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

ICANN IRD Study Team
ird@viagenie.ca

Motivation and Scope

- Multiple ICANN WGs looking at internationalized registration data (IRD) requirements
- This study documents current practices and transformation possibilities to inform them
  1. Look into practices of handling IRD
     1. Electronic merchants and online services
     2. Registries and registrars in geographies using local languages
     3. Protocols on submission, storage, transmission and display
  2. Assess accuracy of transforming IRD

Internationalized Registrations Data

- Two categories of data
  - Contact Data (Registrant, Registrar)
  - Transactional Data (Automatic)
- Contact Data subset in local language:
  - Person and Organizational Name
  - Address
  - City and State
  - Country

Some Definitions

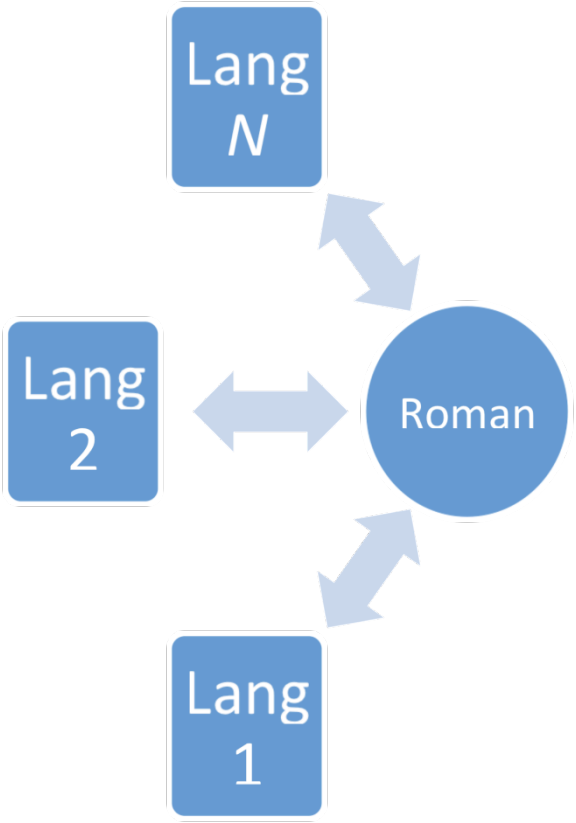
- **Translation:** expressing meaning, presented in a source language, in the words of a target language
  - Casablanca (Spanish for Arabic Dār al-Bay.dā'); Lake Como (English for Italian Lago di Como)
- **Transcription:** a method of phonetic names conversion between different languages
  - Turkish Ankara Greek Αγκαρα; Russian Щукино English Shchukino; Arabic جـ French Djabaliya
  - Transcription is not normally a reversible process
  - Pinyin romanization of Chinese is regarded as transcription
- **Transliteration:** a method of names conversion between different scripts, in which each character of the source script is represented in the target script
  - must be accompanied by a transliteration key.

  – al-Qāhirah (Cairo); Владивосток Vladivostok القاهرة ;חיפה Hefa)
- Generically referred to as **Transformation**

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)

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

Survey of E-Merchants

Name	Country	Script	Language
Amazon	USA/ Global	All	All
Alibaba	China/ Global	All	All
Rakuten	Japan	Kanji, Hiragana, Katakana	Japanese
Homeshop18	India	Local Various	Local Various
LDLC	France	Latin	French
eMall	Saudi Arabia	Arabic	Arabic

Survey Results

- Websites allow data in local languages
- Verify the contact data only to a limited extent
- Just accept the user input, putting the onus of verification of addresses on the user
- Even active in markets where they do not support the dominant script or language used

Survey of Registries and Registrars

- Separate surveys for registries and registrars
- The registry survey responded by twelve registries
  - large gTLDs and ccTLDs
  - covering multiple languages and scripts, such as Arabic, Han, Cyrillic, Japanese, German, French and English
- The registrar survey has been responded by two registrars in the time frame of the study
  - one is a very large registrar
  - conclusions should not be generalized, but may still provide insights

Survey Results

- Collect information in local languages
- Sometimes in both local language and its romanized form
  - romanized form is required by the Registrar Accreditation Agreement (RAA) even for IDN registrations
  - Consistency between the two versions is not verified
- None transform the contact information
  - where multiple language data is collected, provided directly by the registrant
- Support of IRD is variable across the processes and systems

Survey of Relevant Protocols

- WHOIS only supports ASCII
- EPP supports UTF-8 encoding for transmitting and receiving data, without language specification
- EPP does not record multiple linguistic versions of the same data
- RDAP can encode language information and can handle multiple versions in parallel
- These protocols do not record the method and history of (any) transformation(s) data may have undergone to get to its current form

Transformation

- Data
  - **Individual or Entity names**, including family and given names, organization names, etc.
  - **Addresses**, including proper names, generic terms (which should not be transformed), abbreviations (where applicable), punctuation, digits, etc.
  - **City** and state/province names
  - **Country** names, including full and short forms
- Scripts (and Languages)
  - **Han** (**Chinese** using Traditional and Simplified Chinese writing)
  - **Devanagari** (**Hindi**, Marathi)
  - **Arabic** (**Arabic**, Persian, and Urdu)
  - **Cyrillic** (Bulgarian, **Russian** and Ukrainian)

Transformation Testing Data Details

Script	Type	No. of Items	No. of Words	No. of Characters	Notes
Han	Name	12	27	136	Data covers Chinese language (both Traditional and Simplified)
	Address	12	129	818	
	City /State	5	5	38	
	Country	5	11	65	
Devanagari	Name	22	22	180	Data covers (mostly) Hindi and Marathi languages
	Address	12	73	430	
	City /State	26	37	295	
	Country	-	-	-	
Arabic	Name	20	20	115	Data covers (mostly) Arabic, Urdu and Persian languages
	Address	15	49	320	
	City /State	10	13	77	
	Country	10	14	100	
Cyrillic	Name	20	21	150	Data covers (mostly) Russian, Ukrainian and Bulgarian languages
	Address	14	30	216	
	City /State	11	19	174	
	Country	10	10	67	

Measures

- Accuracy - binary
  - exact match between transformed and manual transformation
  - best = 100%
- Levenshtein Distance – non-binary
  - the number of edits (insertion, deletion and substitution) between two strings
  - For Cyrillic Russian Вельов, is “Velyov” but get “Viel'ov”, distance = 2 (delete i and substitute ‘ with y)
  - exactly same strings = zero edits; maximum distance = length of the longer string
  - Best = 0%

Tools

- Some general translation tools
  - Ace Translator (http://www.acetools.biz/)
  - Babylon (http://translation.babylon.com/)
  - Google Translate (https://translate.google.com/)
  - Microsoft Translate (http://www.microsoft.com/en-us/translator/)
  - Power Translator (https://www.lec.com/power-translator-software.asp)
  - Systrans (http://www.systransoft.com/)
  - Translution (http://www.translution.com/default.asp)

Tools

- Some general transliteration or transcription tools
  - Google Input Tools (http://www.google.com/inputtools/)
  - IBM ICU Transliteration (http://demo.icu-project.org/icu-bin/translit; also see http://userguide.icu-project.org/transforms/general)
  - JUnidecode (http://www.ippatsuman.com/projects/junidecode/index.html)
  - Microsoft Transliteration Utility (http://msdn.microsoft.com/en-us/goglobal/bb688104.aspx)
  - Ok-board.com (http://ok-board.com/)
  - Unidecode (https://pypi.python.org/pypi/Unidecode)
  - Yahoo Transliterate (https://transliteration.yahoo.com/)

Tools

- Some specialized for transformation of various parts of contact information
  - Address Doctor (http://www.addressdoctor.com/en/)
  - Basis Technology Rosette Name Translator (http://www.basistech.com/text-analytics/rosette/name-translator/)
  - Experian Data Quality (http://www.qas.com/contact-data-quality.htm)
  - IBM Global Name Recognition (http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=GPA&htmlfid=897/ENUS207-295)
  - Loqate (http://www.loqate.com/technology/transliteration/)
  - Trillium Software (http://www.trilliumsoftware.com/products/data-types/customer-data/)

Summary of Transformation Results

Levenshtein Distance (%) across Tools	Han	Devanagari	Arabic	Cyrillic	Average
Name	26	28	38	25	29
Address	61	51	60	49	55
City/State	19	30	27	45	30
Country	63	-	72	33	56
Average	42	36	49	38	

Summary of Transformation Results

Over all contact information across all languages	Type	% Overall Accuracy	Average of % Accuracy	% Over all Lev. Dist.	Average of % Lev. Dist
Transliteration1	Transliteration	10	16.3	50	53.7
Transliteration2		9		55	
Specialized1		30		56	
Translation1	Translation	66	66	37	38.5
Translation2		66		40	

Results/Findings

- The following information needed for transformation
    - Current language and script
    - Method of obtaining current data (manual or transformed)
- For transformed data, additional information needs to be recorded:
    - Source language and script
    - Type of transformation (translation or transliteration)
    - Mechanism of transformation (manual or automated)
    - Standard used for the transformation (for transliteration)

Results/Findings

- One tool may not work for all contact information
- Transliteration is usable for scripts which fully specify consonants and vowels, not work for scripts where consonants or vowels are under-specified
- Ad hoc transformation
  - using translation systems
  - give an arbitrary output; not predictable
  - more readable and independent of the scripts of the language pair
  - perform better from an end-user perspective
  - limited set of language pairs which have mature automatic translation systems
  - new translation system for a language pair is very challenging

Results/Findings

- Consistent transformation is possible through transliteration
  - compromises the comprehensibility of the information; especially between scripts which encode information differently
  - still inconsistent if different standards or tools are used
- Accurate transformation is not possible through automated processes
  - requires manual effort, including registrant verification
- Pivoting through romanization interesting possibility to provide local language to local language transformation
  - two levels of transformation involved make output inaccurate for effective use, given variation in transformation techniques and tools