Exiv2 - Feature #1311

Support Unicode in metdata

05 Sep 2017 20:09 - Arnold Wiegert

Status:	New	**Start date:**	05 Sep 2017
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	metadata	**Estimated time:**	0.00 hour
Target version:	1.0		

Description

Images can contain Unicode metadata in IPTC-IIM fields
hopefully exiv2 will be able to as well

History

#1 - 06 Sep 2017 14:32 - Robin Mills

- File metadata-test.jpg added

- Category set to metadata

- Target version set to 1.0

There is a sample file and analysis here: http://dev.exiv2.org/boards/3/topics/2944?r=2954#message-2954

I don't know what it means ***"to support UNICODE in metadata"***. For sure, when I extract the XML from metadata-test.jpg (and reformat it in ascii), on Windows, I see:

```
C:\Users\...> exiv2 -pX metadata-test.jpg | xmllint --format -
<?xml version="1.0"?>
<?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc9d"?>
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="XMP Core 5.5.0">
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:lr="http://ns.adobe.com/lightroom/1.0/"
 rdf:about="">
      <dc:subject>
        <rdf:Bag>
          <rdf:li>Places</rdf:li>
          <rdf:li>Germany</rdf:li>
          <rdf:li>Baden-W&#xFC;rttemberg</rdf:li>
          <rdf:li>Umlaut&#xDF;&#xFC;&#xF6;&#xE4;&#xDC;&#xD6;&#xC4;</rdf:li>
        </rdf:Bag>
      </dc:subject>
      <lr:hierarchicalSubject>
        <rdf:Bag>
          <rdf:li>Places</rdf:li>
          <rdf:li>Places|Germany</rdf:li>
          <rdf:li>Places|Germany|Baden-W&#xFC;rttemberg</rdf:li>
          <rdf:li>Places|Germany|Umlaut&#xDF;&#xFC;&#xF6;&#xE4;&#xDC;&#xD6;&#xC4;</rdf:li>
        </rdf:Bag>
      </lr:hierarchicalSubject>
    </rdf:Description>
  </rdf:RDF>
</x:xmpmeta>
<?xpacket end="w"?>

C:\Users\rmills\gnu\github\clanmills\exiv2\contrib\cmake\msvc>
```

The output on the Mac is identical:

```
658 rmills@rmillsmm:~/gnu/github/clanmills/exiv2 $ exiv2 -pX metadata-test.jpg | xmllint --format -
<?xml version="1.0"?>
<?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc9d"?>
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="XMP Core 5.5.0">
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:lr="http://ns.adobe.com/lightroom/1.0/"
 rdf:about="">
      <dc:subject>
        <rdf:Bag>
```

```
          <rdf:li>Places</rdf:li>
          <rdf:li>Germany</rdf:li>
          <rdf:li>Baden-W&#xFC;rttemberg</rdf:li>
          <rdf:li>Umlaut&#xDF;&#xFC;&#xF6;&#xE4;&#xDC;&#xD6;&#xC4;</rdf:li>
        </rdf:Bag>
      </dc:subject>
      <lr:hierarchicalSubject>
        <rdf:Bag>
          <rdf:li>Places</rdf:li>
          <rdf:li>Places|Germany</rdf:li>
          <rdf:li>Places|Germany|Baden-W&#xFC;rttemberg</rdf:li>
          <rdf:li>Places|Germany|Umlaut&#xDF;&#xFC;&#xF6;&#xE4;&#xDC;&#xD6;&#xC4;</rdf:li>
        </rdf:Bag>
      </lr:hierarchicalSubject>
    </rdf:Description>
  </rdf:RDF>
</x:xmpmeta>
<?xpacket end="w"?>
659 rmills@rmillsmm:~/gnu/github/clanmills/exiv2 $
```

When the data is output to the terminal (without the filter: ***xmllint --format -***), Windows and the Mac do not display bytes 128-255 with the same glyphs. However, that is the different behaviour by the Mac's Terminal and Window's cmd.exe console.

To my knowledge, metadata strings (in IPTC and Exif) is stored as count + binary sequence of bytes. The standards do not say that it has to be stored as UTF-8, or anything else. On the exiv2 command-line, there is a switch -n (--encoding) to define the Charset for User comments.

One matter that could be investigated is to modify the Windows samples to operate in UNICODE. Be default, we build the Exiv2 library without enabling EXV_UNICODE_PATH and therefore only offer ascii interfaces into the file-system. When EXV_UNICODE_PATH is enabled, the only sample which takes advantage of this is exifprint. A useful project could be to:

1) Always build msvc libraries with EXV_UNICODE_PATH enabled.
2) Convert all sample applications to use wmain() for msvc builds.
3) Respect -n --encoding option to tell the samples how to convert wstring to binary.
4) Study what should be done for Cygwin and msgs/2.0

This could be a good project for Google Summer of Code. If we recruit a Chinese or Indian student, they would have the language to effectively test this.

Files

metadata-test.jpg	70.6 KB	06 Sep 2017	Robin Mills