Ryzom - Bug # 1523

Status:	Assigned	**Priority:**	High
Author:	DuDraig	**Category:**	Build
Created:	02/07/2013	**Assignee:**	kervala
Updated:	02/12/2013	**Due date:**	
Subject:	Source files are encoded as windows-1252 or French locale.		

Description

Since this is an open source project that is worked on internationally, the first thing that should have been done is to re-encode all of the source files and any data files that do not have an explicit encoding definition to Unicode UTF-8.

As it is now (on Windows), you must open any source file in a text editor with the windows-1252 encoding. If you are using Visual C++ 2008 or 2010 Express, the only was to view the source files without error is to set your computer language for non-Unicode applications to "French (France)". VC does not provide a per project/solution configuration for source encoding outside of the system default or UTF-8. Changing the default language encoding affects other applications and requires a reboot.

It is a relatively simple task to re-encode all the source to UTF-8 and should not impact compilation or the run-time outside of some possible makefile alterations if the current encoding is specified. It would be of tremendous help to those of us outside of France and is considered a prerequisite in the open source community.

While it is an easy fix, it should be considered a high priority since it has such a fundamental impact on international development environments.

History

#1 - 02/08/2013 07:39 am - DuDraig

<sigh> After a night of coding my English skills turn to rot. Let me clarify that mud in the previous post.

When most projects are turned open source, since most code is written using only the 7-bit ASCII characters in English, conversion to UTF-8 is not necessary since the first 128 codes overlap.

However, the Ryzcom source uses in comments a lot of French characters that are included in the ISO/IEC 8859-1 Western European encoding, which on Windows is a subset of windows-1252. If a user has their OS set to a default language with an encoding that does not overlap that encoding, like any non Western European language, they must jump many troublesome hoops to read or edit the source files without creating problems on their machine or with any committed source.

Conversion of these files to UTF-8 Unicode solves that problem and allows the source to be browsed, edited, and compiled in MSVC++. It also makes my jEdit editor much happier.

I ran into this because my OS is normally set to Japanese (windows-932, a superset of Shift-JIS). It causes me no problems for open source projects that are either pure ASCII or that have been converted to UTF-8 Unicode, but the Ryzcom source went all blowy-uppy.

I wrote and ran a small utility that trod through every .c, .cpp, and .h file, converting them to UTF-8 Unicode as needed. Everything compiles fine but there are probably many other files that should also be converted that do not require another specified encoding.

There were only two source file exceptions:

.../code/nel/tools/3d/object_viewer_qt/src/plugins/object_viewer/widgets/edit_range_widget.h
This file has three copies of the following line:
@brief The widget provides a horizontal slider and 3 QSpinBox(to set start/end value range and current value from this range.).

The 'c' in "current" is encoded in UTF-8 as U+0441 (CYRILLIC SMALL LETTER ES). I changed each to an ASCII 'c'.

.../code/nel/tools/3d/plugin_max/nel_patch_converter/script.cpp
This file has two instance of the character '･' (Katakana Middle Dot) encoded in Shift-JIS (probably due to the very problem this issue addresses - someone committing a file from a machine configured for Shift-JIS). I changed both of those to UTF-8 Unicode encoded U+00B7 (Greek Middle Dot).

If y'all want these converted files, or you'd like to give me guidance towards what other files should also be UTF-8 Unicode, just let me know.

I hope this clarifies the previous mud.

#2 - 02/10/2013 08:42 pm - kervala

We begun to translate all French comments in English and I'm sure it's not finished :) By doing that, all sources will be in plain ASCII. I don't see any advantage to convert sources with french comments in UTF-8, comments will still be not understandable by most of people.

#3 - 02/11/2013 10:10 pm - DuDraig

Understandability is not the point. As long as non ASCII characters exist in the sources not encoded in UTF-8, they cause editing, compiling, and committing problems for any host not set to ISO/IEC 8859-1. The script.cpp file with the Shift-JIS character is an example. If the affected source files were UTF-8, the Shift-JIS encoding would not have happened.

Since it will take more time to translate the French comments, it would seem prudent and helpful to developers to encode the affected files now, something that is easy and quick to do that does not impede the continuing translation.

#4 - 02/11/2013 10:35 pm - kervala

I just fixed some files, I searched for some French accents but I can't detect invalid characters such as Cyrillic ones.

Please could you help us ? You did a script to check the files with a bad encoding, didn't you ?

So please could you post the list of all files with a bad encoding ? So we'll be able to fix the problem :) Thanks !

If I prefer to translate from French to English, that's because I have no way to find out what files have French comments, if I had a list I could do that really quick.

#5 - 02/12/2013 05:00 pm - DuDraig
- File UTF-8 Files.7z added
- File Encoding notes.txt added

Easily done. That is what I offered.

I have attached an archive of ISO-8859-1 encoded .h, .c, and .cpp source files containing non US-ASCII characters that have been re-encoded to UTF-8. Just extract it in the Ryzom source root directory. There is also a text file listing all of the files in the archive.

Two files were already UTF-8 so are not included but need translation.
Four I've been edited to US-ASCII since they had either a BOM or non US-ASCII characters that look like US-ASCII characters.
127 were re-encoded to UTF-8 and need translation.

The converted source tree on Windows 7 in Japan locale (Shift-JIS encoding) with VC++ 2010 Express now displays source files correctly and compiles without error.
I do not know about any other files that might also need re-encoding or translation, like makefiles or text files, but this should at least make things usable to international developers.

I hope you find this of some help.

#6 - 02/12/2013 05:09 pm - kervala

- Status changed from New to Assigned
- Assignee set to kervala
- % Done changed from 0 to 10

Thank you so much for your contribution :)

I'm working on it !

#7 - 02/12/2013 05:29 pm - DuDraig

- File UTF-8 Files.7z added

<sigh> 7-zip has betrayed me. It did not append the custom edited 5 files. It replaced the archive with just those 5 files.

Attached is the correct archive with all the files.

Files

UTF-8 Files.7z	15.7 kB	02/12/2013	DuDraig
Encoding notes.txt	10.6 kB	02/12/2013	DuDraig
UTF-8 Files.7z	377.4 kB	02/12/2013	DuDraig