LeechCraft - Bug #1891

Проблемы с автоопределением кодировок тегов локальных файлов.

03/02/2015 01:17 AM - Mellon

Status:	Closed	**Start date:**	03/01/2015
Priority:	Low	**Due date:**	
Assignee:	0xd34df00d	**% Done:**	100%
Category:	Plugins: LMP	**Estimated time:**	7.00 hours
Target version:	0.6.75		
Reported in:	master		

Description

В общем, не работает.

Эта штука работает с потоками, как и показано в #1250 и #1271 , но с локальными файлами дело обстоит иначе.
Должна работать. но не работает.

Воспроизвести можно так:
0. Включить в настройках автоопределение кодировки. Выбрать китайский язык.
1. Загрузить в плейлист прилагаемый файл, предварительно распаковав его.

По идее, всё это должно перекодировать теги по схеме utf8 → cp1252 → libguess → utf8
Вот пример того, как это должно работать:

```
% GST_DEBUG=*id3*:5 gst-launch-0.10 filesrc location="/tmp/□□□□□□□□/02 □□□□□.mp3" ! id3demux ! fake
sink -t >! /tmp/in
здесь можно обратить внимание на "Read 1 fields from Text ID frame of size ## with encoding 0. " о
тсюда растут ноги cp1252

% cat /tmp/in | iconv -f utf-8 -t cp1252 -c > /tmp/in1
при этом, без указания ключа игнорирования ошибок '-c', iconv вывалится с ошибкой на illegal input
 sequence at position 215

% guessconv -l chinese -i /tmp/in1 -o /tmp/in2
Input file: /tmp/in1
Output file: /tmp/in2
Input encoding: GB2312
Output encoding: UTF-8

% cat /tmp/in1 | iconv -f GB2312 -t utf8 -c
Setting pipeline to PAUSED ...
Pipeline is PREROLLING ...
FOUND TAG      : found by element "id3demux0".
    album artist: □□□□
           album: □□□□□□□□
           title: □□□□
    track number: 2
           image: buffer of 103492 bytes, type: image/jpeg, width=(int)500, height=(int)500, sof-m
arker=(int)0, image-type=(GstTagImageType)GST_TAG_IMAGE_TYPE_UNDEFINED
     ID3v2 frame: buffer of 13 bytes, type: application/x-gst-id3v2-tcmp-frame, version=(int)3
container format: ID3 tag
Pipeline is PREROLLED ...
Setting pipeline to PLAYING ...
New clock: GstSystemClock
Got EOS from element "pipeline0".
Execution ended after 48213465 ns.
Setting pipeline to PAUSED ...
Setting pipeline to READY ...
Setting pipeline to NULL ...
Freeing pipeline ...
```

Причиной того, что этого не происходит, скорее всего, являются как раз те самые проигнорированные ошибки

рекодирования.
Ибо, если рекодировать напрямую, то корректным результатом должно стать:

```
% id3info "/tmp/□□□□□□□□/02 □□□□□.mp3" | grep '===' | rcc-recode -l zh
=== TPE2 (Band/orchestra/accompaniment): □□□□
=== TALB (Album/Movie/Show title): □□□□□□□□
=== TIT2 (Title/songname/content description): □□□□
=== TRCK (Track number/Position in set): 2
=== APIC (Attached picture): ()[, 0]: image/jpeg, 103492 bytes
=== TCMP ():  frame
```

Здесь следует обратить внимание на различие:

```
title: □□□□
=== TIT2 (Title/songname/content description): □□□□
```

Выводы.

1. Ожидать хорошего функционирования перекодировки тегов, при работе с ними через gstreamer, было бы ошибкой. Gstreamer следуя спецификациям стандарта, занимается рекодингом опираясь на соответствующие заголовки ID3 читая теги и переменные среды окружения — их выдавая.

2. Несмотря на продвинутость libguess, эта билиотека часто путает схожие кодировки, что может привести к ошибкам рекодирования. проект librcc, также используя libguess делает более правильные предположени, основываясь на своих правилах.

Что можно попробовать сделать:

1. Допускать игнорирование ошибок рекодинга // Очень сомнительная затея

2. Добавить в настройку возможность указания конкретной кодировки. их соответствие языкам можно взять из приложенного файла (в дальнейшем его будет не сложно дополнить).

Associated revisions

Revision 4c71d22d - 03/05/2015 01:39 AM - 0xd34df00d

[LMP] Reworked stream tags recoding settings for #1891.

Revision bd41e84a - 03/05/2015 01:54 AM - 0xd34df00d

[LMP] Immutable FixEncoding() for #1891.

Revision 168194ff - 03/05/2015 01:54 AM - 0xd34df00d

[LMP] Exposed FixEncoding() for #1891.

Revision 39305e4b - 03/05/2015 02:07 AM - 0xd34df00d

[LMP] Added an option for local tags recoding for #1891.

Revision 8e6ffe38 - 03/05/2015 02:07 AM - 0xd34df00d

[LMP] Recode local tags for #1891.

Revision 414a69cb - 03/05/2015 02:10 AM - 0xd34df00d

[LMP] Flush LocalFileResolver cache on recoding changes.

This finally fixes #1891.

History

#1 - 03/02/2015 06:27 PM - DA

У меня и русских половину не декодирует :) Могу приложить, коль понадобится.

#2 - 03/03/2015 04:16 AM - 0xd34df00d

- Status changed from New to Assigned

А, естественно. Рекодинг происходит только тогда, когда теги берутся от гстримера. Теги берутся от гстримера только для нелокальных источников, иначе используется taglib.

#3 - 03/04/2015 09:17 AM - Mellon

Ок. чтение тегов локальных файлов ведется (при загрузке/сканировании файлов) не через гстример.
Теперь taglib.
Taglib страдает той же херью, и если в ID3v2 не указана кодировка тегов, то чтение производится в iso8859-1. Вообще-то, скорее всего не "тоже страдает", а скорее всего taglib является бекэндом id3demux гстримера, а может и нет, но не важно.

Так вот. Надо либо все теги кормить libguess, либо, раз работает через taglib, делать чутка анализ. Например, в следующих случаях, есть вероятность, что там крякозябры:
1. Есть только ID3v1
2. В ID3v2 нет указания кодировки тегов.
Если соблюдаются эти условия, то можно проверить на соответствие символов диапазону ANSI, либо сразу кормить libguess.

#4 - 03/04/2015 09:26 AM - Mellon

кстати, на счет гстримера. он же выдает в дебаг все подробности того, какие теги и с какой кодировкой, а значит, эти данные всё-таки можно получить. и использовать.

#5 - 03/05/2015 02:17 AM - 0xd34df00d

- Estimated time set to 7.00 h

- Target version set to 0.6.75

#6 - 03/05/2015 02:17 AM - 0xd34df00d

- % Done changed from 0 to 100

- Status changed from Assigned to Resolved

Applied in changeset main|414a69cb086a80609725354ca2fd70d9a7c6d82f.

#7 - 03/05/2015 02:20 AM - 0xd34df00d

В итоге я не парюсь и сразу кормлю либгессу.

> есть вероятность, что там крякозябры:
> 1. Есть только ID3v1
> 2. В ID3v2 нет указания кодировки тегов.

Кроме mp3 есть и другие аудиофайлы. ИМХО надёжнее просто всегда прогонять через либгесс, да.

> кстати, на счет гстримера. он же выдает в дебаг все подробности того, какие теги и с какой кодировкой, а значит, эти данные всё-таки можно получить. и использовать.

Не прокатит для локальных файлов на момент обновления коллекции и добавления в плейлист.

#8 - 03/05/2015 06:09 AM - Mellon

DA, проверь, пазязя, как там с русским.

Касательно восточно-азиатских кодировок, LC теперь работают корректно (при соблюдении некоторых условий). Если и есть ошибки (а их много), то только из-за несовершенства libguess.

И да, Проблемы с перекодировкой тегов на этом не исчерпываются. То есть будет продолжение. Но уже не тут :3

#9 - 03/06/2015 03:34 AM - Mellon

0xd34df00d wrote:

> Не прокатит для локальных файлов на момент обновления коллекции и добавления в плейлист.

https://taglib.github.io/api/classTagLib_1_1ID3v2_1_1TextIdentificationFrame.html#ae4d221c60dcfb097e471c86fbc6efbae
http://taglib.github.io/api/classTagLib_1_1ID3v1_1_1StringHandler.html
http://taglib.github.io/api/classTagLib_1_1ID3v2_1_1Latin1StringHandler.html

Это типа такой ботолурк был.

#10 - 03/06/2015 03:39 AM - Mellon

Хм... то есть, намекаю на то, что, если в тегах указана кодировка и она не Latin1/iso8859-1, то натравливать libguess может не стоит, например.

#11 - 03/15/2015 10:29 PM - Mellon

- Status changed from Resolved to Closed

на моей русне вроде работает.
если что не так, пожалуйста переоткройте.

Files

rcc.xml	10.2 KB	03/01/2015	Mellon
song.tar.xz	8.56 MB	03/01/2015	Mellon