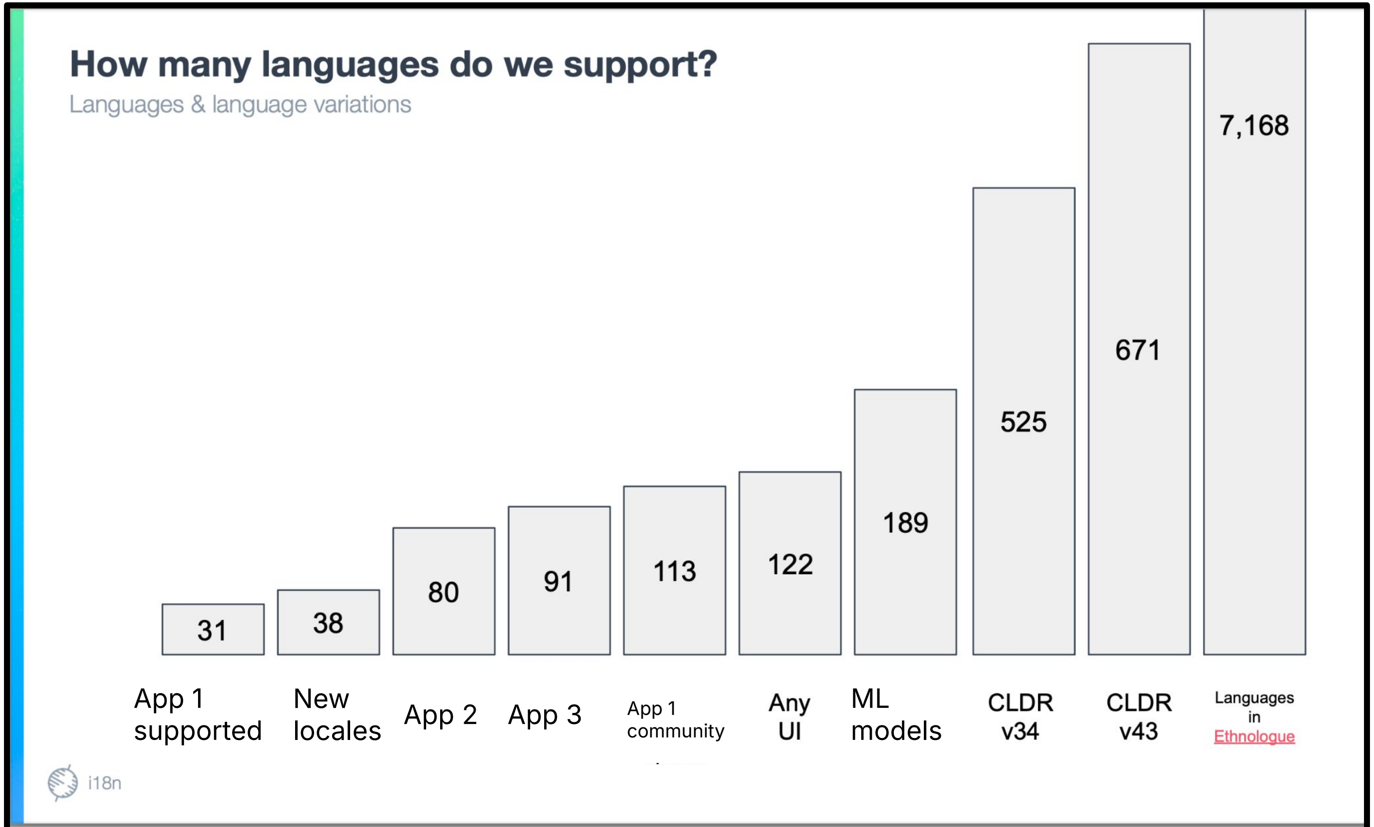
How many locales?

How many population data records & in the future how many XMLs, how many modern languages?

Conrad Nied

2024-10-23

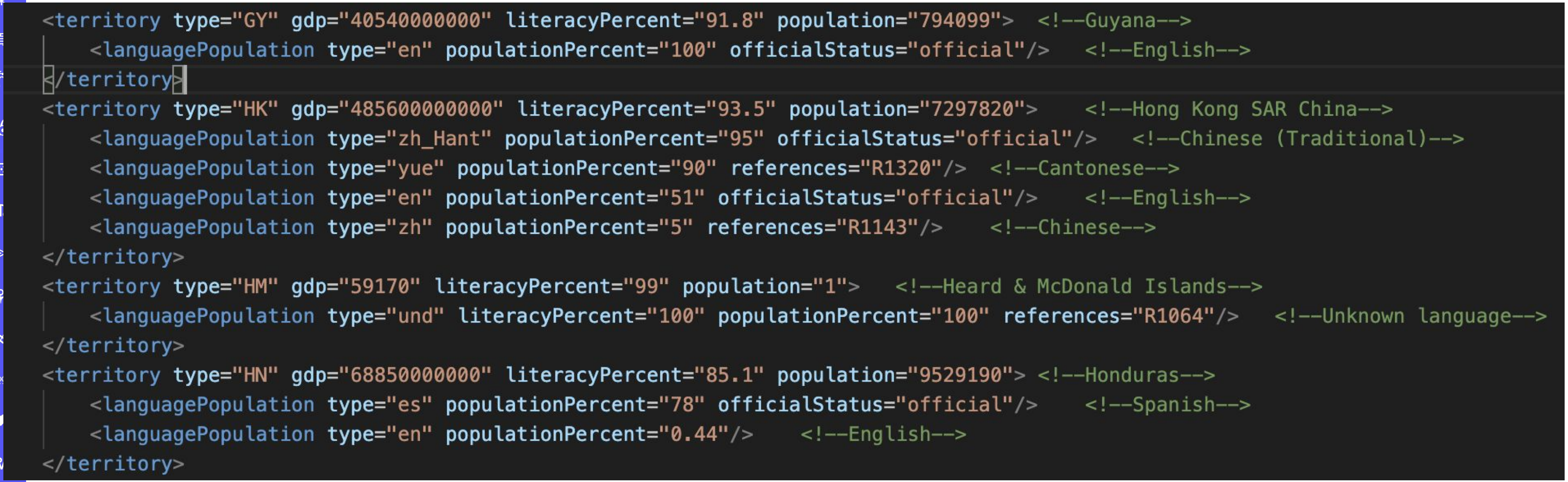
Origin of the question



How many languages do we support?

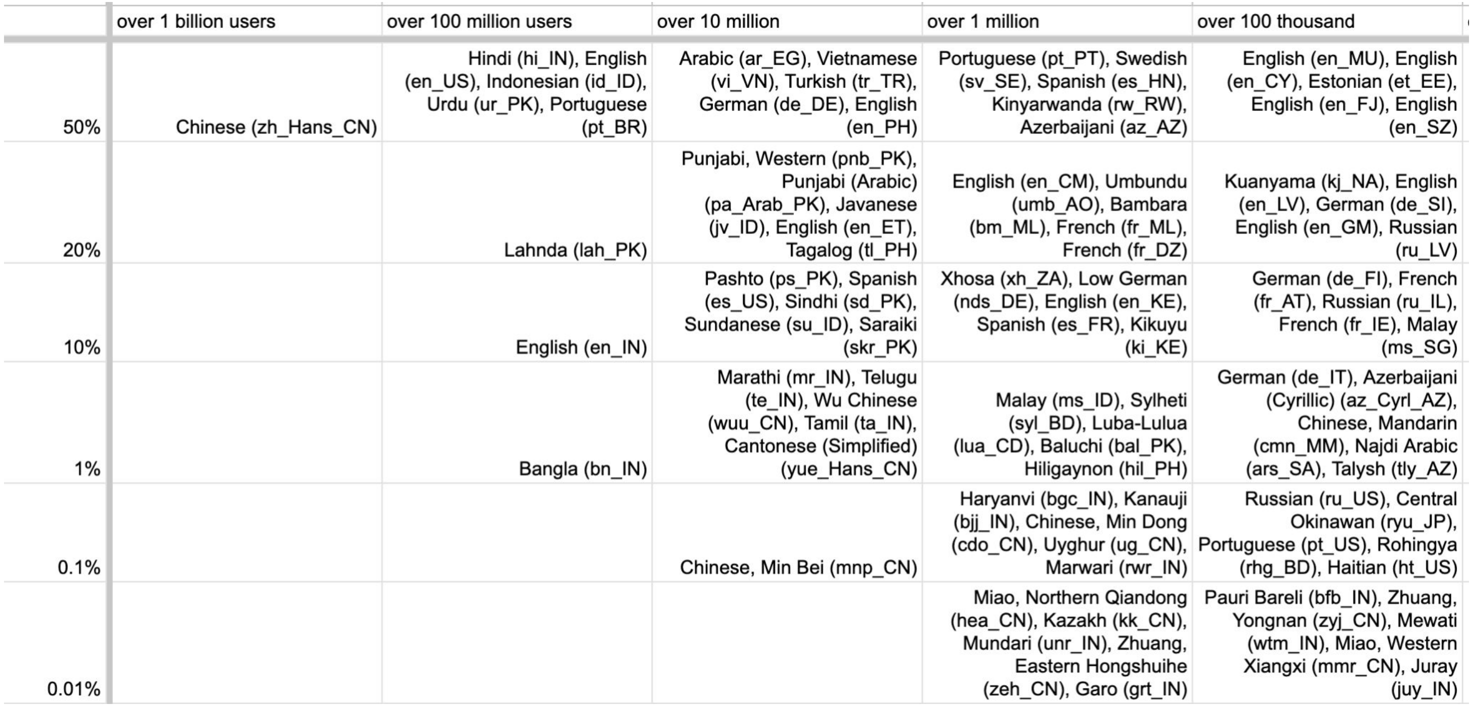
Languages & language variations

supplementalData.xml territoryInfo



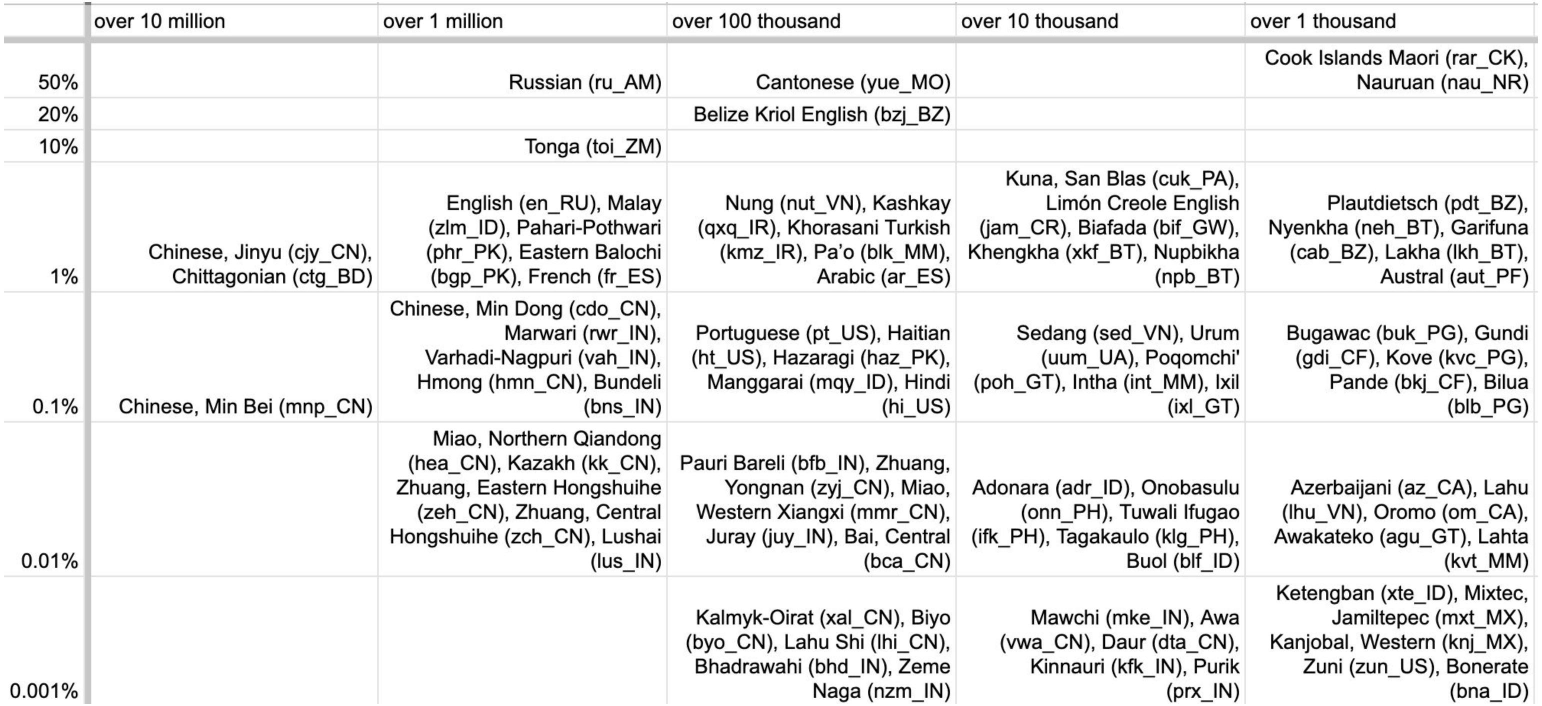
```
<territory type="GY" gdp="40540000000" literacyPercent="91.8" population="794099">  <!--Guyana-->
    <languagePopulation type="en" populationPercent="100" officialStatus="official"/>   <!--English-->
</territory>
<territory type="HK" gdp="485600000000" literacyPercent="93.5" population="7297820">    <!--Hong Kong SAR China-->
    <languagePopulation type="zh_Hant" populationPercent="95" officialStatus="official"/>   <!--Chinese (Traditional)-->
    <languagePopulation type="yue" populationPercent="90" references="R1320"/>  <!--Cantonese-->
    <languagePopulation type="en" populationPercent="51" officialStatus="official"/>    <!--English-->
    <languagePopulation type="zh" populationPercent="5" references="R1143"/>    <!--Chinese-->
</territory>
<territory type="HM" gdp="59170" literacyPercent="99" population="1">   <!--Heard & McDonald Islands-->
    <languagePopulation type="und" literacyPercent="100" populationPercent="100" references="R1064"/>   <!--Unknown language-->
</territory>
<territory type="HN" gdp="68850000000" literacyPercent="85.1" population="9529190"> <!--Honduras-->
    <languagePopulation type="es" populationPercent="78" officialStatus="official"/>    <!--Spanish-->
    <languagePopulation type="en" populationPercent="0.44"/>    <!--English-->
</territory>
```

Top locales by % in country and absolute number of users



	over 1 billion users	over 100 million users	over 10 million	over 1 million	over 100 thousand
50%	Chinese (zh_Hans_CN)	Hindi (hi_IN), English (en_US), Indonesian (id_ID), Urdu (ur_PK), Portuguese (pt_BR)	Arabic (ar_EG), Vietnamese (vi_VN), Turkish (tr_TR), German (de_DE), English (en_PH)	Portuguese (pt_PT), Swedish (sv_SE), Spanish (es_HN), Kinyarwanda (rw_RW), Azerbaijani (az_AZ)	English (en_MU), English (en_CY), Estonian (et_EE), English (en_FJ), English (en_SZ)
20%		Lahnda (lah_PK)	Punjabi, Western (pnb_PK), Punjabi (Arabic) (pa_Arab_PK), Javanese (jv_ID), English (en_ET), Tagalog (tl_PH)	English (en_CM), Umbundu (umb_AO), Bambara (bm_ML), French (fr_ML), French (fr_DZ)	Kuanyama (kj_NA), English (en_LV), German (de_SI), English (en_GM), Russian (ru_LV)
10%		English (en_IN)	Pashto (ps_PK), Spanish (es_US), Sindhi (sd_PK), Sundanese (su_ID), Saraiki (skr_PK)	Xhosa (xh_ZA), Low German (nds_DE), English (en_KE), Spanish (es_FR), Kikuyu (ki_KE)	German (de_FI), French (fr_AT), Russian (ru_IL), French (fr_IE), Malay (ms_SG)
1%		Bangla (bn_IN)	Marathi (mr_IN), Telugu (te_IN), Wu Chinese (wuu_CN), Tamil (ta_IN), Cantonese (Simplified) (yue_Hans_CN)	Malay (ms_ID), Sylheti (syl_BD), Luba-Lulua (lua_CD), Baluchi (bal_PK), Hiligaynon (hil_PH)	German (de_IT), Azerbaijani (Cyrillic) (az_Cyrl_AZ), Chinese, Mandarin (cmn_MM), Najdi Arabic (ars_SA), Talysh (tly_AZ)
0.1%			Chinese, Min Bei (mnp_CN)	Haryanvi (bgc_IN), Kanauji (bjj_IN), Chinese, Min Dong (cdo_CN), Uyghur (ug_CN), Marwari (rwr_IN)	Russian (ru_US), Central Okinawan (ryu_JP), Portuguese (pt_US), Rohingya (rhg_BD), Haitian (ht_US)
0.01%				Miao, Northern Qiandong (hea_CN), Kazakh (kk_CN), Mundari (unr_IN), Zhuang, Eastern Hongshuihe (zeh_CN), Garo (grt_IN)	Pauri Bareli (bfb_IN), Zhuang, Yongnan (zyj_CN), Mewati (wtm_IN), Miao, Western Xiangxi (mmr_CN), Juray (juy_IN)

Top locales not yet in CLDR



	over 10 million	over 1 million	over 100 thousand	over 10 thousand	over 1 thousand
50%		Russian (ru_AM)	Cantonese (yue_MO)		Cook Islands Maori (rar_CK), Nauruan (nau_NR)
20%			Belize Kriol English (bzj_BZ)		
10%		Tonga (toi_ZM)			
1%	Chinese, Jinyu (cjy_CN), Chittagonian (ctg_BD)	English (en_RU), Malay (zlm_ID), Pahari-Pothwari (phr_PK), Eastern Balochi (bgp_PK), French (fr_ES)	Nung (nut_VN), Kashkay (qxq_IR), Khorasani Turkish (kmz_IR), Pa’o (blk_MM), Arabic (ar_ES)	Kuna, San Blas (cuk_PA), Limón Creole English (jam_CR), Biafada (bif_GW), Khengkha (xkf_BT), Nupbikha (npb_BT)	Plautdietsch (pdt_BZ), Nyenkha (neh_BT), Garifuna (cab_BZ), Lakha (lkh_BT), Austral (aut_PF)
0.1%	Chinese, Min Bei (mnp_CN)	Chinese, Min Dong (cdo_CN), Marwari (rwr_IN), Varhadi-Nagpuri (vah_IN), Hmong (hmn_CN), Bundeli (bns_IN)	Portuguese (pt_US), Haitian (ht_US), Hazaragi (haz_PK), Manggarai (mqy_ID), Hindi (hi_US)	Sedang (sed_VN), Urum (uum_UA), Poqomchi' (poh_GT), Intha (int_MM), Ixil (ixl_GT)	Bugawac (buk_PG), Gundi (gdi_CF), Kove (kvc_PG), Pande (bkj_CF), Bilua (blb_PG)
0.01%		Miao, Northern Qiandong (hea_CN), Kazakh (kk_CN), Zhuang, Eastern Hongshuihe (zeh_CN), Zhuang, Central Hongshuihe (zch_CN), Lushai (lus_IN)	Pauri Bareli (bfb_IN), Zhuang, Yongnan (zyj_CN), Miao, Western Xiangxi (mmr_CN), Juray (juy_IN), Bai, Central (bca_CN)	Adonara (adr_ID), Onobasulu (onn_PH), Tuwali Ifugao (ifk_PH), Tagakaulo (klg_PH), Buol (blf_ID)	Azerbaijani (az_CA), Lahu (lhu_VN), Oromo (om_CA), Awakateko (agu_GT), Lahta (kvt_MM)
0.001%			Kalmyk-Oirat (xal_CN), Biyo (byo_CN), Lahu Shi (lhi_CN), Bhadrawahi (bhd_IN), Zeme Naga (nzm_IN)	Mawchi (mke_IN), Awa (vwa_CN), Daur (dta_CN), Kinnauri (kfk_IN), Purik (prx_IN)	Ketengban (xte_ID), Mixtec, Jamiltepec (mxt_MX), Kanjobal, Western (knj_MX), Zuni (zun_US), Bonerate (bna_ID)

Proposed Criteria:

1) Never remove a locale (1502 locales)
2) Accept all volunteers, if they will provide the data add their locale.
   a) Only hold off on minors variations of existing languages
3) Official or Recognized, Nationally or Regionally (+140 locales)
4) Indigenous languages (+4047 locales)

a)	>100,000	398 languages
b)	10,000 to 99,999	777 languages
c)	10 to 9,999	2409 languages
d)	Dying or extinct	646 languages

5) Large Presence in a Country (+85 locales)
6) Add otherwise found in area (+715 locales)

How many locales to include

Criteria	# of Locales in Group	Cumulative # of Locales	10 Largest Locales
1: Keep existing languages	1502	1502	Chinese (zh_Hans_CN), Hindi (hi_IN), English (en_US), English (en_IN), Indonesian (id_ID), Urdu (ur_PK), Portuguese (pt_BR), Bangla (bn_BD), Russian (ru_RU), Japanese (ja_JP)
3a: Add Official or Recognized	139	1641	Mongolian, Peripheral (mvf_CN), Hunsrik (hrx_BR), Dogri (dgo_IN), Chinese (zh_US), Russian (ru_AM), Arabic (ar_US), Tonga (toi_ZM), Kazakh (kk_CN), Bajjika (vjk_NP), Lushai (lus_IN)
3b: Add Indigenous >=10,000	1028	2669	Chinese, Jinyu (cjy_CN), Chittagonian (ctg_BD), Chinese, Min Bei (mnp_CN), Chinese, Min Dong (cdo_CN), Marwari (rwr_IN), Varhadi-Nagpuri (vah_IN), Hmong (hmn_CN), Bundeli (bns_IN), Malvi (mup_IN), Malay (zlm_ID)
4: Add otherwise Indigenous	3009	5678	Wanukaka (wnk_ID), Mamboru (mvd_ID), Western Pantar (lev_ID), Spiti Bhoti (spt_PK), Nda'nda' (nnz_CM), Baima (bqh_CN), Muzi (ymz_CN), Siwi (siz_EG), Palu'e (ple_ID), So'a (ssq_ID)
5: Add Large Presence in Country, >=100,000	85	5763	English (en_RU), French (fr_ES), Sylheti (syl_IN), Arabic (ar_IN), English (en_AF), Bengali (bn_PK), Indian Sign Language (ins_IN), Hmong (hmn_VN), Khmer (km_VN), Urdu (ur_AF)
6: Add otherwise found in area	715	6478	Burmese (my_US), Jamaican Creole English (jam_US), Purik (prx_IN), Pushto (ps_US), Iloko (ilo_US), Kirghiz (ky_RU), Kannada (kn_US), Serbian (sr_US), Hebrew (he_CA), Georgian (ka_RU)