Uri’s, curies en reguliere expressies

|  |  |  |
| --- | --- | --- |
| Joop Ringelberg | 02-12-17 | Versie: 1 |

# Introductie

De contexten, rollen en properties van Perspect identificeren we met Uri’s of met curies. De Uri’s zijn ofwel standaard namespaces (zie hieronder), ofwel opgesteld volgens ons eigen *proprietary[[1]](#footnote-1)* model scheme. Dat schema luidt:

model:DomeinNaam#ContextNaam

Curies ontstaan door prefixes als afkorting voor domeinnamen te introduceren en te combineren met context-, rol- en propertynamen.

We gebruiken ook de default prefix voor de Perspect namespace.

Tenslotte nog een conventie:

1. de namen van contexten, rollen en properties beginnen we altijd met een hoofdletter, behalve
2. in property-assignment expressies en rol-binding expressies.

# Reguliere expressies

Een voor de hand liggende reguliere expressie voor een domeinnaam zou zijn:

/model\:\w+\#\w/

De bijzondere karakters (: en #) zijn ge-escaped met \. De domeinnaam (en de naam achter de #) wordt gematched met de \w character class. Deze groep karakters omvat alle gewone ‘woordletters’. Althans, voor het Engels, Nederlands en nog een paar westerse talen. Sinds de invoering van Unicode zijn er veel meer, gestandaardiseerde, ‘woordletters’. Helaas worden die niet standaard ondersteund in reguliere expressies.

EcmaScript standaardiseert ook reguliere expressies. Inmiddels ligt er een voorstel[[2]](#footnote-2) klaar voor uitbreiding van die standaard voor Unicode. Dit voorstel is nog niet definitief opgenomen in de standaard maar een aantal grote Javascript implementaties heeft hem toch alvast opgenomen. Daaronder bevindt zich V8, de javascript machine achter Chrome en NodeJs.

De uitbreiding zal beschikbaar zijn in versie 64 van Chrome. Deze versie is inmiddels als Canary beschikbaar en verwacht wordt dat hij eind januari 2018 in het *stable channel* van Chrome wordt vrijgegeven.

# Implementatie

Per December 2017 implementeren we in de versie van Ace die Perspect ondersteunt, de reguliere expressies zonder deze unicode uitbreiding. Maar in de code van de module Perspectives.ContextRoleParser staan in commentaren de betreffende unicode reguliere expressies in commentaar.

# Curies en uries in de parser

De ContextRoleParser gebruikt parsers die wel degelijk op de unicode property classes zijn gebaseerd die ook in de uitgebreide reguliere expressies van Ecmascript zijn opgenomen. Per December 2017 accepteert de parser dus een veel groter aantal curies en uries dan de reguliere expressies die voor syntax coloring, indenting en folding in Ace zorgen.

# Appendix

Zie ook: <https://mathiasbynens.be/notes/es6-unicode-regex>. Unicode-aware regular expressions in ECMAScript 6.

1. Proprietary betekent: gepatenteerd. In dit verband: gedeponeerd als schema. Dat moeten we nog doen. [↑](#footnote-ref-1)
2. <https://github.com/tc39/proposal-regexp-unicode-property-escapes>. ECMAScript proposal: Unicode property escapes in regular expressions. [↑](#footnote-ref-2)