Study to Evaluate Available Solutions for the Submission and Display of Internationalized Contact Data

Activities Update

ICANN Singapore Meeting

March 2014

#ICANN49

Key areas of study

1. Document the submission and display practices of internationalized registration data
2. Assess availability and cost of open source and commercial solutions for transliterating and translating contact data
3. Evaluate the accuracy implications for transliteration and translations of contact data

Methodology

- Survey existing practices for collecting contact data in local languages and scripts
- Study translation and transliteration requirements and methods for languages and scripts (registrars, registries, e-merchants)
- Identify tools for transliteration and translation and evaluate their availability, cost and accuracy

Terminology 1/3

- **Toponym** or **Place Name** is a proper noun for geographical names
- **Exonym** is the name used in a language for a geographical feature situated outside the area where that language has official status, e.g. Londres; *UN recommends minimizing exonyms in international usage* (vs. **Endonym**: Beijing vs. Peking)
- **Allonym** or **Alternate Name** or **Variant Name** : Johannesburg and Egoli
- **Generic Term** is a common noun for topographic feature in terms of its characteristics, e.g. mountain, sierra, wadi, river, mer

Source: *Glossary of Terms for the Standardization of Geographical Names.* United Nations Group of Experts on Geographical Names, Department of Economic and Social Affairs, Statistics Division, United Nations. 2002.

Terminology 2/3

- **Name Transformation** covers the translation and **conversion** (transcription and transliteration) of toponyms
- **Translation** is process of expressing meaning, presented in a source language, in the words of a target language or result thereof; Examples: Mer Noire (Fr) = Čornoje More (Ru); Mount Fuji (En) = Fuji San (Jp)
- **Transcription** is a method of phonetic names conversion between different languages, or result thereof; Examples: Ankara (Tur) Αγκαρα (Gr); جبلية (Ar) Djabaliya (Fr). *Transcription is not normally reversible.* **Retranscription** might result in a form differing from the original, for example Agkara, دجبلية

Terminology 3/3

- **Transliteration** is a method of names conversion between different alphabetic scripts and syllabic scripts or a result thereof; *distinct from transcription, it aims at (but does not necessarily achieve) complete reversibility,* and must be accompanied by a transliteration key. The reverse process is called **Retransliteration**. Example:القاهرة al-Qāhirah (Cairo)
- **Reversibility** permits a transliterated item to be reconverted back into the source script, the result being identical with the original
- **Romanization** is conversion from non-Roman into Roman script. Examples: Αθήνα Athina; Москва Moskva; بيروت Bayrūt; תל-אביב Tel-Aviv; ニホン Nihon

Levels of Transformation

- Requiring ***accurate transformation*** (e.g. valid in a court of law, matching information in a passport, matching information in legal incorporation, etc.)
- Requiring ***consistent transformation*** (allowing matching, e.g. to match address of a registrant on a Google map, etc.)
- Requiring ***ad hoc transformation*** (allowing informal or casual version of the information in another language)

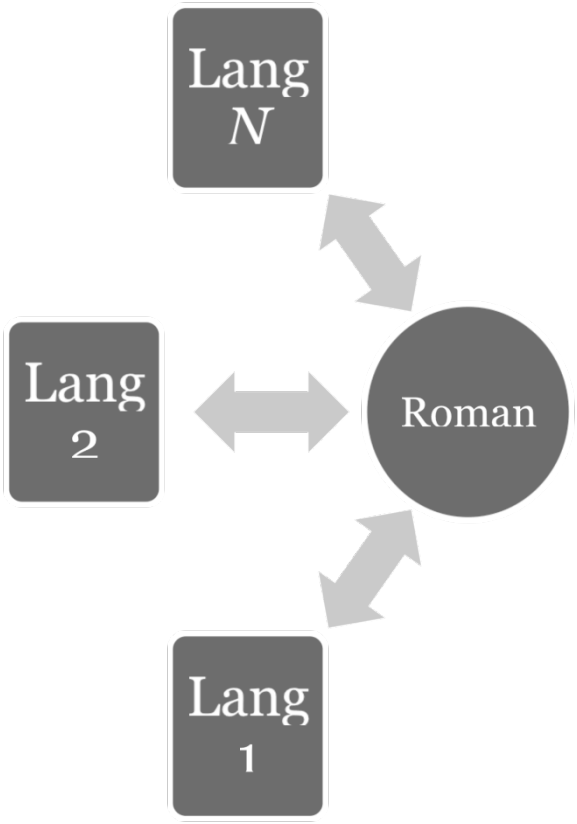
Levels of Transformation

- **Accurate transformation**
  - Translation + Transcription + Transliteration
  - Manual
    - 金人庆 Jin Renqing (China)
    - 金大中 Kim Dae-jung (Korea)
  - Mohammad, Mohammed, Muhammad, :محمد ... ,Mohamed
- **Consistent transformation**
  - Transliteration
  - Automatic; script specific challenges, e.g. Readability for Arabic script: Mhmd
- **Ad hoc transformation**

Int'l Standards Organizations

UNGEGN	United National Group of Experts on Geographical Names
ISO	International Organization for Standardization
UPU	Universal Postal Union
Unicode	Unicode Consortium

Pivoting for Transliteration from *All* Languages to *All* Languages



"The Roman script (also referred to as Latin script) has been adopted as a base for international use by the United Nations, and the Group of Experts strongly recommends the development of a single romanization (that is to say, transliteration) system for each non-Roman script"

"Non-Roman scripts can then be converted via their romanization into other scripts for national and international use"

For consistency, this requires the transliteration into Latin script to be reversible

Fall Back for Missing Languages

- "Progressively… with priorities being the target, source, and variant, in that order" (Unicode)
  - Russian-English/UNGEGN
  - Russian-English [/alternate option]
  - Cyrillic-English/UNGEGN
  - Cyrillic-English [/alternate option]
  - Russian-Latin/UNGEGN
  - Russian-Latin[/alternate option]
  - Cyrillic-Latin/UNGEGN
  - Cyrillic-Latin[/alternate option]

Submission and Display Practices

- Survey a limited no. of registries and registrars covering multiple scripts and geographic regions
  - How is data collected from the registrant in local language/scripts
  - Is data maintained in more than one language/script
  - Is there any translation/transliteration? What is the role of registrant in the process
  - Is there enhancements to tools used in practice (e.g. for EPP and WHOIS services)
  - How is contact data displayed in local language/script

Submission and Display Practices Survey

- Creation of survey [DONE]
- Pilot test [DONE]
- Survey Administration
  - Started in Mid Feb, 9 responses from Registries, 1 response from registrars
  - Finalize in Mid April

Evaluation of Transformation Tools

- Limited to tools which cover a breadth of languages and not limited to transformation between single language pair
- Number of languages and scripts covered
- Standards used for the transformations
- Accuracy for representative language pairs
- Licensing information (open source or proprietary)
- Reversibility of such transformations

Possible Transliteration Tools

- Global Name Recognition (GNR)
- International Components for Unicode (ICU)
- Rosette Name Translator
- Microsoft Transliteration Utility
- Google Translate API
- Microsoft Translator
- Address Doctor
- Text::Unidecode
- Junidecode
- Xlit


Thanks
#ICANN49
ICANN
Singapore