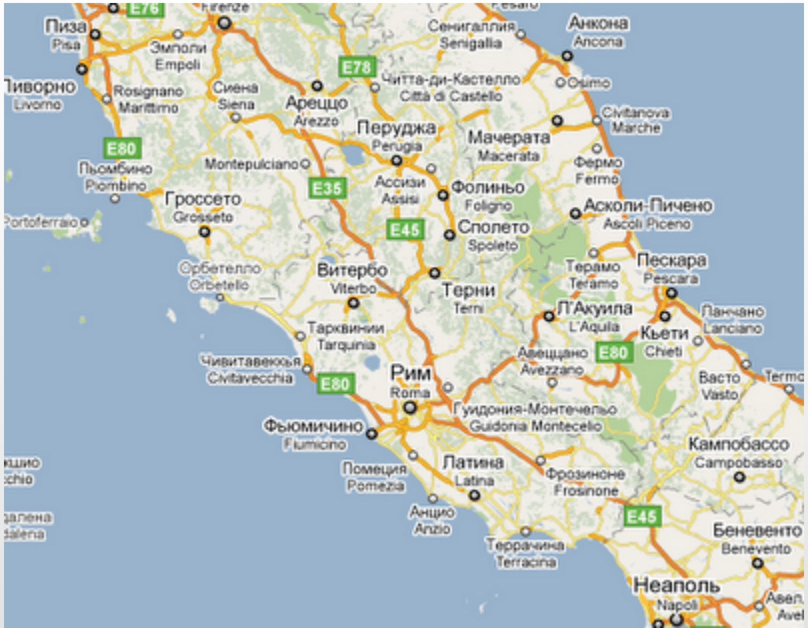
Unicode BCP47 Extensions

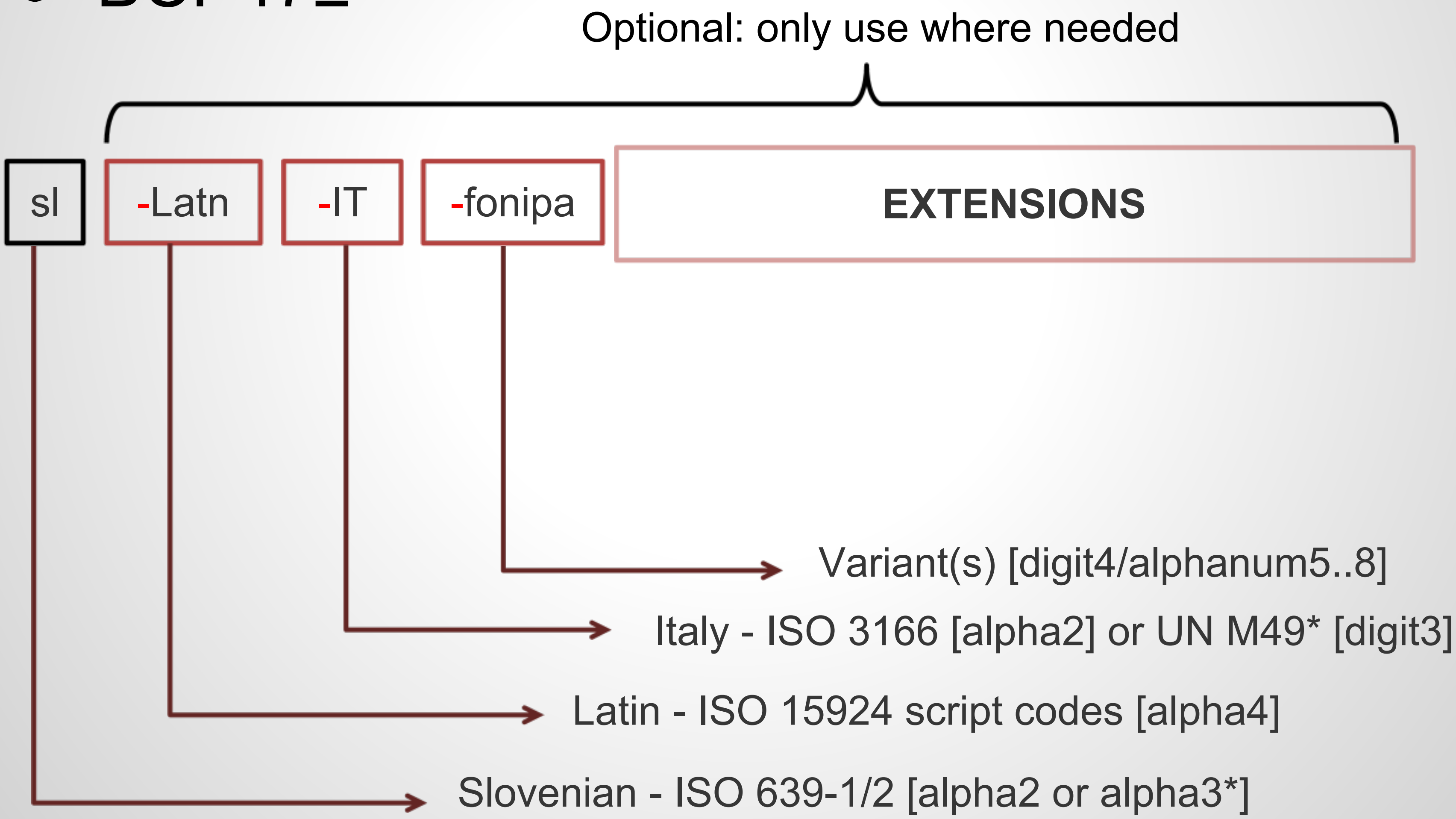
Mark Davis

http://goo.gl/owbBk



Unicode Locale/Lang ID

- BCP47±



Optional: only use where needed

sl -Latn -IT -fonipa EXTENSIONS

Variant(s) [digit4/alphanum5..8]

Italy - ISO 3166 [alpha2] or UN M49* [digit3]

Latin - ISO 15924 script codes [alpha4]

Slovenian - ISO 639-1/2 [alpha2 or alpha3*]

Extension U: Unicode Locales

- RFC6067
- Two-letter keys…
  - ca - bcp47/calendar.xml
  - nu - bcp47/number.xml
  - co - bcp47/collation.xml
    - + specialized collation settings: ka,…
  - cu - bcp47/currency.xml *(compat)*
  - tz - bcp47/timezone.xml *(compat)*
- … + values

U Examples

- th-u**-ca-buddhist**
    - Thai with Buddhist calendar
- de-u**-co-phonebk-ka-shifted**
    - German using Phonebook sorting, ignore punct.
- ar-u**-nu-native**
    - Arabic with native digits (٠١٢٣٤…)
- ar-u**-nu-latn**
    - Arabic with Western digits (01234…)

Extension T - Transforms

- RFC6497
- General
  - Transliterations, transcriptions, translations, etc.
  - For unstructured interchange, only locale ID avail.
- Examples
  - ja-t-it
  - ja-Kana-t-it
  - und-Latn-t-und-cyrl

Extension T - Specialized

- m0 - Mechanisms *(typically authorities)*
  - und-Latn-t-ru**-m0-ungegn-2007**
- i0 - Input Method Transformation
  - zh-t**-i0-pinyin**
- k0 - Keyboard Transformation
  - en-t**-k0-dvorak**
- t0 - Machine Translation
  - ja-t-de**-t0-und**
- x0 - Private Use
  - ja-t-de-t0-und**-x0-medical**

Resources

- Choosing a language tag
  - http://w3.org/International/questions/qa-choosing-language-tags.en
  - http://cldr.unicode.org/index/cldr-spec/picking-the-right-language-code
- Extension fields/subfields
  - Last Release:
    - http://unicode.org/repos/cldr/tags/release-21-0-2/common/bcp47/
  - Latest snapshot:
    - http://unicode.org/repos/cldr/trunk/common/bcp47/
  - Requesting registrations:
    - http://tools.ietf.org/html/rfc6497#section-2.6
    - http://unicode.org/cldr/trac/newticket

Discussion

Background slides

Unicode Locale/Lang ID (2)

- UTS #35 *Unicode Locale Data Markup Language* (LDML)

- Based on BCP 47 + RFC 6067 + language-subtag-registry

- Some restrictions & extensions
  - Both '_' and '-' as separators
  - No extlang, no irregular (grandfathered) tags
    - Uses “zh” for compatibility, not “cmn”, etc.
  - Private use codes defined
    - “ZZ” for Unknown Region