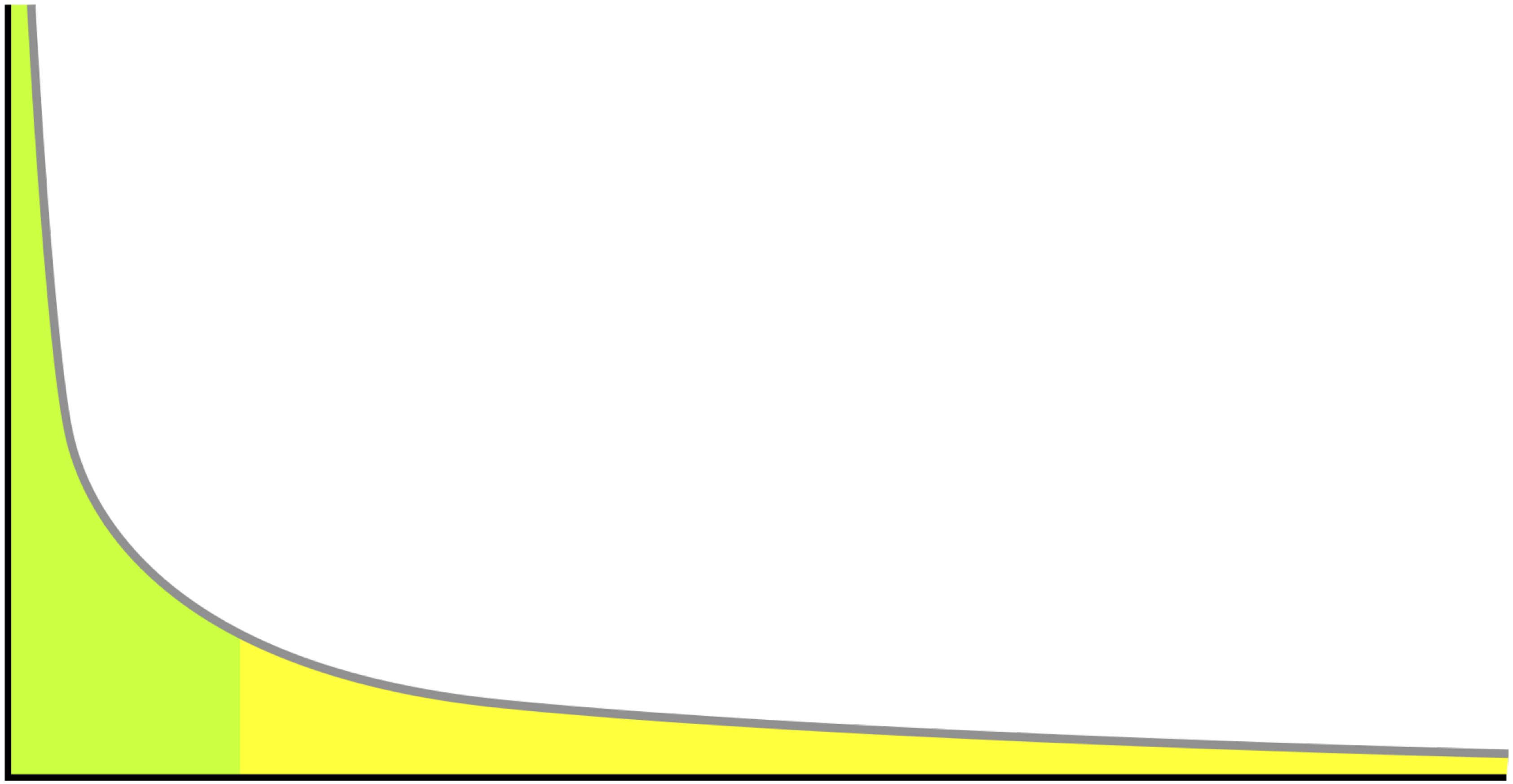
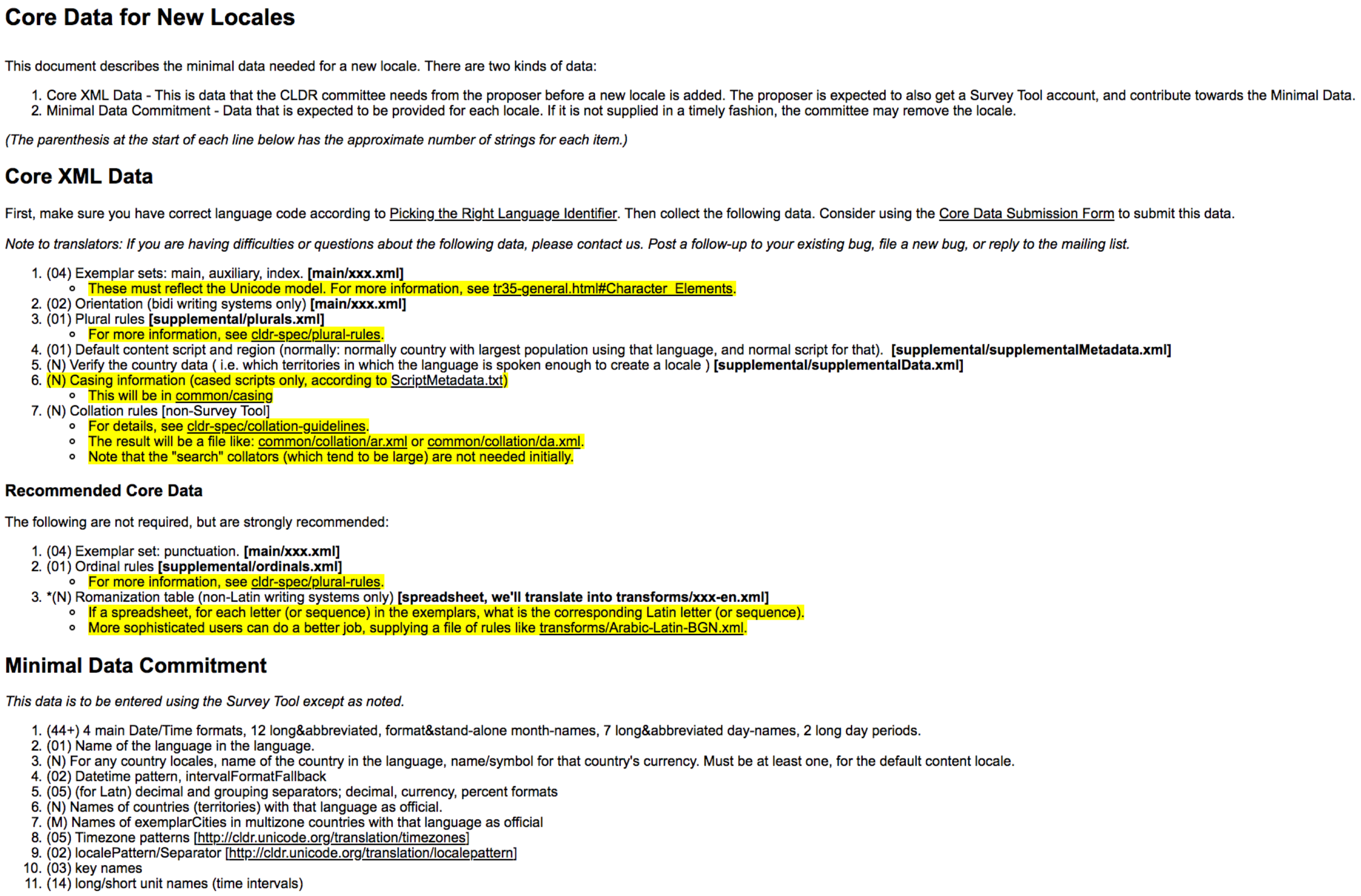
Revitalizing CLDR for the Long Tail



http://cldr.unicode.org/index/cldr-spec/minimaldata



Core Data for New Locales

This document describes the minimal data needed for a new locale. There are two kinds of data:

1. Core XML Data - This is data that the CLDR committee needs from the proposer before a new locale is added. The proposer is expected to also get a Survey Tool account, and contribute towards the Minimal Data.
2. Minimal Data Commitment - Data that is expected to be provided for each locale. If it is not supplied in a timely fashion, the committee may remove the locale.

(The parenthesis at the start of each line below has the approximate number of strings for each item.)

Core XML Data

First, make sure you have correct language code according to Picking the Right Language Identifier. Then collect the following data. Consider using the Core Data Submission Form to submit this data.

Note to translators: If you are having difficulties or questions about the following data, please contact us. Post a follow-up to your existing bug, file a new bug, or reply to the mailing list.

1. (04) Exemplar sets: main, auxiliary, index. **[main/xxx.xml]**
   - These must reflect the Unicode model. For more information, see tr35-general.html#Character_Elements.
2. (02) Orientation (bidi writing systems only) **[main/xxx.xml]**
3. (01) Plural rules **[supplemental/plurals.xml]**
   - For more information, see cldr-spec/plural-rules.
4. (01) Default content script and region (normally: normally country with largest population using that language, and normal script for that). **[supplemental/supplementalMetadata.xml]**
5. (N) Verify the country data (i.e. which territories in which the language is spoken enough to create a locale) **[supplemental/supplementalData.xml]**
6. (N) Casing information (cased scripts only, according to ScriptMetadata.txt)
   - This will be in common/casing
7. (N) Collation rules [non-Survey Tool]
   - For details, see cldr-spec/collation-guidelines.
   - The result will be a file like: common/collation/ar.xml or common/collation/da.xml.
   - Note that the "search" collators (which tend to be large) are not needed initially.

Recommended Core Data

The following are not required, but are strongly recommended:

1. (04) Exemplar set: punctuation. **[main/xxx.xml]**
2. (01) Ordinal rules **[supplemental/ordinals.xml]**
   - For more information, see cldr-spec/plural-rules.
3. *(N) Romanization table (non-Latin writing systems only) **[spreadsheet, we'll translate into transforms/xxx-en.xml]**
   - If a spreadsheet, for each letter (or sequence) in the exemplars, what is the corresponding Latin letter (or sequence).
   - More sophisticated users can do a better job, supplying a file of rules like transforms/Arabic-Latin-BGN.xml.

Minimal Data Commitment

This data is to be entered using the Survey Tool except as noted.

1. (44+) 4 main Date/Time formats, 12 long&abbreviated, format&stand-alone month-names, 7 long&abbreviated day-names, 2 long day periods.
2. (01) Name of the language in the language.
3. (N) For any country locales, name of the country in the language, name/symbol for that country's currency. Must be at least one, for the default content locale.
4. (02) Datetime pattern, intervalFormatFallback
5. (05) (for Latn) decimal and grouping separators; decimal, currency, percent formats
6. (N) Names of countries (territories) with that language as official.
7. (M) Names of exemplarCities in multizone countries with that language as official
8. (05) Timezone patterns [http://cldr.unicode.org/translation/timezones]
9. (02) localePattern/Separator [http://cldr.unicode.org/translation/localepattern]
10. (03) key names
11. (14) long/short unit names (time intervals)

The First Seven Steps to ’Core XML Data’

- Exemplar characters (in proposal)
- LTR, RTL, other direction (in proposal)
- Plural rules
- Script/country data (in proposal)
- More country data (in proposal)
- Casing (in proposal)
- Collation (in proposal)

The ‘Plurals’ Long Straw

Only a Few Translations Away

“Standard Time”

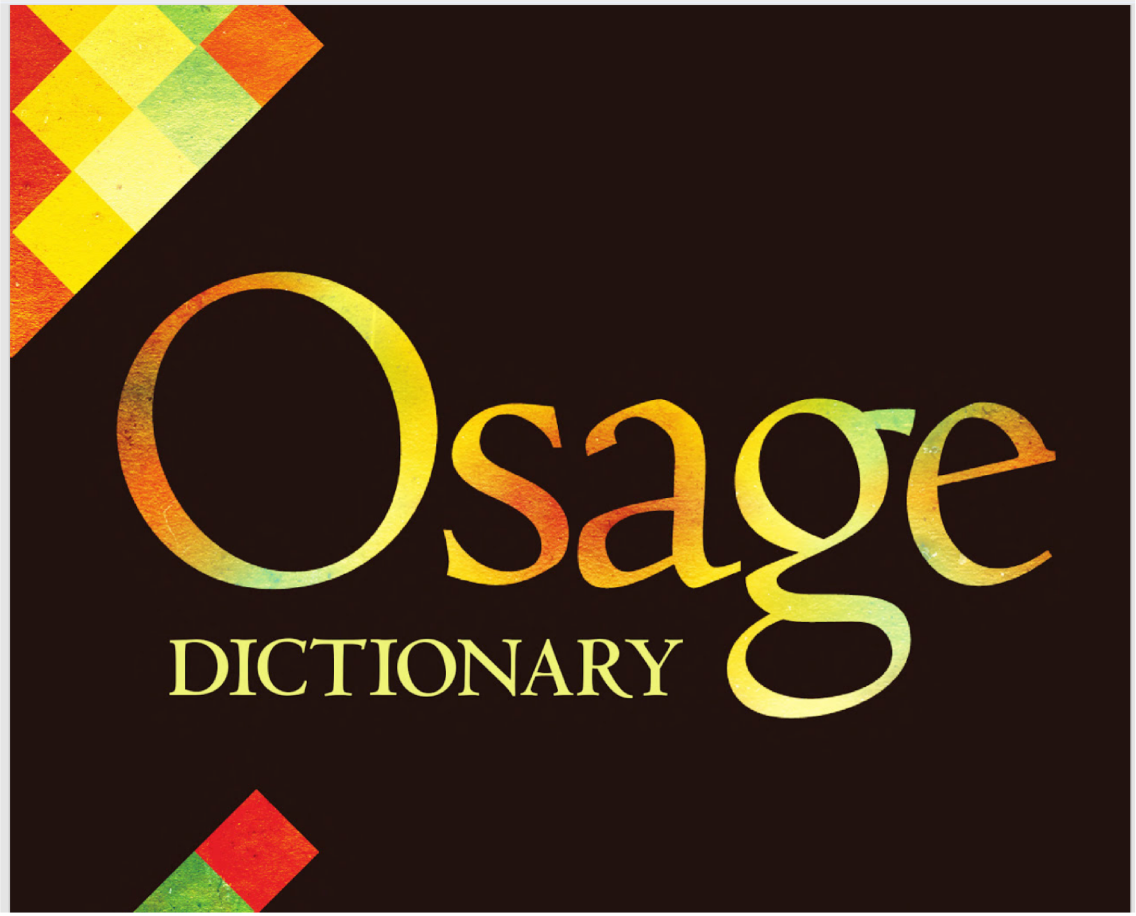
“Daylight Time”

“Gregorian Calendar”

“Standard Sort Order”

“Western Digits”

Traditional Sources

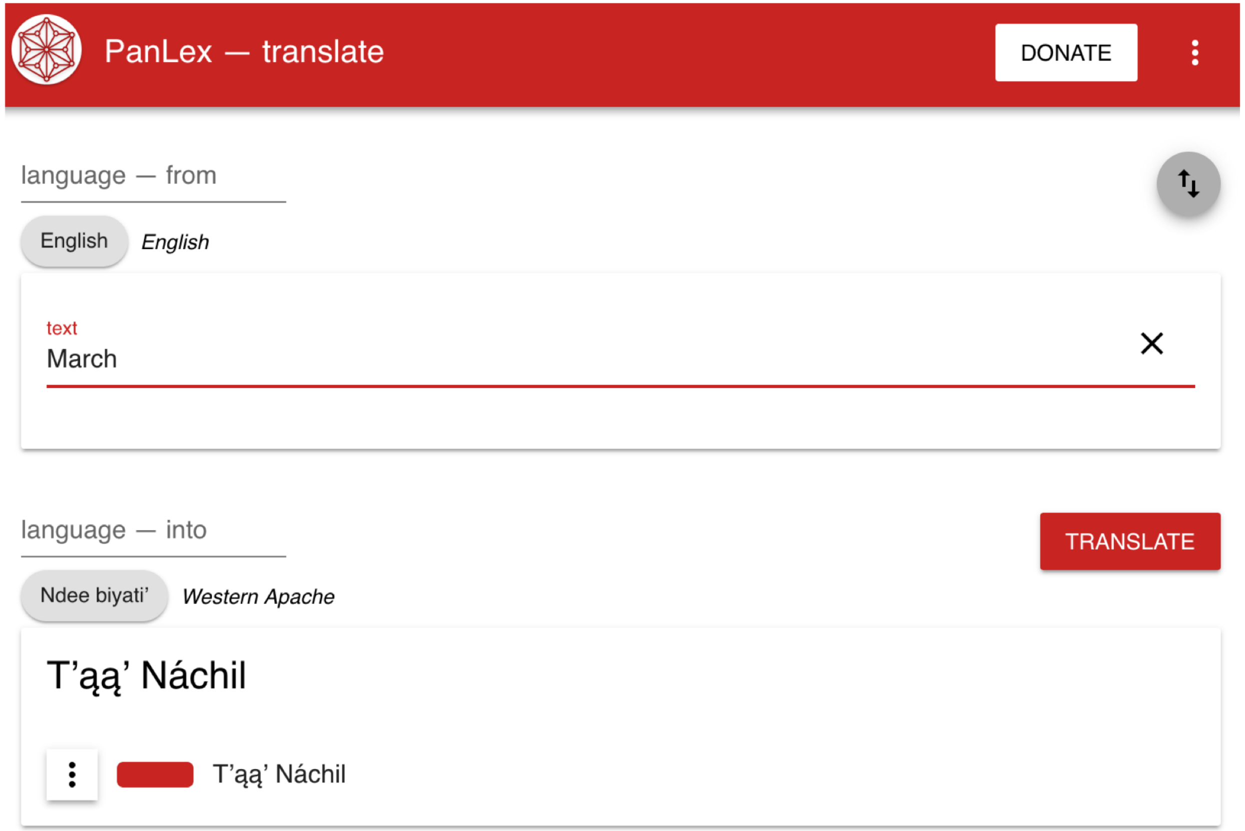


The Easier Remainder

- 50 other easy-to-find translations
  - (1) Name of language in language - easy
  - (38) Some 12 month names and 7 day names require translations
    - Relatively easy to find translation in dictionaries or other good references.
  - (7) for day, week, month, year, hour, minute, second

- With caveats
  - Some translation may need a little more sophistication depending on use
  - Abbreviations may not be a concept in a given language

Automated Sources?



PanLex — translate

DONATE

language — from

English *English*

text

March

language — into

TRANSLATE

Ndee biyati’ *Western Apache*

T’ąą’ Náchil

T’ąą’ Náchil

```
curl http://api.panlex.org/v2/expr -d '{ "uid": "eng-000", "txt": "January" }'

curl http://api.panlex.org/v2/expr -d '{ "uid": "apw-000", "trans_expr": 452754
}
```

```
January   = “Dził Bilátahgai"

February  = ”Múhshchii’"

March     = "T’ąą’ Náchil"

April     = "T’ąą’ Náchoh"

May       = "Itsáh Hashkēē"

June      = "Itsáh Bizhaazh"

July      = "Nii’ Dichíhé"

August    = "Binest’ánts’ǭǭsé"

September = "Binest’ánchoh"

October   = "Gha˛a˛zhį’"

November  = "Kǫ’ Bąąh Náłk’as"

December  = "Zas Nłt’ees"
```