Translation & Transliteration Draft Recommendations

Chris Dillon
Version 8, 5 September, 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. They are all listed here: https://community.icann.org/display/tatcipdp/4+Proposed+Questions+and+Taxonomies

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

Main two 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:

1. The main 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.
2. 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.
3. The WG believes that data will be most accurate if registrants can enter contact information in its language/script. So, for example, Thai contact information would be entered in the Thai script.
4. 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 a non-Roman script. Therefore, labelling of the language/script used in the various fields in the database is important.
5. 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.
6. 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.

7. 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. Work is now required to support stakeholders who will transform contact information so that a future RDS will have the functionality they require.

The United Nation's (UN's) recommendation should be followed concerning the common language if contact information is transformed: **"[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"** From: *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 O5 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. If WHOIS were to be replaced by a system without IRD functionality, there would be pressure from the non-Roman script world for that system to be replaced.

The WG notes that some stakeholders are in favour of mandatory transformation.

However, 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 address fields.
#4 The WG recommends that registrants provide their contact information in a language or script appropriate for the region in which that address is located. The WG believes that this will provide data that are as accurate as possible. The WG notes that this recommendation does not prevent registrars/registries from providing 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 transform 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 transformation 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 issues raised in the charter:

C3. Transformation – benefits vs costs

1. 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.
2. 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.
3. Searching contact information is easier in one language.

However,

4. 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 *.
5. A registrant should be able to submit contact information in the language of the contact information. This should be the basic requirement.
6. 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.

* Accurate 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.

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

C4. Impact of transformation on WHOIS validation as set out under the 2013 Registrar Accreditation Agreement

As costs are only incurred by stakeholders requiring transformed contact information 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 on transformation come into effect?

As this working group's recommendations are not binding in the case of stakeholders who carry out transformation, the policy may come into effect as soon as stakeholders transform data. The recommendations presume the existence of a system which can handle internationalized registration data.

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*: "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.
See also: https://community.icann.org/display/tatcipdp/1+What+is+contact+information+and+What+Taxonomies+are+Available

O2. Who gets access to what information?
This question is beyond the remit of this PDP. As regards the current WHOIS, whether contact information is original language/script or transformed does not affect stakeholders’ access rights to it. 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.

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 and *The final report from the EWG on gTLD Directory Services*.

O4. 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 languages of the contact information. For example, Singaporean contact data could be entered in English, Mandarin, Malay or Tamil. 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.

O5. In cases when contact information is to be transformed, how should it be done?

Addresses should be transliterated except for **country names**, which should be selected from a drop-down list of English names.

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. 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 Devanagari.

Note that:

1. 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.
2. Transliteration is easier to automate than translation. Many reliable systems already exist for alphabetic scripts and it is relatively quick to develop more.
3. 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.
4. For **organizational and personal names**, the Romanized forms preferred by the organization and individual should be used. When those are not available, transliteration should be used.
5. The contact information described in these recommendations would be usable for postal purposes.

Example primary record

```
Status: 検証済み JA
Date: 2014年8月28日 JA
Registered name holder: 岡崎太郎 JA
Organization: 国立情報学研究所 JA
Postal address: 日本 テ101-8430 東京都千代田区一ツ橋2－1－2 JA
Email address: 岡崎.太郎@グーグル.日本 JA
```

Example transformed record

```
Status: validated EN
Date: 28 Aug 2013 EN
Registered name holder: Ted Okazaki EN
Organization: National Institute for Informatics EN
Postal address: 2-1-2 Hitotsubashi Chiyodaku Tōkyō 101-8430
Japan EN
Email address: 岡崎.太郎@グーグル.日本 JA
```

Notes

1. The language of the contact information is the primary, **authoritative** version.
2. It is possible that three of even more languages would be required in the directory – original, Romanized and then other local language(s).
3. Other statuses will include `legacy` for data imported from older systems.
4. In this case the transformed data may be useless, as they are a year old.
5. Mr Okazaki's name is actually pronounced Tarô, but he uses "Ted" when speaking English.
6. The data need to be tagged for language, e.g. JA, EN, so that it is clear which transformation should be used if it is required.
7. Acronymns (e.g. NII) are not used, unless there is no long form. If an organization name has no official English form, then a transliterated form will appear, in this case: `Kokuritsu Jōhōgaku Kenkyūjo`. Note that there is no automated transliteration system for Japanese and so spaces and capital letters not in the original can appear in the transliterated version.
8. In fact the postal address would be split into various fields. There would be similar records for the technical contact and administrative contact. 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.

O6. Do IRD and transformed versions need to 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.

Appendix A: Chart of charter questions and recommendations

Recommendation	Covered	Agreed	Some disagreement	Under discussion
Is it desirable to translate contact information to a single common language or transliterate contact information to a single common script?	Y	N	Y	Y
#1 The WG recommends that it is not desirable to make transformation of contact information mandatory.	Y	N	Y	Y
#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.	Y	N[1]	Y	Y
#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 address fields.	Y	Y	N	N

[1] Some stakeholders believe transformed contact information is undesirable as it is difficult to match with the original language contact information.

Recommendation	**Covered**	**Agreed**	**Some disagreement**	**Under discussion**
#4 The WG recommends that registrants provide their contact information in a language or script appropriate for the region in which that address is located. The WG believes that this will provide data that are as accurate as possible. The WG notes that this recommendation does not prevent registrars/ registries from providing best practice transformation.	Y	Y	N	Y
#5 The WG recommends that there is no need to determine who bears the costs as no mandatory transformation is recommended (see #1 above).	Y	Y	Y[2]	Y
#6 The WG recommends that IRD becomes the basic requirement for directories of DNRD.	Y	Y	N	N
#7 This WG recommends that there should be no requirement for registrars or registry operators to support English.	Y	Y	N	N

[2] At least at an earlier stage some stakeholders expressed a need for transformed contact information to counter phishing.

Appendix B: The case for mandatory transformation

1. It is desirable to make transformation mandatory as the availability of contact information in a single common language/script makes it easier to contact registrars in the event of legal, security etc. issues.
2. The costs of transformation could be spread among the stakeholders requiring it. Costs could be reduced in the case of alphabetic scripts such as the Cyrillic and Greek alphabets by using automatic transliteration. In the case of non-alphabetic scripts such as Arabic, Chinese and Japanese, optional transformation by registrants could reduce the costs. Data consistency would be an issue, but in most cases the data would enable the registrars to be contacted.