Subject: Re: what about compressed RailML files?
Posted by Nilo Menezes on Mon, 15 Oct 2012 10:49:32 GMT
View Forum Message <> Reply to Message

Hello Dirk,

This is my first post on the RailML group. I wrote an internal tool that reads RailML 2.1 files and provide some operations on it (time table extraction and track export based on route). I work at Multitel, Belgium, at the Certification Laboratory.

Regarding your message, may I suggest using Gzip instead of zip?

Why:
1) GZip is streaming friendly, you can read the compressed file directly, no need to decompress first. This also make GZip files very welcome on command line applications.
2) You can only add a single file to it. In fact, GZip does not specify internal files, all you have a single stream. To get the file name, we process the .gz file name itself.
3) The overhead is very small.
4) Most software libraries and languages provide GZip compression/decompression (Python, Ruby, C/ZLib, Java, C#, etc).

For the file extension:
..railml for uncompressed files
..railml.gz for gzipped RailML files (following Unix tradition like ..tar.gz or .tar.bz2)

Regarding the points you listed:
On 05/07/2012 18:39, Dirk Bräuer wrote:

> There are some questions we should consider:
> - Do we recommend file extensions and if so, which?
It is a very good idea. Anything different from .xml would be nice. I have a lot of problems opening large RailML xml files with the wrong tools on Windows. With .xml it is harder to create a specific file association too.

> - Do we enforce Deflate compression algorithm or do we allow others?
If we use gzip, this question would be already answered.

> - Do we allow more than one RailML file in one ZIP file?
I recommend only one file. If the user needs more files, he can create a tar or use another program for that.
Maybe I'm missing something here, but what do you mean by more than one file? Would they share the same references? Is this grouping a kind of context somehow?

> - Do we enforce UTF-8 file names in the ZIP file or do we allow also the
> older but default Ansi-437 ? (Bit 11 of GeneralPurposeBitFlag of the
> CommonFileHeader of ZIP would allow to distinguish between both).
UTF-8 is widely spread. Enforcing ANSI-437 can be annoying for international use. The European page for example is the 850. I'm not sure if these code pages are ANSI standards, I think they are just code pages created by IBM and Microsoft.
UTF-8 is welcome on Windows, Mac OS X and Linux. So I think we would make everybody happy. If we adopt the GZip format, any problems regarding file name encoding would be solved by a simple rename.

> - Do we 'allow' or 'recommend' the compressed RailML files?
It is very easy to accept both. On my tool, if you decide to use .gz, it will change very few lines of code. Uncompressed files are great when we are tweaking them. Compressed files are great for transmission and storage. I work with 150Mb XML files... I would not like to compress and uncompress them every time I change a letter or something.

Best Regards,

Nilo Menezes
