# Translation & Transliteration Draft Recommendations

Chris Dillon

Version 6, 24 July, 2014

The aim of this draft is to provide, for the purpose of discussion, draft recommendations and explanations for the questions in the *Translation and Transliteration of Contact Information Policy Development Process (PDP) Working Group* Charter. This straw man also addresses those additional questions the Working Group has identified during its meetings, and those the *Expert Working Group to Define Requirements for Internationalized Registration Data and Corresponding Data Model for gTLD Registries* asked the Working Group to respond to. They are all listed here:

[https://community.icann.org/display/tatcipdp/4+Proposed+Questions+and
+Taxonomies](https://community.icann.org/display/tatcipdp/4%2BProposed%2BQuestions%2Band%2BTaxonomies)

In the below “transformation” is used as shorthand for "translation or transliteration".

See [community.icann.org/display/tatcipdp/1+What+is+contact+information
+and+What+Taxonomies+are+Available](https://community.icann.org/display/tatcipdp/1%2BWhat%2Bis%2Bcontact%2Binformation%2Band%2BWhat%2BTaxonomies%2Bare%2BAvailable) and O1 below for definitions of technical terms.

## Main questions in the Working Group's charter (marked C for Charter):

**C1. Is it desirable to translate contact information to a single common language or transliterate contact information to a single common script?**

### WG Deliberations:

* The purpose of transformed data is to allow those not familiar with the original script to contact the registrant and thus accuracy of the transformed data is paramount.
* It would be very difficult if not impossible to maintain consistency if transformations across millions of data entries in a very large number of scripts/languages were to take place.
* The WG believes that data will be most reliable if registrants can enter contact information in their own language/script or the language/script used by their chosen registrar.
* It is important though that those wishing to contact a registrant have a clear idea which part of the registration directory data entry is the name, street, town and email address even if those are provided in non-Roman scripts. Therefore, labelling of the language/script used is important.
* The costs of transformation of all non-Roman scripted registration directory entries would be much higher than can be justified in view of its potential usability, especially considering accuracy and the language capabilities of registrants.

The WG notes that even if a transformed version of the data is available, it is unlikely that communication in Latin script with a registrant who has provided IDN registration data would be effective, rendering mandatory transformation ineffectual.

However, it is not the intention of the *Translation & Transliteration of Contact Information PDP Working Group* to discourage best-practice transformation by registries, registrars or even registrants. The WG notes that the proposal for a next generation gTLD Directory Service, as outlined in *The final report from the EWG on gTLD Directory Services*, currently has no Internationalized Registration Data (IRD) functionality. The WG is of the opinion that such functionality needs to be available for registrants so that they can enter IRD in their native language/script or – at least – the language/script their registrar operates in. The WG believes that if registrants are able to enter data in their own languages/scripts, the accuracy of the data is likely to increase.

The United Nation’s (UN’s) recommendation should be followed concerning the common language: **“[t]he 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”**

*Manual for the national standardization of geographical names* (UNGEGN, 2006). *Group of Experts* refers to the *United Nations Group of Experts on Geographical Names*. This is their practical solution to the challenge of finding the right place consistently.

See E2 below for how non-Roman script contact information should be transformed.

### Draft Recommendations

**#1 The WG recommends that it is not desirable to make transformation of contact information mandatory.** However, the current WHOIS and the future, new Registration Directory Service (RDS) databases should be capable of receiving input in the form of non-Roman script contact information.

Even if money were no object, it would be undesirable to transform all non-Roman script contact information as it would not be consistent and accurate enough for many purposes without checking.

**#2 The WG recommends that any future gTLD directory service should be capable of storing non-Roman script data and a transformed version in Roman script to accommodate the possibility of a ‘best practice’ transformation service.**

**#3 The WG recommends that as part of the PDP on the purpose of gTLD Registration Data, the need to add IRD capability (see #2) to any new gTLD Directory Service is addressed. Crucially this should include tag fields to indicate the languages used in the addresses and date fields.**

**#4 The WG recommends that registrants provide their contact information in the language/script that their registrar operates in. The WG believes that this will provide data that is as accurate and reliable as possible. The WG notes that this recommendation does not prevent registrars/registries from providing voluntary best practice transformation.**

**C2. Who should decide who should bear the burden of translating contact information to a single common language or transliterating contact information to a single common script?**

### Observations

The WG notes that this question relates to the concern expressed by the *Internationalized Registration Data Working Group* (IRD-WG) in its report that there are costs associated with providing translation and transliteration of contact information. For example, if a policy development process determined that the registrar must translate or transliterate contact information, this policy would place a cost burden on the registrar.

However, as the WG has concluded that it would not be desirable to require translation and transliteration of contact information, there is no need to make a general decision on the distribution of the financial burden. Stakeholders who decide to transform contact information, will have to bear the costs themselves.

### Recommendations:

**#5 The WG recommends that there is no need to determine who bears the costs as no mandatory transformation is recommended (see #1 above).**

## Other questions in the charter:

**C3. What exactly the benefits to the community are of translating and/or transliterating contact data, especially in light of the costs that may be connected to translation and or transliteration?**

* Transformation would to some extent facilitate communication among stakeholders not sharing the same language. Good communication inspires confidence in the Internet and makes bad practices more difficult.
* English is currently the de facto language for intercultural communication and business transactions. It is the language likely to benefit the greatest number. Moreover, if these recommendations are followed, the transformed data are in the Roman alphabet, making them to some extent accessible by speakers of other lingua francas such as French and Spanish.
* Searching contact information is easier in one language.

However,

* these benefits are outweighed by the financial burdens that would be imposed on stakeholders. Such burdens would be substantial enough to make the expansion of the Internet and provision of its benefits considerably more difficult in the developing world. This is the main reason for this PDP Working Group’s recommendation #1.
* A registrant should be able to submit contact information in his/her own language to the registrar, or at the very least in the operating language/script of a registrar. This should be the basic requirement.
* An additional burden would be achieving accuracy in transforming a very large number of scripts and languages – mostly of proper nouns – into a common script and language.

**#6 The WG recommends that IRD becomes the basic requirement for directories of DNRD.**

**C4. What impact will translation/transliteration of contact data have on the WHOIS validation as set out under the 2013 Registrar Accreditation Agreement?**

As costs are only incurred by stakeholders requiring transformed data for their needs, it is unlikely that the 2013 RAA would be affected. If some effect were to come to light, transformation could not affect the legal provisions in the 2013 RAA: *Registrar shall implement internationalized registration data publication guidelines according to the specification published by ICANN following the work of the ICANN Internationalized registration Data Working Group (IRD-WG) and its subsequent efforts, no later than 135 days after it is approved by the ICANN Board.*

Future RAAs should be written in the light of the policy in this PDP Working Group’s final report. For example, recommendation #6 of IRD as the basic requirement could affect future RAAs.

**C5. When should any new policy relating to translation and transliteration of contact information come into effect?**

This policy should come into effect as soon as possible – and at the latest with the introduction of the new internationalized RDS. To impose policy on stakeholders before then may force them to increase the number of temporary solutions and any transferral of data between systems may have to take place twice instead of once.

## Other questions the group believes to be important (and marked O) are:

**O1. What is contact information and what taxonomies are available?**

Contact Information as defined in the *Final Issue Report on the Translation and Transliteration of Contact Information* based on the definition in the *Registrar Accreditation Agreement 2013*: <http://gnso.icann.org/en/issues/gtlds/transliteration-contact-final-21mar13-en.pdf>

"In the context of these issues, “contact information” is a subset of Domain Name Registration Data. It is the information that enables someone using a Domain Name Registration Data Directory Service (such as WHOIS) to contact the domain name registration holder. It includes the name, organization, and postal address of the registered name holder, technical contact, as well as administrative contact. Domain Name Registration Data is accessible to the public via a Directory Service (also known as the WHOIS service). The Registrar Accreditation Agreement (RAA 3.3.1) specifies the data elements that must be provided by registrars (via Port 43 and via web-based services) in response to a query, but it does not require that data elements, such as contact information, must be translated or transliterated."

See also: [https://community.icann.org/display/tatcipdp/1+What+is+contact+
information+and+What+Taxonomies+are+Available](https://community.icann.org/display/tatcipdp/1%2BWhat%2Bis%2Bcontact%2Binformation%2Band%2BWhat%2BTaxonomies%2Bare%2BAvailable)

**O2. Who gets access to what information?**

This question is beyond the remit of this PDP. As regards the current WHOIS, whether data are original language/script or transformed does not affect stakeholders’ access rights to them. The question is addressed in *The final report from the EWG on gTLD Directory Services*. The policy as described in the final report presumes that only those with the right may access data and that data protection and freedom of information principles have been correctly implemented.

This raises the question of access to transformed data added by other stakeholders. This should be allowed within the limits just outlined.

**O3. Who are the stakeholders — who is affected and what do they want?**

The stakeholders include all Internet users, registrants, registrars, registries, ICANN, security organizations et al.

For *what do they want*, see: [https://community.icann.org/display/tatcipdp/13+Community+Input](https://community.icann.org/display/tatcipdp/13%2BCommunity%2BInput) and *The final report from the EWG on gTLD Directory Services*.

**O4. How much would a particular feature cost and how to weigh the costs versus the benefits?**

Accurate, reliable transformation is expensive. Existing automated systems for transformation are inadequate. They do not provide results of sufficient quality for purposes requiring accuracy and cover fewer than 100 languages. Developing systems for languages not covered by transformation tools is slow and expensive, especially in the case of translation tools. For purposes for which accuracy is important, transformation work often needs to be done manually.

See *Study to evaluate available solutions for the submission and display of internationalized contact data* for further information.

## Questions from the Expert Working Group to Define Requirements for Internationalized Registration Data and Corresponding Data Model for gTLD Registries:

See: [www.icann.org/en/news/announcements/announcement-08jul13-en.htm](http://www.icann.org/en/news/announcements/announcement-08jul13-en.htm)

10 April 2014: Interim Report (Page 17)

The EWG noted the following issues related to translation and transliteration and recommended that this PDP Working Group would consider them:

**E1. If registrants are allowed to submit localized registration data, what languages or scripts are registrars or registry operators expected to support?**

Registrars' and registry operators' systems must at least support the input of contact data in one of the official languages of the countries in which they are based. An ability to support users in those languages will be beneficial to business.

**#7 This WG recommends that there should be no requirement for registrars or registry operators to support English.**

**E2. If registrants are required to submit a single common script version of the internationalized registration data, are users expected to submit a translated version, a transliterated version, or “either” (provided there is a convention or method to distinguish between the two)?**

In cases when transformed data are input, the official English translation of organizational names should be used where one exists, otherwise transliteration. Addresses should be transliterated except for country names, which should be selected from a drop-down list of English names.

Translation should use official translations. If there is no official (used for example on the organization's headed paper) translation of an organizational name, it should be transliterated.

Transliteration should follow the rules in a national standard of the language where one exists and failing that in a national standard of a related language using the same script. Addresses are transliterated (except for the country name which is given in English). There may be issues with letters that do not exist in the related language or with letters that are transliterated differently depending on the language. It may be possible reliably to pivot (automatically transliterate) between some alphabetic scripts: for example, Roman, and Cyrillic and Greek, but not, for example, Arabic and Neo Brahmi.

Note that:

* If this solution is implemented, English only occurs in two fields (organization name and country) and the latter list is relatively short and easy to translate.
* Transliteration is easier to automate than translation. Many reliable systems already exist for alphabetic scripts and it is relatively quick to develop more.
* Some parts of addresses would ideally be translated; for example the translated Bangkok is more useful internationally than the transliterated Krung Thep. However, the transliterated Beijing is much more useful than the translated Northern Capital. It is not easy for automated systems to know when to translate such cases as Krung Thep.
* Such contact information would be usable for postal purposes.

**E3. If registrants are required to submit a single common script version 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)?**

As transformation is required by other stakeholders depending on their needs, there should be no burden on registrars or registry operators.

**E4. If registrants are required to submit a single common script version and an internationalized version of their registration data, should there be a requirement to detect whether both literally match each other?**

If transformation is required, accuracy (involving matching) will be required for some purposes, for example legal purposes and validation. It is possible to have many kinds of translation and many kinds of literal translation. It would be possible to answer the question of whether an official translation of an organizational name was being used or not.

As long as the same transliteration is being strictly used for a language, it should be possible to match two transformations of the same data.

**E5. If there are two versions of the registration data, which version should be considered primary or authoritative if there is a mismatch?**

If there is a mismatch, the form in the original language is the primary form and should be regarded as authoritative.

**E6. If translated / transliterated versions of the data are required, how will data be maintained simultaneously in multiple languages/scripts? Should there be additional meta-level information?**

* Language tags are required to indicate the language of the addresses. These are supported in RDAP but not in EPP.
* Each language version needs to indicate the validation status of the data. This is not possible with the existing implementations of EPP and RDAP.
* Addresses need to have a date field to indicate when the data was input. This functionality is not supported in EPP or RDAP.
* It is possible that three of even more languages would be required in the directory – original, English and then local language(s).

**E7. For company and individual names, should translation or transliteration be required?**

* For organizational names, official translations should be used if they exist, otherwise transliterations. A tag should indicate the language of the organizational name.
* For individual names, the form preferred by the individual should be used. When that is not available, transliteration should be used.

## Relevant information not covered in the questions:

### Foreign language field names

When data are not transformed, the provision of translated field names in the future RDS would at least indicate the relevant parts of foreign language contact data. The data need to be tagged for language so that it is clear which transformation should be used if it is required.

## Issues

* If WHOIS is replaced by a system without IRD functionality, there will be pressure from the non-Roman script world for that system to be replaced.
* Work is now required to support stakeholders who will transform contact information so that a future RDS will have the functionality they require.