convert documents based on the encoding specified in the Encoding entry in the [XMLFILTER] section of the **astoria.ini** file. The following settings can be used for this entry:

- **1:** On Windows, this setting causes Astoria to show a dialog box that allows you to specify which encoding to use. On Solaris, this setting specifies that ISO-8859-1 (Latin 1) be used during export and edit.
- **2:** This setting specifies that UTF-8 encoding be used during export and edit.
- **3:** This setting specifies that UTF-16 encoding be used during export and edit.
- **4:** This setting specifies that ISO-8859-1 (Latin 1) be used during export and edit