A Brief History of PCRE

Philip Hazel

http://www.quercite.dx.am

Perl-
Compatible
Regular
Expressions

My CS Career

- 1971: Programming for the Titan; all software home-grown
- 1972: First IBM mainframe; much support software still needed
- 1970’s & 1980’s: Text editors, typesetters, comms, and other things

↓ ↓ ↓ Fast forward ↓ ↓ ↓

- 1995: After several years of rapid change

  Unix servers replace IBM mainframe

  Internet replaces X25 network

  More flexible Mail Transfer Agent needed

- Development of the Exim MTA

  Experimental Internet Mailer

Why Regular Expressions?

- A regex is a pattern for matching in a string
  - Often a string of text but could be a string of binary
  - Based originally on mathematical set theoretical regular expressions
- An MTA can use regexes for
  - Selecting domains or hosts for routing (or blocking)
  - Validating local email addresses (e.g. CRSIDs)
  - Spam detection
- Henry Spencer library for simple regexes
  - Repetition, alternation, classification
- Perl was extending the power of its patterns
- 1997: PCRE written in the summer, bundled with Exim
  - Many releases September–December

I thought it was all over

- ASCII is not enough
    - Locale support (ISO 8859 and other encodings)
    - UTF-8 support for Unicode character set
    - Unicode property support
    - Even EBCDIC support
- Portability: Windows & CMake support
- Getting ahead of Perl
    - Recursive patterns
    - Possessive quantifiers from Java
    - Named groups from Python
- Other extensions
    - Partial matching
    - Newline options (CR, LF, CRLF, etc)
- User patches for performance enhancement
    - String relocations (270 reduced to 22)
    - Unicode property 2-stage inline lookup

2007: I retired, but no let-up

- 2011: Just-in-time (JIT) compiler support

  Followed by 16-bit & 32-bit support

- 2015: API overhaul; release of PCRE2 (versions 10.*xx*)

  All-functional interface

- 2017: Major code refactoring

  Compile: added prepass identifies all groups

  Match: use heap instead of stack for backtracking

- And all the time Perl continues to develop

  Example: addition of script run detection (2019)

- PCRE is now a standard package in many Linux distributions

  Used by Apache, PHP, KDE, Safari, Mathematica (to name but a few)

- Testing: ASAN, valgrind, build farm, fuzzers, Coverity Scan

Regex Entertainments

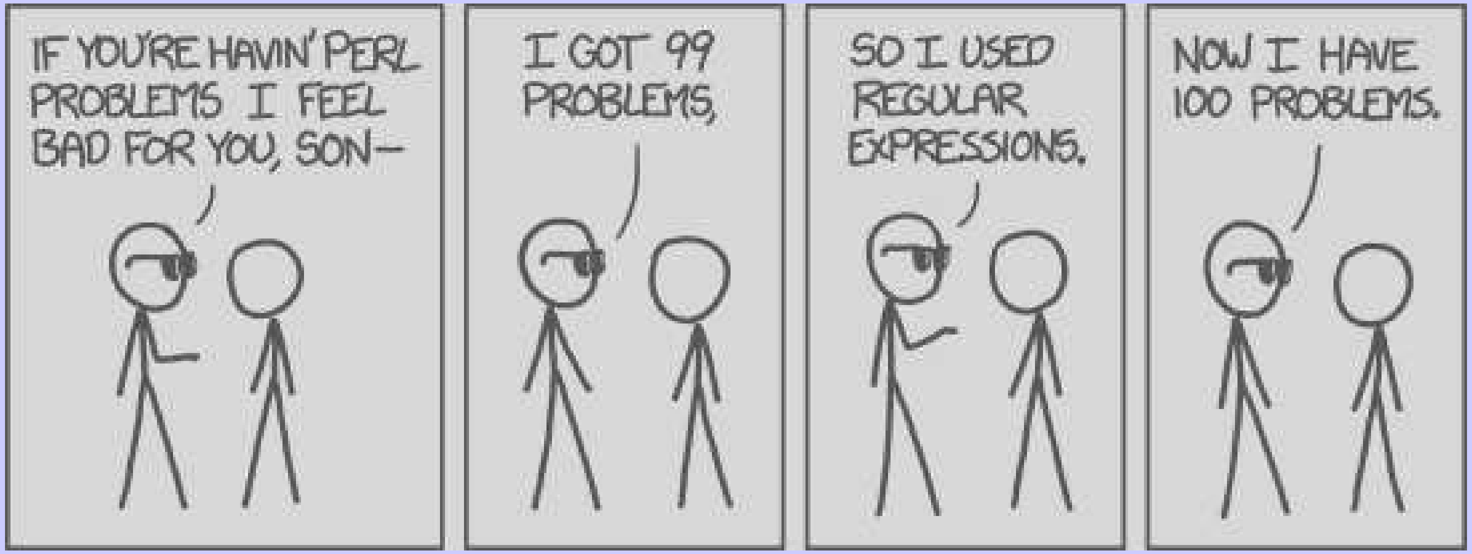
- Palindrome

```
/^\W*+(?:((.)\W*+(?1)\W*+\2|)|((.)\W*+(?3)
   \W*+\4|\W*+.\W*+))\W*+$/i
```

- Pangram

```
/^(?=.*(?=(([A-Z]).*(?(1)\1)))(?!.+\2)){26}/i
```

- *Mastering Regular Expressions* (3rd Edition) by Jeffrey Friedl
- *http://www.rexegg.com/* Tutorials and discussions
- *https://regex101.com/* Interactive regex tester


IF YOU'RE HAVIN' PERL PROBLEMS I FEEL BAD FOR YOU, SON–
I GOT 99 PROBLEMS,
SO I USED REGULAR EXPRESSIONS.
NOW I HAVE 100 PROBLEMS.