ITDevCon

Delphi European Conference

Regular Expressions – Friend or Foe?

Primož Gabrijelčič

October 25/26 2012 VERONA

bit Time software

Primož Gabrijelčič

programmer, consultant, speaker, trainer

Delphi / Smart Mobile Studio

Email: primoz@gabrijelcic.org
Twitter: @thedelphigeek
Skype: gabr42

The Delphi Geek – http://www.thedelphigeek.com
Smart Programmer – http://www.smartprogrammer.org

ITDevCon

The Delphi Geek

andom ramblings on Delphi, programming, Delphi programming, and all the rest

Tuesday, October 16, 2012

ITDevCon 2012 is almost here ...

.. and I'm busy preparing my presentations. Come join me in Verona and listen to following talks:

Regular expressions - friend or foe?

Regular expressions are one of the most underused features of the Delphi RTL. While in the past we could attribute this to the lack of built-in support, Delphi XE introduced the RegularExpressions unit which greatly simplifies the use of the regular expressions engine. This session will present basic ideas behind the regular expressions, examine the RegularExpressions unit and in particular the main workhorse - the TRegEx class - and continue with practical examples which will show how and when to use regular expressions - and particularly when to stay away from them.

"Hands On": Parallel programming with OmniThreadLibrary

In the past few years, OmniThreadLibrary has become "de facto" standard for Delphi multithreaded programming. Still, the main stumbling block for programmers to "go multithreaded" is the grasp of patterns and practices for multithreaded development. This "hands on" session will take different practical examples, deconstruct them into basic operations and then show how to build simple parallel solutions based on the initial analysis.

"Hands-On": Developing for Windows and OS X

Multiplatform development is slowly taking hold in the Delphi world. While we can argue that the tools for the mobile platform are still in the infancy stage, the situation on the desktop is quite different. Delphi is a valid development tool for the OS X-based computers. The main topic of this session will be hassle-free multiplaftorm development - how to write your code that it "simply works" on both platforms and what to do when this is not possible.

(That is, if you'll not be listening to other great presentations running at the same time!)

Pages

Presentations

Introduction to regular expressions

ITDevCon

Introduction

“In computing, a regular expression provides a concise and flexible means to "match" (specify and recognize) strings of text, such as particular characters, words, or patterns of characters.”

-Wikipedia

“A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules.”

-About.com

History

- Originated in the Unix world
- Many flavors
  - Perl, PCRE (PHP, Delphi), .NET, Java, JavaScript, Python, Ruby, Posix …

Usage

- Testing (matching)
- Searching
- Replacing
- Splitting

Limitations

- Slow(ish)
- Can use lots of time and memory
- Unsuitable for some purposes
  - HTML parsing
- UTF-8

Tools

- Editors
- grep/egrep/fgrep
- Online tools
  - regex.larsolavtorvik.com
- RegexBuddy, RegexMagic
  - www.regexbuddy.com/regexmagic.html

Delphi

- RegularExpressions, RegularExpressionsCore
  - Since XE
- TPerlRegex
  - Up to 2010
- PCRE flavor

Example

- Search for "Handel", "Händel", and "Haendel"
  - `H(ä|ae?)ndel`
  - `Handel|Händel|Haendel`

- `if TRegEx.IsMatch(s,'H(ä|ae?)ndel') then`

Syntax

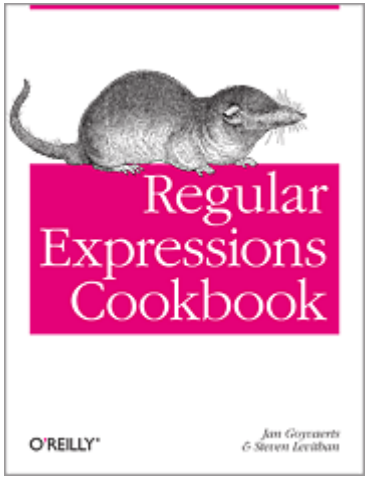
Literals and Metacharacters

- Metacharacters
  - `$()*+.?[\^{|`
- Literals
  - Everything else
- Escape
  - `\`
- Nonprintable
  - `\n, \r`

Tutorial

- www.regular-expressions.info/tutorial.html
- www.regular-expressions.info/delphi.html

- Jan Goyvaerts, Steven Levithan – Regular Expressions Cookbook (Amazon, O'Reilly)



Character class, Alternatives, Any

- One-of
  - `[abc]`
  - `[a-fA-F0-9]`
  - `[^a-fA-F0-9]`
- Alternatives
  - `Delphi|Prism|FreePascal`
- Any
  - `.`

Shorthands

- \d, \D
  - `[0-9]`, `[^0-9]`
- \w, \W
  - `[a-zA-Z0-9_]`, `[^a-zA-Z0-9_]`
- \s, \S
  - `[ \t\r\n]`, `[^ \t\r\n]`

Anchors

- Start of line/text
  - ^, \A
- End of line/text
  - $, \Z, \z
- Word boundary
  - \b, \B

Unicode

- Single grapheme
  - `\X`
- Unicode codepoint
  - `\x{2122}` ™
- \p{category}
  - `\p{N}`
- \p{script}
  - `\p{Greek}`

Groups

- Capturing group
  - `(\d\d\d)`
- Noncapturing group
  - `(?:\d\d\d)`
- Named group
  - `(?P<digits>\d\d\d)`

Group references

- Unnamed reference
  - `\1`, `\2`, … `\99`
- Named reference
  - `(?P=digits)`

- Example
  - `(\d\d\d)\1`

Repetitions

- Exact
  - `{42}`
- Range
  - `{17,42}`
  - `[a-fA-F0-9]{1,8}`
- Open range
  - `{17,}`

Repetition shortcuts

- `?`
  - `{0,1}`
- `+`
  - `{1,}`
- `*`
  - `{0,}`

Repetition variations

- Non-greedy
  - `*?`, `+?`
- Possesive
  - `*+`, `++`, `?+`, `{1,3}+`

Modifiers

- Case-insensitive
  - `(?i)`, `(?-i)`
- Dot matches line breaks (‘single-line’)
  - `(?s)`, `(?-s)`
- ^ and $ match at line breaks (‘multi-line’)
  - `(?m)`, `(?-m)`

Search & Replace

- `\1..\99` reference to a group
- `\0` all matched text
- `(?P=group)` reference to a named group
- `\`` left context
- `\'` right context

Examples

- Username
  - `[a-z0-9_-]{3,16}`
- Email (simplified)
  - `([a-z0-9_\.-]+)@([\da-z\.-]+)\.([a-z\.]{2,6})`
- IP (dotted, v4)
  - `([0-9]{1,3}\.){3}[0-9]{1,3}`

Bad examples

- File name
  - (?i)^(?!^(PRN|AUX|CLOCK\$|NUL|CON|COM\d|LPT\d|\..*)(\..+)?$)[^\\\./:*\?\"<>\|][^\\/:*\?\"<>\|]{0,254}$
- Parsing HTML with RegEx
- Catastrophic backtracking
  - (x+x+)+y

Delphi IDE

ITDevCon

Search & Replace

- Different flavor
- docwiki.embarcadero.com/RADStudio/XE3/en/Regular_Expressions
- Groups
  - {}
- ?, | are not supported

Search & Replace demo

- Find {$IFDEF} and {$IFNDEF}
  - \$IFN*DEF
- Replace {$IFN?DEF WIN64} with {$IFN?DEF CPUX64}
  - \{\$IF(N*)DEF WIN64\}
  - \{$IF\1DEF CPUX64\}

Code Examples

Questions?

ITDevCon