LGR Toolset:
A tale of implementing an LGR processor

audric.schiltknecht@viagenie.ca, marc.blanchet@viagenie.ca
Viagénie
wil@cloudregistry.net
Cloud Registry
LGR Toolset

Tool to help LGR designers create their LGR:
• Web front-end with a Python backend
• Open source
• Define and manage code points and variants
• Validations
• Labels to test against, …
• LGR XML format can be complicated for some use cases and is cumbersome for non-XML savvy people
Unicode dependency

- LGR files can use whatever Unicode version
- Language/(3rd-party) libraries are generally linked to a specific Unicode version
- Use existing regex engine or develop from scratch?
Regex Engine

• Existing:
  − Need a *shim* to abstract Unicode management
  − Dependant on library release cycle for future Unicode updates
  − Enjoy the existing validation and tests
  − Not all RECOMMENDED properties supported

• Scratch:
  − Complex (understand: cost more)
  − Stick to your own release cycle
Label eligibility

• “Differed” label eligibility:
  – Label must be valid per LGR (all code points in LGR + context rules)
  – Compute label disposition with reflexive mappings

• Clarifications added in -03
Variant generation

- Depending on LGR, variant space can be large, especially if there are sequences/null variants.
- Duplicate variants: multiple occurrences of the same variant label with different disposition. These must be detected: need to keep variant list!
- Try to limit label length to mitigate potential DoS
Duplicate variants

• From the draft:
  
  `<char cp="0061">`
  `<var cp="0061" type="allocatable"/>`
  `</char>`
  `<char cp="0062"/>`
  `<char cp="0061 0062">`
  `<var cp="0061 0062" type="blocked"/>`
  `</char>`
  
  • With input label “ab”, two variants:
    
    `{a}{b} (allocatable), {ab} (blocked)`
Variants space stats

• Latest Arabic LGR:
  – Number of code points: 128.
  – Total number of variants: 192.
  – Average number of variants per code point: 3.

• Average number of variants per label length on a set of 161 labels:
  – 5 -> average # of variants: 193 (max: 5120)
  – 8 -> average # of variants: 3806 (max: 12800)
Conclusion

• Discussions on ML to clarify draft (label eligibility, add warnings regarding variant generation)

• Guidelines for LGR writers to optimize processing (eg. rule ordering)

• Need to implement mechanism(s) to limit label length to prevent resources exhaustion

• More info: audric.schiltknecht@viagenie.ca