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

1. Overview

As part of the process to implement the internationalized registration data recommendations of the ICANN WHOIS review team, ICANN seeks to commission a study to document and evaluate the potential solutions submitting or displaying contact data in non-ASCII (American Standard Code for Information Interchange) character sets. In this Terms of Reference, we provide the background and outline the requirements for the study.

2. Background

Domain Name Registration Data (DNRD) is accessible to the public via a Directory Service (also known as the WHOIS service). It Domain Name Registration Data (DNRD) – refers to the information that individuals or organizations ("registrants") submit when they register a domain name. Domain name registrars or registry operators collect these data, and some of the data is made available for public display or for use by applications. The data elements that registrants must submit to ICANN Accredited Generic Top Level Domain (gTLD) registrars (and accurately maintain), are specified in the ICANN Registrar Accreditation Agreement (RAA). For Country Code Top Level Domains (ccTLDs), the operators of these TLD registries define their own data elements or follow their country's policy regarding the request and display of registration information.

Much of the currently accessible domain name registration data (DNRD) is encoded free form in US-ASCII. This legacy condition is convenient for WHOIS service users who are sufficiently familiar with languages that can be submitted and displayed in US-ASCII to be able to use ASCII script to submit DNRD and make and receive WHOIS queries using that script. However, these data are less useful to the WHOIS service users who are only familiar with languages that require character set support other than US-ASCII for correct submission or display. It is important to note that the latter (currently underserved) community is likely to continue growing and is likely to outnumber the former group in a matter of years.[1]

---

1. Internet Corporation for Assigned Names and Numbers . "*Final Report of Internationalized Registration Data Working Group*", 2012. Available at http://gnso.icann.org/en/issues/ird/final-report-ird-wg-07may12-en.pdf

Internationalized Registration Data (IRD) refers to DNRD that are represented in languages or scripts whose characters cannot be represented in US-ASCII. To support internationalized registration data, many registries have developed specific conventions that are documented in the GNSO-SSAC Internationalized Registration Data WG final report. However, these conventions are not standardized, as the WHOIS protocol has no mechanism to signal character encoding.

The WHOIS Policy review team, formed as a result of the Affirmation of Commitments (AoC), was charged to "assess the extent to which the WHOIS policy is effective and its implementation meets the legitimate needs of law enforcement and promotes consumer trust." In its Final Report, it highlights the needs to define requirements and evaluate solutions for internationalized registration data, with following recommendations:

> "ICANN should task a working group within six months of publication of this report, to determine appropriate internationalized domain name registration data requirements and **evaluate available solutions**; at a minimum, the data requirements should apply to all new gTLDs, and the working group should consider ways to encourage consistency of approach across the gTLD and (on a voluntary basis) ccTLD space; working group should report within a year.[2]"

Last November, the ICANN Board adopted an Action Plan[3] in response to the WHOIS Review Team's Final Report, that instructs Staff to conduct this Study to address this recommendation. In particular, the Board directed staff to: 1) task a working group to determine the appropriate internationalized domain name registration data requirements, evaluating any relevant recommendations from the SSAC or GNSO; 2) produce a data model that includes (any) requirements for the translation or transliteration of the registration data, taking into account the results of any PDP initiated by the GNSO on translation/ transliteration, and the standardized replacement protocol under development in the IETF's Webbased Extensible Internet Registration Data Working Group; 3) evaluate available solutions (including solutions being implemented by ccTLDs). See http://www.icann.org/en/groups/board/documents/resolutions-08nov12-en.htm.

In order to meet these recommendations, ICANN needs to define the following requirements for internationalized registration data, and the registrars and registry

---

[2] Internet Corporation for Assigned Names and Numbers. "*WHOIS Policy Review Team Final Report*", 2012. Available at http://www.icann.org/en/about/aoc-review/whois/final-report-11may12-en.pdf

[3] Internet Corporation for Assigned Names and Numbers. "Action Plan to Address WHOIS Policy Review Team Report Recommendations." 2012. Available at: http://www.icann.org/en/groups/board/documents/briefing-materials-1-08nov12-en.pdf

operators need to deploy systems and processes when dealing with internationalized registration data:

Submission

1) Whether to allow users to submit internationalized registration data?
2) If users are allowed to submit internationalized registration data, what languages or scripts are registrars or registry operators expected to support?
3) If users are allowed to submit internationalized registration data, whether to require that users submit corresponding US-ASCII versions of the internationalized contact data?
4) If users are required to submit US-ASCII versions of the internationalized registration data, are users expected to submit a translated version, a transliterated version, or a transcribed version, or "either" (providing there is a convention or method to distinguish between the two)?
5) If users are required to submit US-ASCII versions of the internationalized registration data and the user is unfamiliar with or unable to submit such a transformation, are registrars or registry operators expected to provide assistance (and if so, how would such assistance be manifested?)

Storage & Transmission

6) Are registrars and registry operators expected to store US-ASCII and internationalized versions of DRND? Is this always "two" sets of DNRD?
7) If not US-ASCII, does each ICANN accredited registrar choose which language/script to store independently? How would this affect accuracy of registration data?

Display

8) Since the WHOIS protocol has not been internationalized, how does the registry/registrar support the display of internationalized registration data?

The purpose of the study is to document and evaluate the available solutions for internationalized contact data to aid in these implementation decisions. It is expected that community consultation to take place in relation to these implementation decisions.

3. Requirements

The study should cover the following subjects or issues:

3.1. Document the submission practices of internationalized registration data at a representative set of gTLD and ccTLD registries and registrars.

3.2. Document the display practices of internationalized registration data at a representative set of gTLD and ccTLD registries and registrars.

3.3 As electronic merchants and online service providers in other industries often have to accommodate submission or display of their content in multiple languages, investigate and document how other e-merchants or web sites manage internationalized contact data.

3.4. Consider and assess the cost and functionality of commercial, open source, or other known but as yet not widely implemented solutions for 1) transliterating internationalized contact information to US-ASCII, 2) translating internationalized contact information to English, 3) transcribing internationalized contact information to US-ASCII, or 4) a mixture of translation, transliteration and transcription. For example, see the following table.

	Internationalized Data	Transliteration	Translation[4]	Mixture of Translation and Transliteration
Name	技术联系人	ji shu lian xi ren	Technical Contact	
Organization	中国互联网络信息中心	Zhong guo hu lian wang luo xin xi zhong xin	China Internet Network Information Center	
Address Information	北京中关村南四街四号, 100080	Beijing zhongguan cun nan si jie si hao, 100180	4 South 4th Street Zhongguancun beijing, Beijing 100080, CN	4 South 4th Street Zhongguancun beijing, Beijing 100080, CN

3.5 Consider and assess the accuracy implications for transliteration and translation of the internationalized contact data, for exampled as noted in [1], translation or transliteration system has the following challenges:

- According to RFC 6365, many language translation systems are inexact and cannot be applied repeatedly to translate from one language to another. Thus there will be problems with both consistency and accuracy. For example:

[4] Note, in some cases it is impossible to determine the language of the name of a person.

- Translation/transliteration may vary significantly across languages using the same script.

- Two people may translate/transliterate differently even within a language and the same person may translate/transliterate differently at different times for the same language.

- As a concrete example, محمد (U+0645 U+062D U+0645 U+062F) is a commonly used name in the Arabic script based languages (270 million pages for محمد found on Google on 19th Feb. 2011). It is translated/transliterated to English in the many ways (some listed below): Mohammed, Mohamed, Muhammed, Muhamed, Mohammad, Mohamad, Muhammad, Muhamad. So if محمد is the name of a monolingual registrant (a likely possibility), which spellings should Registrar A choose? Will Registrar B choose the same spelling? Also, how would a registrar/registry determine which particular spellings to use for a particular registrant? How would the monolingual registrant ever verify such information even if presented such data by the registrar or by a third organization that does the translation/transliteration?

- According to RFC 6365, many script transliterations are exact. There are also official and unofficial transliteration standards, most notably those from ISO TC 46[5] and the U.S. Library of Congress.

    - However, for a given script, there may exist multiple systems for transliteration into Latin. In the case of Chinese, these systems are not only quite different from each other, but most of them use particular Latin characters to represent phonemes that are quite different from the most common phoneme-character pairings in European languages.

Based on practices documented in 3.1 and 3.2 and understanding the issues raised in 3.5 and best practices by other e-merchants in 3.3, what are the common best practices registry/registrar could do to minimize these variations so that translation, transliteration or transcription are done in an un-ambiguous way across

[5] For example: ISO 9:1995 Cyrillic -> LATIN, ISO 233:1984 Arabic -> LATIN, ISO 233-2:1993 Arabic -> LATIN, simplified, ISO 259:1984 Hebrew -> LATIN, ISO 843:1997 Greek -> LATIN, ISO 3602: 1989 Japanese -> LATIN, ISO 7098:1991 Chinese -> LATIN, ISO 9984:1996 Georgian -> LATIN, ISO 9985:1996 Armenian -> LATIN, ISO 11940:1998 Thai -> LATIN, ISO/TR 11941:1996 Korean-> LATIN, ISO 15919:2001 Denanagari -> LATIN.

all registrars/registries. For example, one such practice could be to have automatic translation + user confirmation/validation, if possible.

The final product is to be in the form of a report. The interim deliverables include 1) a study proposal (along with detailed methodology), 2) preliminary report to be posted for public comment, 3) summary of public comments in the standard ICANN form, and 4) final report after incorporating the community feedback gleaned from the public comments received.