- DRAFT DRAFT DRAFT -
Extended Outline for the IRD-WG Final Report
July 13 2011

EXECUTIVE SUMMARY-TBA

1. INTRODUCTION

- Accommodating the submission and display of internationalised registration data in WHOIS is seen as an important evolutionary step for Whois services.
- Quote RFC 4690
- The SSAC called attention to these issues in SAC037, *Display and usage of Internationalized Registration Data, Support for Characters from Local Languages or Scripts*.[1] In the report the SSAC recommended that the ICANN Board of Directors should form a working group to study the feasibility and suitability of introducing submission and display specifications to deal with the internationalization of registration data.
- At the request of the ICANN Board of Directors, the GNSO and the SSAC created the IRD-WG to study this issue.

1.1 IRD-WG objective and Goals

- The board asked the IRD to study: 1) the feasibility and suitability of introducing submission and display specifications to deal with the internationalization of Registration Data; and 2) Engage participation from all ICANN Supporting Organizations and Advisory Committees as well as Country Code top level domain (ccTLD) operators, to ensure broad community input.

- The WG interprets Board's request into two issues:
    - Suitability: Is it suitable or desirable to have internationalized registration data?

    - Feasibility: Is it feasible to introduce submission and display specifications to deal with IRD?

[1] Security and Stability Advisory Committee, "SAC037, Display and usage of Internationalized Registration Data, Support for Characters from Local Languages or Scripts," 21 April 2009, <http://www.icann.org/en/committees/security/sac037.pdf>.

- ***IRD-WG Membership:*** Edmon Chung (GNSO) and James Galvin (SSAC) co-chair the IRD-WG. The international representation in the IRD-WG includes 17 participants, 5 staff support, 5 countries (China, Tunisia, New Zealand, Russia, and the USA), 3 ccTLDs (.cn, .nz, .ru) and 3 ICANN Supporting Organizations and Advisory Committees (ALAC, GNSO, SSAC).[2]

- 1.2 Terminology

- WHOIS related terminologies

  1) Domain Name Registration Data – refers to the data that registrants provide when registering a domain name and that registrars or registries collects. All of a portion of this set of data is made available to the public. For interaction between ICANN Accredited Registrars and registrants, these data elements are specified in the Registrar Accreditation Agreement and individual Registry agreements with ICANN.

  2) Registration Data Access Protocol – refers to the elements of a (standard) communications exchange – queries and responses - that make access to registration data possible. Of these protocols, the WHOIS protocol, as defined in RFC 812/954/3912, is the primary one used today.

  3) Registration Data Directory Service – refers to the service(s) offered by registries and registrars to implement the protocol defined in (2) and to provide access to (potentially a subset of) the registration data.

- IDN related terminologies
  - A-label | U-label

- Cite any relevant definitions from RFC 3536 (currently in its final call stage): https://www1.tools.ietf.org/html/draft-ietf-appsawg-rfc3536bis-06

2. BACKGROUND

(NOTE: This section should include all the facts we need to support our findings. Most of this is in our interim report.)

[2] For a list of the IRD-WG members and the Charter, see the IRD-WG wiki at <https://st.icann.org/int-reg-data-wg/index.cgi?internationalized_registration_data_working_group>.

2.1. What is Domain Name Registration data?

ICANN requires that a registrant provide certain information when registering a domain name, and that registrars or registries make this registration information available for public scrutiny. The Registrar Accreditation Agreement (RAA 3.3.1) specifies the following data elements that must be provided by registrars in response to a query:

3.3.1.1 The Registered Name;

3.3.1.2 The names of the primary name server and secondary name server(s) for the Registered Name;

3.3.1.3 The identity of the Registrar (which may be provided through Registrar's website);

3.3.1.4 The original creation date of the registration;

3.3.1.5 The expiration date of the registration;

3.3.1.6 The name and postal address of the Registered Name Holder;

3.3.1.7 The name, postal address, e-mail address, voice telephone number, and (where available) fax number of the technical contact for the Registered Name; and

3.3.1.8 The name, postal address, e-mail address, voice telephone number, and (where available) fax number of the administrative contact for the Registered Name.

Generally the data can be considered in the following categories:

- Domain name
- Registration status
- Nameserver information
- Names (owner, admin, technical contact)
- Addresses (owner, admin, technical contact)
- Phone/fax numbers (admin, technical, owner phone/fax)
- Dates (creation date, expiration date, update date)

In ccTLDs, there is no such requirement, and each ccTLD sets its own policy regarding what constitute this data, usually a subset of the data above is displayed in ccTLD WHOIS services.

2.2. Where different registration data elements are collected, stored, managed, and displayed?

In gTLD environment:

For Submission:

- Registrant provide the information when registering the domain with registrars / resellers.

- Registrars store the information, and in the case of thick registries, submit a copy of the information to the registry through the Extensible Provisioning Protocol (EPP)

For Query:
- End users query the registration data directory service for domain name, contact information or name server information.
    - For thick registries, the query displays the data from registry's WHOIS service.
    - For thin registries, the query continues to query registrar's WHOIS service and displays the data from registrar's WHOIS service.

2.3. Current practices by registrars and registries and ccTLDs to support the display of internationalizes data.

- No standards, conventions, or policy requirements exist
- Support of ASCII7 CRLF is the only obligation
- Whois services commonly only support submission and display using ascii7
- Certain registries have developed various ad hoc or TLD-specific conventions (Jay's Daley identified enough of these in his paper/email for us to give examples)
- Conventions that are adopted for Whois/43 clients may not be used or useful for web based services (i.e., if the web based whois service accesses registrar or registry data doesn't "proxy" queries onto a port 43 query…)

2.4. The WHOIS protocol

Current Protocol's capability to support IRD

- the WHOIS protocol does not specify any protocol means to signal encoding" and "the WHOIS protocol imposes no restrictions on what data are transferred in a query or response, only that the end of data must be signaled by the use of an ASCII CR/LF. This has been loosely interpreted by many server developers as "WHOIS is ASCII7 only" and by other server developers as "I'll formulate my own way to signal to the client that I'm sending UTF-8.

- No formal data schema: Functionally this means that: 1) it is difficult for third parties who want to translate or transliterate to discriminate between label and data.

3. INTERNATIONAL STANDARDS

(NOTE: This could be a part of the background information but my current thinking is that it is better to elevate to a major section)

In this section we summarize all the international standards and standard practices that exist for internationalizing the various elements of existing registration data.

3.1 IETF standards

- IDNA
- EPP Mappings RFC 5730-5734
- DREG data schema (RFC 3982)

3.2 UPU standards

- UPU E.123 (for telephone/fax number)
- UPU S.42 (address templates)

3.3 Standards for transliteration and translation

- 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

4. FINDINGS

(NOTE: In this section we list the conclusions we can draw from all the facts stated previously)

4.1 Is it desirable to represent domain name registration data in non-US ASCII?

- Much of the original and currently accessible domain registration data are encoded in US-ASCII. This legacy condition is convenient for Whois service users that are familiar with languages that can be submitted and displayed in US-ASCII7. It is also convenient for registrants, registrars and registries and installed base of operational Whois services that display US-ASCII.

- However, these data are less useful to the Whois service users that are only familiar with languages that require character set support other than US- ASCII7. It is important to note that it is likely that the latter (underserved) community will continue to grow and could outnumber the former in a matter of years.

- Many registrants are monolingual, which is the expectation and motivation behind internationalized domain names. Therefore it is unreasonable to assume they know or can enter the registration data in languages other than their local language.

- Thus it is desirable to represent domain name registration data in non-US ASCII.

- However, this desirability should be balanced against other uses of the data. While a domain registrant may intend to only use their domain "locally" or interact with people in their native script, the nature of the Internet itself means that any domain provisioned on it is available universally. In some cases, such as law enforcement and electronic crime investigators (see http://forum.icann.org/lists/ird-wg-report/msg00004.html)

4.2 Is it feasible to introduce submission and display specifications to deal with IRD

Yes, it is feasible, but there are several barriers that needs to be overcome:

- Port 43 WHOIS protocol has not been internationalized: It imposes no formal encoding or composition requirements on data exchanged between a client and a server and hence no method of signaling encodings, and there may be different implementations from TLD to TLD on how this is done. This would create a non-uniform experience for users and interoperability problems. If ICANN continue to require Port 43 WHOIS, this issues needs to be addressed urgently with the technical community. (NOTE: Solutions exist today to accommodate monolingual users in the web environment.)

- At a minimum, the current WHOIS data needs to be tagged with language/script. In addition, the community would benefit from a standard registration data schema (for example in XML) with language tag (RFC 4646):

    - A formal data schema for registration data (for example in XML) would enable end user clients to localize the data label and data better if they choose.
    - A formal data schema for registration data with language tag information would allow better processing of the data.
    - Registrars may, in their policies, allow for more multiple languages in the contact data (for example, they could allow for a Arabic Registrant living in America to put his/her name in Arabic, but his/her address in English). If this is needed, the language-tag data needs to be at a data element level.

- There are recognized standards for internationalizing many of the elements of domain registration data, and to the extent possible these standards should be followed. (See previous sections on these standards)

- One key issue is internationalizing contact information. These data elements include Names (owner, admin, technical contact) and Addresses (owner, admin, technical contact).

    - The central issue here is to balance the need / capabilities of the local registrant versus the need of the (potential) global user of this data.

    - Several options have been discussed. These are:

        - Registrants provide domain contact data in "Must Be Present" script.
        - Registrants provide data in any registrar-accepted script and registrars provide point of contact for transliteration or translation.
        - Registrants provide data in any registrar-accepted script and registrars provide transliteration tools to publish in "Must be Present" script.
        - Registrants provide data in any registrar accepted language and registrars provide translation tools to publish in "Must be Present" script.
        - Registrants Provide data in script and language of their choice, but must specify language (locale(?)) of the data. No requirement for transliteration/translation of IRD (no "must-be-present" script).

While not within the remit of this working group to choose what is the best model, we observe there are several important policy questions to consider, noted below.

- Who should bear the cost?
- Who is in the best position to address this issue most effectively?
- ???

We recommend a subsequent policy development effort examine these issues. The WG offer the following suggestions based on past deliberations:

- There are standards exists for transliteration and translation, however translation and transliteration could be inexact in some cases and the quality/accuracy of the data may suffer.

- Registrants can be monolingual. This is intended to highlight the problem of who does the translation or transliteration and what it means to responsibility for quality and compliance.

- ???
- ???

4. RECOMMENDATIONS

Recommendation 1: ICANN staff to develop, in consultation with the community, a data model for domain registration data. The data model should include a formal data schema that incorporates the standards that the working group has agreed on for internationalizing various registration data elements. This data model should also include tagging information for language/scripts.

Recommendation 2: ICANN staff to develop an issues report on internationalizing contact information. The issues report should consider whether it is desirable to translate contact information to English or transliterate contact information to Latin. The issues should consider policy questions raised in this document, it should also determine whether to start a policy development process (PDP).

Recommendation 3: ICANN to bring the issues raised about the limits of Port 43 WHOIS protocol to IETF's attention and seeking work from the IETF to address this issue.

Appendix: Jim's original outline

I am going to propose a path to closure in this message. It represents my understanding of our last meeting on 18 April (transcript has been available for some time), our public meeting during ICANN San Francisco (transcript has been available for some time), and a few private conversations I have had since and between those two meetings.

I am submitting this proposal as an individual. I welcome discussion on its merits and completeness both in the meeting on 16 May and on this mailing list.

Speaking as co-Chair, I am going to press to move forward with this plan, incorporating feedback and suggestions from our discussion in our next meeting on 16 May as well as any future discussions in meetings and on this mailing list. I will interpret silence as agreement with this plan and its evolution.

In the rest of this message I am speaking as an individual.

As a reminder, the mission of this working group is as follows:

The IRD-WG shall study the feasibility and suitability of introducing display specifications to deal with the internationalization of Registration Data.

In our interim report we have evolved 4 models and we sought community input on the efficacy of the those models. We did get a few well reasoned comments but it is fair to say that we did not receive anything close to a community consensus on how to choose between the models. I would like to propose something different than choosing between the 4 models, which we discussed during our last meeting.

In my opinion, the models are trying to address the problem of executing translation and transliteration. Model 1 is status quo, i.e., we stick with the system we have and require US-ASCII to be present at all times. The other models distribute the translation and transliteration services in various ways. I do not think we need to solve this problem. I think we identify this as the problem that needs further study.

Specifically, I suggest the outline below for our final report.

This is an expanded outline insofar as I try to say a bit about what I would expect to be in each section. It is probably not explained as well as it could be but I do hope it gets the point across. I did not want to make this message any longer than it already is. I also was not trying to write the report since I do want some discussion about this approach first.

The model for the outline is we state what we have, we make some observations about what we have, and we propose further study of a few specific issues.

1. INTRODUCTION - Mostly boilerplate information including problem statement and details about the formation of this group. We can re-purpose a great deal of what is in the interim report.

2. BACKGROUND - This should include all the facts we need to support our findings. Most of this is in our interim report.

a. what we know various registrars and registries are doing today to support the display of internationalizes data.

b. what we know about the existing WHOIS protocol.

c. what we know about the definition of registration data.

d. what we know about where different registration data elements are collected, stored, managed, and displayed.

3. INTERNATIONAL STANDARDS - This could be a part of the background information but my current thinking is that it is better to elevate to a major section. In this section we summarize all the international standards and standard practices that exist for internationalizing the various elements of existing registration data. Most of this is in our interim report.

4. FINDINGS - In this section we list the conclusions we can draw from all the facts stated previously.

a. WHOIS is insufficient. It has no structure and hence no method of signaling encodings.

a.1. Registration data has multiple purposes and internationalization requirements are different depending on the purpose. To the extent the data is already represented in XML, e.g., within EPP between registrars and registries, internationalization is primarily ensuring the data is properly tagged with the script that is in use.

a.2. The lack of structure in WHOIS excludes any signaling mechanism, thus the data can not be correctly tagged and further it can not be correctly displayed.

a.3. There are recognized standards for internationalizing many of the elements of registration data but in many cases the data would need to be translated or tranliterated for use with the current WHOIS.

b. Registrants are monolingual. This is intended to highlight the problem of who does the translation or transliteration and what it means to responsibility for quality and compliance.

c. Quality of data is not a well defined phrase. Registrants are expected to provide high quality data but can it be verified? Even if could what happens to the quality after translation and transliteration and who is responsible for that?

d. Registration data is itself undefined. WHOIS services do vary. WHOIS requirements vary between registrars and registrants as evidenced by the contracts.

4. RECOMMENDATIONS

a. Seek a plan to define registration data, who collects it, who stores it, who is responsible for it, and specify its purpose.

b. Seek a plan to replace WHOIS. In other words, although the data can probably be internationalized, displaying it is problematic with the current system. This study would need to consider if registration data should be translated or transliterated, who should do it, what it means to the overall registration data infrastructure, and what it means to the quality of the data.

c. As an interim solution, given the continued use of WHOIS, as much as possible, all parties in the lifecyle of the registration data should adopt the international standards noted above for registration data elements wherever they can.