Final Report of the ICANN Internationalized Registration Data Working Group

7 May 2012

Table of Contents

1.   Executive Summary

On 26 June 2009 the Board of Directors of the Internet Corporation for Assigned Names and Numbers (ICANN) approved a resolution (2009.06.26.18) requesting that the Generic Names Supporting Organization (GNSO) and the Security and Stability Advisory Committee (SSAC), in consultation with staff, convene an Internationalized Registration Data Working Group to study the feasibility and suitability of introducing display specifications to deal with the internationalization of registration data.[1] Subsequently, the SSAC and the GNSO formed the working group (hereinafter the IRD-WG) to study the issues raised by the ICANN Board.

In November 2010 the IRD-WG produced an Interim Report requesting community input on several questions relating to possible models for internationalizing Domain Name Registration Data.[2] On 03 October the IRD-WG posted a draft Final Report in the Public Forum for comment for 45 days. After considering the comments received, this final Report is the IRD-WG's response to the Board request regarding the feasibility and suitability of introducing display specifications to deal with the internationalized registration data.

2.   Introduction

With the increasing use of the Internet in all geographic regions and by diverse linguistic groups, the demand for a multilingual Internet has intensified. To satisfy the demand, many Internet applications are now able to accept and to display characters from a broad range of languages, character sets and scripts.

Access to domain name registration information (often called WHOIS data) is provided by several applications. Accommodating the submission and display of internationalized registration data is seen as an important evolutionary step for WHOIS services. The following statement from the Internet Architecture Board (IAB), found in Request for Comment (RFC) 4690, summarizes the issues associated with this evolution:

> "In addition to their presence in the DNS, IDNs introduce issues in other contexts in which domain names are used. In particular, the design and content of databases that bind registered names to information about the registrant (commonly described as "WHOIS" databases) will require review and updating. For example, the WHOIS protocol itself [Daigle 2004][3] has no standard capability for handling non-American Standard Code for Information (non-ASCII) text: one cannot search consistently for, or report, either a DNS name or contact information that is not in ASCII characters. This may provide some

---

[1] See ICANN Board Resolutions, 26 June 2009, "Display and Usage of Internationalized Registration Data," <http://www.icann.org/en/minutes/resolutions-26jun09.htm#6>.
[2] See <http://gnso.icann.org/issues/ird/ird-wg-final-report-15nov10-en.pdf>.
[3] L. Daigle, "RFC 3912: WHOIS Protocol Specification," Network Working Group, Internet Engineering Task Force, Internet Society, September 2004, <http://www.ietf.org/rfc/rfc3912.txt>.

additional impetus for a switch to IRIS but also raises a number of other questions about what information, and in what languages and scripts, should be included or permitted in such databases."[4]

The SSAC also called attention to these issues in SAC037, *Display and usage of Internationalized Registration Data, Support for Characters from Local Languages or Scripts*.[5] In this report, the SSAC recommended that the ICANN Board of Directors 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.

2.1   IRD-WG Objectives and Membership

Text of the Board Resolution [6]

"Whereas, ICANN has been working towards the introduction of Internationalized Domain Names (IDN) with the gTLD and ccTLD communities.

Whereas, support for characters from local languages in domain name registration submission and display is an issue that affects many communities across the GNSO, CCNSO, ALAC and GAC.

Whereas, while standard formats are defined for domain labels, no standard format is required for elements of a domain name registration record (Registration Data), such as contact information, host names, sponsoring registrar and domain name status.

Whereas, members of the community with knowledge and expertise in these areas have identified topics of inquiry in the display and usage of Internationalized Registration Data, including applications and Internet user experience, data reliability, accuracy and operational issues, and security and standardization issues. See:

- SAC037 "Display and usage of Internationalized Registration Data" (21 April 2009) <http://www.icann.org/committees/security/sac037.pdf>;
- SAC033 "Domain Name Registration Records and Directory Services" (22 July 2008) <http://www.icann.org/committees/security/sac033.pdf>;
- SAC027 "Comment to GNSO regarding WHOIS Studies" (7 February 2008) <http://www.icann.org/committees/security/sac027.pdf>.

---

[4]J. Klensin and P. Fältström, "RFC 4690: Review and Recommendations for Internationalized Domain Names (IDNs)," Network Working Group, Internet Engineering Task Force, Internet Society, September 2006, <http://www.ietf.org/rfc/rfc4690.txt>.
[5]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>.
[6] See ICANN Board Resolutions, 26 June 2009, "Display and Usage of Internationalized Registration Data," <http://www.icann.org/en/minutes/resolutions-26jun09.htm#6>.

Whereas, the Board recognizes that discussion and resolution of these issues would be beneficial to the introduction of Internationalized Domain Names.

Resolved (2009.06.26.18), the Board requests that the GNSO and SSAC, in consultation with staff, convene an Internationalized Registration Data Working Group composed of individuals with knowledge, expertise, and experience in these areas to study the feasibility and suitability of introducing display specifications to deal with the internationalization of Registration Data.

The Board directs the Internationalized Registration Data Working Group to solicit input from interested constituencies including ccTLD operators and the CCNSO during its discussions to ensure broad community input.

The Board further directs staff to provide a dedicated staff person and additional staff resources as staff determines to facilitate the work of the Internationalized Registration Data Working Group.

The IRD-WG interprets the Board's request as two broad issues and related questions:

1) *Suitability issues*:
    o   Is it suitable (or desirable) to internationalize Domain Name Registration Data and address the relevant issues raised in IAB RFC 4690?
    o   What data elements are suitable to be internationalized?

2) *Feasibility issues*:
    o   Is the current WHOIS system capable of handling the query and display of Internationalized Domain Name Registration Data?
    o   What specifications are feasible to deal with Internationalized Domain Name Registration Data?

IRD-WG Membership: Edmon Chung from the GNSO and James Galvin from the SSAC chair the IRD-WG.[7] The IRD-WG members include a total of eighteen participants from gNSO, ccNSO, ALAC and SSAC.[8] There is also some geographic diversity in the membership with participants from China, New Zealand, Pakistan, Peru, Russia, Tunisia, and the USA.

---

[7] The IRD-WG would also like to acknowledge that Jeremy Hitchcock was chair of the WG representing SSAC until December 2010.
[8] 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>.

2.2 Terminology

The term "WHOIS" in the ICANN environment could refer to various components of the WHOIS system. To avoid confusion and bring precision to the discussion, we use the following terms as proposed in SAC 051.[9]

Domain Name Registration Data (DNRD) – refers to the information that registrants provide when registering a domain name and that registrars or registries collect. Some of this information is made available to the public. For interactions between ICANN Accredited Generic Top Level Domain (gTLD) registrars and registrants, the data elements are specified in the current Registrar Accreditation Agreement (RAA). For Country Code Top Level Domains (ccTLDs), the operators of these TLDs set their own or follow their government's policy regarding the request and display of registration information.

Domain Name Registration Data Access Protocol (DNRD-AP) – refers to the elements of a (standard) communications exchange—queries and responses—that make access to registration data possible. For example, the WHOIS protocol (RFC 3912) and HTTP (RFC 2616 and its updates) are commonly used to provide public access to DNRD.

Domain Name Registration Data Directory Service (DNRD-DS) – refers to the service(s) offered by registries and registrars to provide access to (potentially a subset of) the DNRD. ICANN Accredited gTLD registries and registrars are required by contracts to provide the DNRD Directory Services via both port 43 and over the web interface. For ccTLDs the TLD registries (or governments) determine which service(s) they offer.

To ensure that discussions regarding internationalized registration data take place in a consistent manner, the IRD-WG uses the following definitions of IDN-related terms. Some of these terms are from the ICANN's IDN glossary,[10] and others are from related RFCs from the Internet Engineering Task Force (IETF) (RFC 6365[11] and RFC 5890[12]). We note that these definitions are informal ones. Both Unicode and Internationalizing Domain Names in Applications (IDNA) require and use very precise definitions to differentiate among types of objects; these informal ones are not a substitute for those more precise ones, which are given by the referenced documents.

ASCII: a character-encoding scheme based on the ordering of the English alphabet. When mentioned in relation to domain names or strings, ASCII refers to the fact that before internationalization, only the letters a-z, digits 0-9, and the hyphen "-" were allowed in domain

---

[9] ICANN Security and Stability Advisory Committee (2011), SSAC Advisory on Domain Name WHOIS Terminology and Structure (SAC 051). Available: < http://www.icann.org/en/committees/security/sac051.pdf>
[10] ICANN, "IDNs Glossary," Retrieved August 10, 2010, <http://www.icann.org/en/topics/idn/idn-glossary.htm>.
[11] P. Hoffman and J. Klensin.,."Terminology Used in Internationalisation in the IETF", RFC 6365, September 2011..
[12] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.

names. US-ASCII is the Internet Assigned Numbers Authority (IANA) preferred character set name for ASCII.

Internationalized domain names (IDNs)*:* IDNs are domain names that include characters used in local scripts that are not written with the twenty-six letters of the basic Latin alphabet. An IDN can contain Latin letters with diacritical marks, as required by many European languages, or may consist of characters from non-Latin scripts such as Arabic or Chinese.

Internationalized Registration Data (IRD)*:* Where DNRD can be represented in different languages and scripts it is referred to as *Internationalized Registration Data*. Where DNRD contains data other than US-ASCII (not just the capacity for it), it is referred to as *Localized Registration Data*.

Translation**:** The process of conveying the meaning of some passage of text in one language, so that it can be expressed equivalently in another language. <RFC6365>

Transliteration: The process of representing the characters of an alphabetical or syllabic system of writing by the characters of a conversion alphabet. <RFC6365>

Transcription*:* The process of systematically writing the sounds of some passage of spoken language, generally with the use of a technical phonetic alphabet (usually Latin-based) or other systematic transcriptional orthography.   <RFC6365>

A-label | U-label: A domain name consists of a series of "labels" (separated by "dots"). The ASCII form of an IDN label is termed an "A-label." An A-label conforms to the Letter-Digit-Hyphen (LDH) constraint on labels as defined by the DNS standards. All operations defined in the DNS protocol use A-labels exclusively. The Unicode form, which a user expects to be displayed, is termed a "U-label." A special form of "ASCII compatible encoding" (ACE) is applied to a U-label to produce a corresponding A-label. The transformation is symmetric: one can derive a U-label from an A-label for the purpose of displaying the domain name using characters from a local script so that a user sees a familiar script rather than a less recognizable A-label.

Thin | Thick Registry: A thin registry only includes data sufficient to identify the sponsoring registrar, status of the registration, nameserver, creation, and expiration dates for a domain registration. Registrars maintain the complete set of registration data for those domains they sponsor. .COM and .NET are examples of thin registries. Thick registries maintain fields to store and display the registrant's contact information and designated administrative and technical information, in addition to sponsoring registrar and registration status information, with the DNRD usually provided by the sponsoring registrar..INFO and .BIZ are examples of thick registries.

3.   Background

3.1   What is Domain Name Registration Data?

As defined earlier, DNRD refers to the information that registrants provide when registering a domain name and that registrars or registries collect. Some of this information is made available to the public. For interactions between ICANN-accredited gTLD registrars and registrants, the data elements are specified in the RAA.[13] For ccTLDs, the operators of these TLDs set policies for the request and display of registration information.

The RAA (RAA 3.3.1) specifies the following data elements that must be provided by registrars (via Port 43 and via web-based services) in response to a query:

- The Registered Name;

- The names of the primary nameserver and secondary nameserver(s) for the Registered Name;

- The identity of the registrar (which may be provided through registrar's website);

- The original creation date of the registration;

- The expiration date of the registration;

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

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

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

Regarding the data, the IRD-WG also notes that:

- Some registries are subject (in their registry agreements with ICANN) to slightly different requirements regarding which data must be publicly accessible.[14]

- In ccTLDs, the operators of these TLDs set their own or follow their government's policy regarding what constitutes this data and what is to be displayed in the DNRD-DS.

---

[13] See Section 3.3 of Registrar Accreditation Agreement (2009). <http://www.icann.org/en/registrars/ra-agreement-21may09-en.htm#3>.
[14] See for example <http://www.icann.org/en/tlds/agreements/biz/appendix-05-08dec06.htm.

3.2   Where Are Different Registration Data Elements Collected, Stored, and Displayed?

Registrants submit information regarding DNRD to a sponsoring entity. In all gTLDs, that entity is either a registrar or a reseller. In ccTLDs, that entity can be a registry, in addition to a registrar or a reseller.

Collection: Registrants submit this information as part of the process of registering a domain name. The sponsoring entity collects the information.

Storage: The sponsoring entity that receives the domain registration data stores a copy. In addition, the registrar submits a limited subset of the information (domain status, nameserver information) to the registry if the sponsoring registry is a thin registry. If the sponsoring entity is a thick registry, registrars submit a more complete set of the information to the registry as required by the RAA with that registry. Under a data escrow provision in the RAA, ICANN accredited registrars are required to regularly deposit a backup copy of their gTLD registration data with an ICANN approved escrow provider.

Display: End users query the registration data directory service for a domain name, contact information, or nameserver information.

- For thin registries, the query displays the data from the registry's DNRD-DS. Some clients could parse the registry output and continue to query the registrar's DNRD-DS.

- For thick registries, the query displays the data from the registry's DNRD-DS.

3.3   The WHOIS Protocol

The **WHOIS protocol** (RFC 3912)[15] is a client-server, query-response protocol. The client:
   1) Connects to the service host at Transmission Control Protocol (TCP) port 43;
   2) Sends a single "command line;" and
   3) Signals the end of the command line with a <CR><LF> (carriage-return and line-feed) character sequence.

The server listening to port 43:
   1) Accepts and parses the query;
   2) Composes a response, again using a <CR><LF> to signal end of response; and
   3) Returns the response to the client. The server closes its connection as soon as the output is finished. The client then displays the response to the standard output or processes the response as otherwise indicated by the user.

---

[15] L. Daigle, "RFC 3912: WHOIS Protocol Specification," Network Working Group, Internet Enginerring Task Force, The Internet Society, September 2004, <http://www.ietf.org/rfc/rfc3912.txt>.

The protocol imposes no constraints on the data it transports. The only constraint imposed on query and message formats is that they must be terminated using an ASCII line feed (LF) and carriage return (CR) character sequence.

According to RFC 3912, "The WHOIS protocol has not been internationalized. The WHOIS protocol has no mechanism for indicating the character set in use. Originally, the predominant text encoding in use was US-ASCII. In practice, some WHOIS servers, particularly those outside the United States, might be using some other character set either for requests, replies, or both. This inability to predict or express text encoding has adversely impacted the interoperability and, therefore, usefulness of the WHOIS protocol."[16] Finally, the WHOIS protocol does not define a structured data schema.

3.4   Current Practices by gTLD Registries / Registrars and ccTLDs to Support the Query and Display of Internationalized Registration Data

Text requests and content returned by DNRD-DS are historically encoded using US-ASCII. While the WHOIS protocol does not specify US-ASCII as the exclusive character set for text requests and text content encoding, and thus gives latitude with respect to protocol encoding, the protocol specification leaves the method of signalling/selecting character sets as a local implementation matter. The current situation is that no standards or conventions exist for all WHOIS protocol implementations to signal support of character sets other than US-ASCII. DNRD-DS are supported by a large and diverse set of providers for an even larger and more diverse set of users who increasingly want to access DNRD using a familiar script or language; thus, the lack of a signalling convention is problematic.

To support internationalized data, many registries have developed specific conventions.   For example, .org uses the following method to display the internationalised domain name:[17]

- The "domain name" field displays the registered internationalized domain name in Punycode (e.g., xn--probestck-w9a.org). Three additional fields follow at the end of the PIR WHOIS query results:

  - IDN-Tag: The IDN-Tag of the IDN as determined by the registrant (e.g., "de" for German).

  - Unicode Hex: The IDN in Unicode Hex format (e.g., U+0070 U+0072 U+006F U+0062 U+0065 U+0073 U+0074 U+00FC U+0063 U+006B).

  - Unicode HTML: The IDN in HTML entity format (e.g., probestück).

---

[16] L. Daigle, "RFC 3912: WHOIS Protocol Specification," Network Working Group, Internet Enginerring Task Force, The Internet Society, September 2004, <http://www.ietf.org/rfc/rfc3912.txt>.
[17] See "How are IDNs Displayed in WHOIS" (2011) Public Interest Registry. Available at <http://www.pir.org/why/global/idn#q10>.

Thus, if users were using a command line client, they would not be able to see the internationalised data, but if they were using a web browser client, they might be able to see the IDN correctly displayed in the native script.

On the other hand, some ccTLDs have adopted other types of signalling, for example:

- .dk uses "--charset=latin-1" or "--charset=utf-8"
- .no use "-c utf-8"
- .jp uses "/e" to switch from ISO-2022-JP to ASCII.

Finally, the IRD-WG notes that conventions that are adopted for WHOIS/43 clients may not be used or useful for web-based services.

ccTLD Practices

In 2010, ICANN staff informally contacted ccTLD operators whose communities used languages other than English to share their practices to support internationalized DNRD.

Staff solicited current WHOIS submission and display practices from 16 ccTLDs by asking the following questions:

1. Does your registry allow users to register domain names using characters from local scripts?
2. Does your registry collect and store registration data in US-ASCII in addition to characters from local scripts?
3. Can users of the web interface choose the display language? What languages does your web interface support?
4. Does your registry provide access to registration information via WHOIS/port 43? Can users choose the display language?

The responses from the 16 ccTLD operators are summarized as follows:

- 10 ccTLDs allow users to register domain names using characters from local scripts;
- 10 of 16 ccTLDs support "English" (US-ASCII) and a local language/script, including Arabic, Chinese, German, Japanese, Lithuanian, Portuguese, Spanish, Swedish;
- All ccTLDs support WHOIS/Port 43; and
- Character set dependencies affect WHOIS client submission and Display (some uses UTF-8, others UTF-16, or ISO-8859).

4. Findings

In this section, we report the findings of IRD-WG discussions. We organize the findings as responses to the Board's questions.

4.1   Is It Suitable To Internationalize Domain Name Registration Data?

Much of the currently accessible domain registration data is encoded 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. Many millions of domain name registrants and Internet users do so today, even though their primary language is written using a different script. This condition is also convenient for registrants, registrars and registries, and the installed base of operational WHOIS services that display US-ASCII.

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.[18] It is important to note that the latter (underserved) community is likely to continue growing and could outnumber the former in a matter of years. Many registrants are monolingual, which is the expectation and motivation behind internationalized domain names.[19] Therefore, it is unreasonable to assume all of them will know or will be able to enter the registration data in languages other than their local language. Thus, it is desirable for registrants to be able to submit DNRD in character sets other than ASCII.

However, this desirability should be balanced against other uses of the data. While domain registrants 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 globally.[20]

4.2   What Data Elements Are Suitable To Be Internationalized?

The registration data required by RAA can be broken down into the following categories:

- Domain name;
- Sponsoring registrar;
- Registration status;
- Nameserver information;
- Names (e.g. owner, admin, technical contact);
- Postal addresses (owner, admin, technical contact postal information);
- Phone/fax numbers (e.g. admin, technical, owner phone/fax);
- Dates (e.g. creation date, expiration date, update date); and
- Email addresses (owner, admin, technical contact email).

The IRD-WG members agreed that various elements of registration data could be separately internationalized, including the following:

---

[18] See usages identified in Section 3 of SAC 023, available at
<http://www.icann.org/en/committees/security/sac023.pdf >
[19] See S. Hussain comment on the Interim Draft Report. <http://forum.icann.org/lists/ird-wg-report/msg00000.html>.
[20] See R. Rassmussen comment on the Interim Draft Report <http://forum.icann.org/lists/ird-wg-report/msg00004.html>.

Domain names (RAA 3.3.1.1): The IRD-WG recommends that WHOIS services should return both A-label and U-label representation for the given IDN domains queried.[21]

Sponsoring Registrar (RAA 3.3.1.3): The IRD-WG recommends that this data element should be in ASCII to aid law enforcement and intellectual property investigations, and to the extent possible, make it available in local languages and scripts. It is important to note that ICANN's application for registrar accreditation requires applicants to submit a transliteration of "any legal name, street, electronic or mailing address which is not in Latin characters."[22]

Nameserver names (RAA 3.3.1.2): Currently all nameservers are in US-ASCII. However, with IDNs, it is possible that some registrants will compose nameserver names using IDN labels. Several alternatives exist:

1. Always display the nameserver name in US-ASCII 7 using the A-label. A supporting argument for this choice is that nameserver name information is generally only of technical interest and should be displayed in same way as it is in the DNS.

2. Display nameserver names in both A-label and U-label (to the extent such information is available). This is consistent with the recommended treatment of the domain name.

The IRD-WG recommended that this field should continue to be displayed in A-label and, to the extent possible, be displayed in the corresponding U-label.

Telephone/Fax (RAA 3.3.1.7,8): The IRD-WG recommended that the International Telecommunications Union Telecommunication Standardization Sector (ITU-T) Recommendation E.123 could be used to internationalize telephone and fax, specifically using the international notation (+31 42 123 4567). [23]

Email address (RAA 3.3.1.7,8): The Email Address Internationalization Working Group in the IETF has introduced a series of specifications that define mechanisms and protocol extensions needed to fully support internationalized email addresses. These specifications recently became Internet Standards (RFC 6530, 6531, 6532, 6533). As such, the IRD-WG recommended that the email address field should conform to these new Internet Standards, specifically RFC 6531, where the syntax of the internationalized email address is specified.

---

[21] IRD-WG members discussed query and display of IDN variants labels in the context of WHOIS. They noted that it is outside the scope of the IRD-WG to define variants or discuss how different languages handle variants. However, the IRD-WG members agree that a WHOIS service query of IDNs variant labels delegated into the DNS should return the domain of which it is a variant in its response, as well as an indication that the label queried is a variant of the original domain. The IRD-WG members agree that this should be consistent across Whois services. The IRD-WG members also agree that defining a Whois service query of a reserved or blocked variant returns is a matter of local policy.

[22] ICANN Registrar Accreditation Application, <http://www.icann.org/en/registrars/accreditation-application.htm>.

[23] International Telecommunications Union Recommendation E.123 <http://www.itu.int/rec/T-REC-E.123/en>.

Dates (RAA 3.3.1.4,5): This includes creation date, expiration date, and update date of the domain. The IRD-WG members discussed this issue and proposed International Organization for Standardization (ISO) 8601-2004[24] as the standard to be used. Internet applications commonly represent the date and time as specified in RFC 3339, which includes a profile of the ISO 8601 standard. As a result, the WG recommend that date and time elements should conform to formats specified in [RFC3339], and represented in UTC with no offset from the zero meridian. For example, 1990-12-31T23:59:60Z.

Registration Status: Registrars and registries often provide the status of the domain registration. The IRD-WG identified several alternatives as follows:

1. Return the status in a US-ASCII representation of the registrar's choosing;

2. Publish the exact Extensible Provisioning Protocol (EPP) status code and leave it to the clients to decide whether to localize or not;

3. Identify a more easily understood representation;

4. Publish the easily understood representation in mandatory and local character sets; or

5. Any combination of these approaches.

The IRD-WG members discussed different opinions and chose option 2, since it gives client the ability to localize this field. Option 2 is also used in the new gTLD Applicant Guide Book .[25] The EPP status codes are described in RFCs 5730-5734, RFC 3735, 3915, and 4310.[26] Finally, the IRD-WG also recommends that ccTLDs that use EPP display it in a similar way.

Entity Names and Address (RAA 3.3.1.6,7,8): This includes names and addresses of registrants, administrative contacts, and technical contacts. In line with the recommendations in section 5.1, the IRD-WG agreed that registrants should be able submit their data element in the user's local language and script. Additionally, the IRD-WG discussed whether discussed whether it is desirable to adopt a "must be present" representation of contact data, in conjunction with local script support for the convenience of local users. This will be covered in detail in the next section.

The table below summarizes the discussions in this section.

Fields	Suitable to Internationalize?	Possible Standards
Domain Names	Yes	Both A-label and U-label

---

[24] See http://www.iso.org/iso/catalogue_detail?csnumber=40874
[25] ICANN, "New TLD Program Application Guidebook," <http://www.icann.org/en/topics/new-gtlds/dag-en.htm>.
[26] Also see a discussion paper by CentralNic regarding these codes. Available at < https://www.centralnic.com/company/labs/epp/status >

Nameserver Names	Yes		A-label, and optionally U-label
Sponsoring Registrar	NO		US-ASCII
Telephone/fax	YES		ITU-T E.123
Email	YES		RFC 6531
Registration Status	N/A		Exact EPP status where applicable
Dates	YES		In UTC as specified in RFC 3339 with no offset from the zero meridian
Entity Names and Addresses	YES		Local languages/scripts

4.3   Is It Suitable To Support The Translation Or Transliteration of Entity Name and Contact Information Into a Single Script / Language?

To balance the needs and capabilities of the local registrant with the need of the (potential) global user of this data, one of the key questions the IRD-WG members discussed is whether DNRD-DS should support multiple representations of the same registration data in different languages or scripts. In particular, the IRD-WG members discussed whether it is desirable to adopt a "must be present" representation of contact data, in conjunction with local script support for the convenience of local users.

In general, the IRD-WG recognizes that the internationalized contact data can be translated or transliterated into the "must be present" representation. As defined in RFC 6365, *Translation* is the process of conveying the meaning of some passage of text in one language, so that it can be expressed equivalently in another language. *Transliteration* is the process of representing the characters of an alphabetical or syllabic system of writing by the characters of a conversion alphabet.

Based on the definition above, and consistent with the current state of domain name registration data, if transliteration were desired, then the "must be present" script would be the Latin script. If translation were desired, then the "must be present" language would be English.

The IRD-WG identified four options:

1. Registrants provide domain contact data in "Must Be Present" script in addition to their local language. Registrars and registries will display both in the DNRD-DS.

15

2. Registrants provide data in any registrar-accepted script and registrars provide point of contact for transliteration or translation.
3. Registrants provide data in any registrar-accepted script and registrars provide transliteration tools to publish in "Must Be Present" script.
4. Registrants provide data in any registrar-accepted language and registrars provide translation tools to publish in "Must Be Present" script.

Appendix B describes these options in detail. A respondent proposed a fifth model in the public forum on the draft Final Report. This proposed model is that registrants would provide domain contact data in any registrar-accepted script, and also optionally could provide the translated/transliterated data in the English language/Latin script.

The IRD-WG examined these models and their impacts on registries, registrars, and end users. In the end, the IRD-WG could not reach a consensus on which model to choose for several reasons:

- 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:

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

    o 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.

    o 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 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[27] and the U.S.

---

[27] 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 ->

Library of Congress

- o However, it is unclear whether transliteration would meet the community's need for the data. For example, 'street' in Chinese is '街'(U+8857) or '路' (U+8DEF). It would be transliterated into "jie" or "lu." However, a non-native Chinese speaker would have no idea what those transliterated words mean. Fundamentally, from a usability point of view, a transliteration system would generate correct strings according to transliteration standards, but nevertheless linguistically nonsensical strings from human perspective.

- o Furthermore, 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.

- Due to its lack of expertise on these issues as well as lack of participation from key stakeholders the IRD-WG could not tackle two important policy questions:

  - o How are contact names and addresses used? This will determine whether the information loss and consistency issues associated with translation/transliteration is acceptable.

  - o If translation or transliteration is desired, who should bear the burden? Should it be the registrant, registrar, or registry? IRD-WG notes that from a policy perspective, those who are in the best position to address this issue in the most cost-effective manner should bear the burden.

The IRD-WG would like to recommend an Issue Report on this subject. In the interim, the IRD-WG recommends that once an alternative DNRD Access Protocol that meets the needs of internalization is adopted (see section 4.5), registrars who accept registration data in local scripts (non US-ASCII) should make registration data in the local script available through DNRD-DS output, and should at least include tags to identify languages and scripts (e.g. RFC 5646). Thus, those who need to translate, transliterate or transcribe this information would at least know what language and script the contact information is in.

4.4 Is It Suitable To Introduce Display Specifications To Address Internationalized Domain Name Registration Data?

The IRD-WG agrees that it is suitable to introduce display specifications to address internationalized DNRD. Specifically, while standard formats are defined for domain labels, no

---

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.

standard format is required for elements of a domain name registration record (Registration Data), such as contact information, host names, sponsoring registrar and domain name status.

The IRD-WG concluded that the community would benefit from a standard registration data schema (for example in Extensible Markup Language (XML)):

- A formal data schema for registration data (for example in XML) would enable end-user clients to better localize the data label.

- A formal data schema for registration data with language tag information would allow better processing of the data.

- In some cases, it is possible that registrars may allow for multiple languages or scripts in the contact data (for example, e.g. Arabic speaking registrants living in the United States to put their names in Arabic language, but their address in English). If this is needed, the language-tag data needs to be at a data-element level.

4.5   Is the Current WHOIS System Capable of Handling the Query and Display of Internationalized Registration Data?

The IRD-WG agreed that the current WHOIS system is not capable of handling the query and display of internationalized registration data, but determined that there are some workarounds and local conventions that can permit exceptions.   In particular, operators that currently provide DNRD-DS using the WHOIS protocol do not support character sets other than US-ASCII in a standard manner, which could present problems if registrants wish to query DNRD in character sets other than US-ASCII. According to RFC 3912, the WHOIS protocol "lacks [a] mechanism for indicating the character set in use … This inability to predict or express text encoding has adversely impacted the interoperability and, therefore, usefulness of the WHOIS protocol."

In absence of protocol specification, various registries/registrars have adopted ad hoc solutions to support internationalized DNRD. Continued deployment of ad hoc solutions will inevitably lead to widespread inconsistent user experience and interoperability issues, greatly reducing the utility of DNRD.

Thus, these issues need to be addressed urgently. The IRD-WG encourages the ICANN community to identify, evaluate and adopt an alternative DNRD Access Protocol that would meet the needs of internationalisation.

4.6   Is it Feasible To Introduce Submission and Display Specifications to Address Internationalized Registration Data?

The IRD-WG agreed that it is feasible to introduce submission and display specifications to address internationalized registration data. In particular, most elements of the DNRD have existing standards that apply to them, and to the extent possible, the IRD-WG recommend those

be considered. In particular, the following schemas/templates have been worked on in the past and should be considered:

-   EPP RFC 5730-5734
-   DREG RFC 3982
-   UPU S.42 (address templates)

5. Recommendations

Recommendation 1: To introduce submission and display specifications, ICANN staff should develop, in consultation with the community, a data model for domain registration data. The data model should specify the elements of the registration data, the data flow, and 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: The GNSO council and the SSAC should request a common Issue Report on translation and transliteration of contact information. The Issue Report should consider whether it is desirable to translate contact information to a single common language or transliterate contact information to a single common script. It should also consider who should bear the burden and who is in the best position to address these issues. The Issue Report should consider policy questions raised in this document and should also recommend whether to start a policy development process (PDP).

Recommendation 3: ICANN staff should work with the community to identify a DNRD Access Protocol that meets the needs of internationalization, including but not limited to the work products resulting from recommendations 1 and 2, and the requirements enumerated in this report.

Recommendation 4: ICANN should take appropriate steps to require gTLD registries and registrars and persuade ccTLD registries and registrars to support the following standards: Domain Names - both A-label and U-label; nameserver Names- A-label, and optionally U-label; Telephone/fax- ITU-T E.123; Email-RFC 6531; Registration Status- Exact EPP status where applicable; Date and time - in UTC as specified in [RFC3339], with no offset from the zero meridian..

Appendix A: IRD Members and Staff Support

IRD-WG members: (sorted by last name)

Erick Iriarte Ahon - ALAC, LACTLD
Scott Austin -- GNSO Intellectual Property Interests Constituency, Commercial
Stakeholder Group, Gordon & Rees LLP
Edmon Chung -- GNSO Registry Stakeholder Group, .ASIA
Jay Daley --- GNSO Registry Stakeholder Group, .NZ
Rafik Dammak -- GNSO Non-Commercial Users Stakeholder Group, Tunisia
Avri Doria -- GNSO Non-Commercial Users Constituency, Non Commercial Stakeholder
Group
Jim Galvin -- SSAC, Afilias
Chuck Gomes -- GNSO Registry Stakeholder Group, VeriSign
Jeremy Hitchcock, GNSO Registrar Stakeholder Group, DYN-DNS
Sarmad Hussain, SSAC
Bob Hutchinson
Yao Jiankang, GNSO Registry Stakeholder Group, CNNIC
Andrei Kolesnikov -- Nominating Committee Appointee, .RU
Mark Kosters – SSAC, ARIN
Steven Metalitz -- GNSO Intellectual Property Interests Constituency, Commercial
Stakeholder Group
Ram Mohan, Board Liaison, SSAC, Afilias
Owen Smigelski, Sunrider International, Intellectual Property Interests Constituency,
Commercial Stakeholder Group
June Seo, GNSO Registry Stakeholder Group, VeriSign

ICANN Staff Support: (sorted by last name)

Francisco Arias, Registry Technical Liaison
Liz Gasster, Senior Policy Counselor
Julie Hedlund, Director, SSAC Support
Dave Piscitello, Senior Security Technologist
Steve Sheng, Senior Technical Analyst, Policy

Appendix B: Different Models that IRD-WG Considered for Internationalizing Contact Information.

Model 1: Provide Directory Service Data in "Must Be Present" Script

Model 1 requires registrants to provide their directory service data in a "Must Be Present" script such as US-ASCII. The registrars have the option of asking registrants to provide their contact information in a local script. If registrants also provide information in their local script, then this information is displayed. Many IRD-WG members thought that that Model 1 was feasible because it would have the least potential impact on registrars and registries. However, they also thought that it would provide the fewest benefits for internationalized registration data since local language display is optional. Figure 1 illustrates this model.

```
$ whois -h idnwhois.registrarX.ru    жук.рф
$ whois -h idnwhois.registrarX.ru   XN--F1AI0A.XN--P1AI

% Registrar X WHOIS server
% This query returned 1 object

domain:        жук.рф
domain-ace:    XN--F1AI0A.XN--P1AI
domain-variant:
domain-v-ace:
contact:       Petr Ivanov (Петр Иванов)
organisation:  OSC «Cicle»
address:       Office 1, Lenin st., Kovrov
address:       Vladimir region, 601900
address:       Russia
phone:         +7 49232 48720
fax-no:        +7 49232 48722
e-mail:        cicle@cicle.ru
```

Figure 1: Model 1 for displaying contact information. In this model, registrants provide data in US-ASCII, and optionally in local script. The registrars display the information both in Cyrillic and US-ASCII.

Model 2: Provide Data in Registrar-Accepted Script and Point of Contact

In Model 2, registrants provide their registration data in a script that can be accepted by the registrar, and registrars provide a point of contact for transliteration and abuse issues on request. The registrars will also forward the same information to the registry. Many IRD-WG members also thought Model 2 was feasible. However, some IRD-WG members wondered whether this

model would create inaccuracies. For example, in this model, registries may not verify the validity of the scripts they receive from registrants and may not take responsibility for the accuracy of the information. If the verification of the script is not performed, it is possible that an entry that combines Cyrillic, simplified Chinese, and Indic scripts could be created as a valid WHOIS entry. In addition, some IRD-WG members were wary of any solution that relies upon registrar provision of a point of contact, whether to the public or to registrants. Figure 2 illustrates this model.

```
$ whois -h idnwhois.registrarX.ru    жук.рф
$ whois -h idnwhois.registrarX.ru   XN--F1AI0A.XN--P1AI

% Registrar X WHOIS server
% This query returned 1 object

domain:        жук.рф
domain-ace:    XN--F1AI0A.XN--P1AI
Registrar:     RU-CENTER LLC
Registrar POC: http://nic.ru
phone:         +7 800 234-5689
fax-no:        +7 800 234-5699
email:         info@nic.ru
contact:       Петр Иванв
organisation:  ОАО Циркуль
address:       ул.Ленина, офис 1, г.Ковров
address:       Владимирская обл. 601900
address:       Россия
phone:         +7 49232 48720
fax-no:        +7 49232 48722
e-mail:        cicle@cicle.ru
```

Figure 2: Model 2 to display contact information. Registrants in this model provide localized information and registrars provide a point of contact to respond to translation issues.

Model 3: Provide Data in Any Script Accepted by the Registrar; Registrar Provides Transliteration Tools to Publish in "Must be Present" Script

In Model 3, registrants provide their registration data in any script accepted by the registrar, and registrars provide tools for publishing the data in a "must be present" script. Many IRD-WG members raised concerns that Model 3 would incur added costs to registrars to produce transliterations. In addition, some IRD-WG members thought that transliteration would not be accurate enough to benefit law enforcement or intellectual property enforcement. Moreover, other members thought that Model 3 represents added value and that the focus on policy should be on baseline behavior, not on added value. Finally, some IRD-WG members were wary of any solution that relies upon registrar provision of transliteration services, whether to the public or to registrants. Figure 3 illustrates this model.

```
$ whois -h idnwhois.registrarX.ru    жук.рф
$ whois -h idnwhois.registrarX.ru   XN--F1AI0A.XN--P1AI

% Registrar X WHOIS server
% This query returned 1 object

domain:       жук.рф
domain-ace:   XN--F1AI0A.XN--P1AI
contact:      Petr Ivanov
organisation: OAO «Tsirkul»
address:      Office 1, Ulitsa Lenina, Kovrov
address:      Vladimirskaya oblast, 601900
address:      Rossiya
phone:        +7 49232 48720
fax-no:       +7 49232 48722
e-mail:       cicle@cicle.ru
```

Figure 3: Model 3 to represent contact information. In this model, registrants provide information in local language, and registrars *transliterate* registrants' submission and display them in WHOIS.

Model 4: Provide Data in Any Script Accepted by the Registrar; Registrar Provides Translation Tools to Publish in "Must be Present" Script

In Model 4, registrants provide their registration data in any script accepted by the registrar, and registrars provide tools for translating and publishing the data in a "must be present" language. Many IRD-WG members raised concerns that Model 4 would incur added costs to registrars as they produce translations. In addition, some IRD-WG members thought that translation would not be accurate enough to benefit law enforcement or intellectual property enforcement. Moreover, other members thought that Model 4 represents added value and that the focus on policy should be on baseline behavior, not on added value. Finally, some IRD-WG members were wary of any solution that relies upon registrar provision of translation services, whether to the public or to registrants. Figure 4 illustrates this model.

```
$ whois -h idnwhois.registrarX.ru     жук.рф
$ whois -h idnwhois.registrarX.ru   XN--F1AI0A.XN--P1AI

% Registrar X WHOIS server
% This query returned 1 object

domain:        жук.рф
domain-ace:    XN--F1AI0A.XN--P1AI
domain-variant:
domain-v-ace:
contact:       Petr Ivanov
organisation:  OSC «Cicle»
address:       Office 1, Lenin st., Kovrov
address:       Vladimir region, 601900
address:       Russia
phone:         +7 49232 48720
fax-no:        +7 49232 48722
e-mail:        cicle@cicle.ru
```

Figure 4: Model 4 to represent contact information. In this model, registrants provide information in their local language, and registrars *translate* registrants' submission and display them in WHOIS.