Summary of email discussions from IRD working group
Prepared by Steve Sheng
1/25/2010

Background: Per Edmund and Jeremy's request, Dave Piscitello provided a list of discussion topics that have been raised in past IRD meetings. There have been lively email exchanges on this topic. This document summarizes these discussions, it also provides some questions for the IRD working group to consider.

To reduce the number of pages, opinions of the committee members are summarized. The original comments can be found in an accompanying document (summary of email exchanges.doc)

Q1a) What do we require from internationalized registration data that a user can submit or have a domain name displayed in the IDN A-label (xn--) format or U-label (local language readable) format?

Definition of "submit": a user or an application would submit a string or domain name a) to query the availability of a label in a given TLD, or b) to query a Whois web, command line or application for the registration information associated with a domain name, or c) to register a domain name via a web form at a registrars website.

Definition of "display": render a visual response to a human on a computer or a machine-readable response to an application or automation, for all of the purposes above.

Response from working group members:

1) For both submission and display, I suggest we only look at WHOIS, with the understanding that there may be different UIs to WHOIS - (command line, web, application). (Jay)
2) We don't need to consider scenario C of submission at all because registrars/registries will just do it. (Jay)
3) There is value in offering users with the ability to "use" U-label or A-label as they choose. Users may most often prefer a U-label since this is more visually recognizable and familiar than "xn--<g1b63r1sh>" strings, but users of a command line who might want to submit and display A-labels when using a command line, especially in situations where he is writing a script and the input or output of a whois command would be "piped" to/from another command (sed/awk/grep). (Dave)
4) Can we start with some axioms that 1) WHOIS must accept a "submit" in either U- or A-label; and 2) WHOIS must "display" both in U- and A-label. (Jay)
5) I prefer U-label. The default format should be U-label. Many "xn---weqasdf asdf" format name is not valid IDN. Based on IETF IDNA standard, whether A-label (xn--) format or U-label is submitted, we should use some algorithm to confirm whether it is a valid IDN.(Jiankang)

Questions for IRD working group members to consider:

1) For question q1a) Would the axioms “that 1) WHOIS must accept a "submit" in either U- or A-label; and 2) WHOIS must "display" both in U- and A-label” close the case?

Q1b) What do we require from internationalized registration data that registration data be extensible to accommodate users who would benefit from the ability to submit and have registration information displayed in "familiar" characters from local languages and scripts?

Response from working group members:

1. Yes, that’s why we use internationalized Whois. (Jiankang)
2. Various elements of registration data could be separately internationalized, and we need to address these elements individually. (Jay)

    a. **domain names:** Standards exist for representation of domain names in U- and A-labels. These are sufficient and thus out of our scope. (Dave)

    b. **registrar:** The registrar is clearly a special case. It may be represented by a code that is then cross-referenced against a different list. (Jay) It's been proposed that the sponsoring registrar name should always be displayed in machine-readable form (meaning, US-ASCII7 subset of the Latin-1 character set). The rationale offered for this is that applications and automation use the sponsoring registrar as a search element in databases of registrar contacts and that these are largely ASCII encoded.

    c. **entity names** include registrant, admin contact name, tech contact name.

    d. **Postal addresses:** could adhere to the conventions the UPU establishes.(Dave), but would strongly recommend that we *do not* establish our own standard. (Jay)

    e. **Email addresses:** could adhere to RFC822-MIME conventions or the more up to date RFCs (4952, 5336).

    f. **Telephone numbers:** Could use the ITU telephony convention. (Dave) The ITU telephony convention is ubiquitous. Recommend that the working group consider this one solved by reference to E.123 internationalised notation for telephone numbers. (http://en.wikipedia.org/wiki/E.123) (Jay)

Questions for IRD working group to consider:

2) Do you agree separating the various elements of registration data and we need to address their internationalization individually?
3) Would you agree with the categories and how they are addressed?
4) Should registrar be represented by a code that is then cross-referenced against a different list? Should it remain in ASCII?
5) How should the entity names be internationalized?

Q1c) Does Q1b imply several representations of the same registration data, but in different languages or scripts?

Responses from Working group members:

1. No, it means that the local languages are compulsive. (Jiankang)
2. It doesn't have to but intellectual property concerns and law enforcement would prefer that. (Jay)
3. To answer the question properly we need to develop some categorisation around constraints. In all cases I assume the registrar is happy with billing details they have for registrant, which may be different from the registration data.(Jay)

    a. Scenario 1: Entirely local registration in a gTLD or ccTLD. In this category the constraints are: registrant is based in country C (meaning the registrant is from country C); data in script S which is a normal script for C; data is meaningful in language L, a normal language of that country, and registrar can read script S and language L.
        i. In this case, it is hard to see any primary reason why the registration data must be in any other script/language as well. Obviously there are secondary reasons - law enforcement, intellectual property protection, and consumer awareness. (Jay)
        ii. The above model could work in a thin registry as well, but is the consequence that much of the registration data are locally understood but potentially unusable by some or all non-local users? (Dave)
            1. "This will always be the case because there is no global language/script. Insisting all data is in both the local language and ASCII will not make it universally usable. There are plenty of people who will not understand either." The same fact shows where the problem is with insisting on local + ASCII - if someone lives in country C and only speaks L and writes it in S then how can they register a domain if they are required to also give the data in ASCII? Either they enter rubbish, expect the registrar to do it or give up, none of which are acceptable. (Jay)

            2. Other contends that "As of today, ascii is the universal method of data representation in whois. […] As for law enforcement, legal guys, tax inspectors and other local/international civil and governmental organizations - they read and understand information in ascii, educational level allows it. So I believe, that universal ascii representation of domain properties (as discussed in this thread) must be required. To better serve local community, data can be represented in local language/script.

        iii. There is no technical restrictions for web representation and few problems with traditional command string (pipe, scripts, etc) for

various *ix platforms. These problems must be resolved due to IDN era."

b. Scenario 2: Local registration through non-local registrar (by which I mean a registrar that cannot speak S or L) in a gTLD. In this category the constraints are: registrant is based in country C; data in script S which is a normal script for C ; and data is meaningful in language L, a normal language of that country. In this case we can assume that any problems with the data cannot be understood by the registrar and so requesting the data also in English/ASCII is a prudent move, and having the WHOIS record show both is also prudent. But then we are treating different registrars differently.
   i. Given that a very large percentage of gTLD registrations are placed through fewer than 10 registrars, this is either likely to be the preponderant case, or registrars with strong incentives to attract and accommodate customers based "anywhere" would adopt your (1) above; specifically, the registrar would seek to "read script S and language L (so they understand the data they are supplying)", no?

c. Scenario 3: Non-local registration in a gTLD. In this category the constraints are: registrant is in country C, data is in script S which is NOT a normal script for C. We would then have to question why this combination should be allowed. If it is not allowed then we start to get into the business of connecting scripts to countries and enforcing that linkage at data entry, and even if this were disallowed, scenario 1 and 2 seem to present significant enough challenges. I recommend this be disallowed. My reason for disallowing this is whilst allowing (1) and (2) is ultimately about the reasonable expectations of end users

d. Scenario 4: Non-local registration in a ccTLD. In this category the constraints are: registrant is in country C which is different from the country of the ccTLD, data is in script C which is a normal script for C but NOT a normal script for the ccTLD It would seem quite appropriate in this case for the registry to insist the data is also supplied in a local script of the ccTLD. In this case, registrars has a second set of conditions beyond (a); for example, if registrar can read script S and language L AND if the data are meaningful in L AND if the TLD the registrant wishes to register the label understands S then proceed?

Questions for the working group:

1. Does the working group agree that registrar billing data and generally, any data a registrar or registry keeps that are unique from the registration data identified in a gTLD agreement or RAA are outside our scope?

2. Should universal ascii representation of domain properties be required.

Q1c-i) Related to c), and to avert a possible Babel effect, is it desirable to adopt a "must be present" representation with optional collection/display of registration data for the convenience of "local users"?

Response from Working group members:

1. **B**oth English and local language are necessary. (Jiankang)
2. Law enforcement, IP, and other businesses that automate/make use of Whois are likely to be interested in a lingua franca, right? So would these parties say that this alternative satisfies the needs of a local community at the expense (and considerable risk) to the global community? (Dave)
3. Other disagree. The reality of the world we live is in that there is no lingua franca and we should not fall into the trap of believing there is one. We already have the Babel scenario and cannot get away from that. In domain names by our use of ASCII we have not created a lingua france but simply excluded those people who do not use it. That is a state of affairs that cannot continue. To be accurate, there are two Babel scenarios that need to be separated out.
   a. - the first is where different self-contained groups have there own languages and scripts which are used within a *local context*. Some people speak/read/write more than one language/script for communications with others outside of that local context.
   b. - the second is where multiple languages scripts are used outside of the local context and communications breaks down.

   My suggestions above have all been about recognising the first scenario and trying to avoid the second. That leads onto the point about law enforcement, IPR etc. It is in my view, entirely unreasonable for us to insist that all registration data be duplicated in ASCII for English speaking law enforcement and law firms to be able to use it. That would exclude all those law enforcement officers and law firms that do not read ASCII or speak English. (Jay)

   The alternative, where we expect registration data in every language is clearly absurd, so we are left with the scenario where some registration data can only be understood in the local context, which mirrors the real world. (Jay)

   The argument that "all registration data is currently in ASCII and so readable by law enforcement, IPR concerns etc and we can't lose that" is completely bogus. We only have that situation because a lot of people have been excluded and we cannot continue that way. (Jay)

Questions for the working group:

We do not have consensus on this? What should we do?

Q1C-ii) Related to i), should we consider adopting a "format for civic address information that's reasonably functional around the globe"? (cf. Thomas Roeffler, in his reply to Steve Crocker)

Responses from Working group members:

1. This is unnecessary and just a lazy way of doing things. If we know the country the postal address data is for then there is a defined format for the local postal address (as well as generally standard local labels for the fields). The "correct" thing to do it is to ask for the data in the appropriate local format. If form designers asked for the country first then this would be trivial to implement once you know the fields/labels for each country. (Jay)
2. We may follow the EPP RFC about contact address. (Jiankang)

(Dave P responding)
So do we have three cases:

- adopt UPU standards
- adopt an unique civic address format?
- only accept local format

If only appropriate local format I imagine this would not satisfy law enforcement, security community, or IP concerns.

> (Jay responding to Dave P)
> Are there differences between UPU standards and local formats?
> You mean English-speaking "law enforcement, security community, or IP concerns" ?

Q1C-iii) Related to c), is it possible to adopt an ICANN policy that both meets the expectations and needs of internet users around the globe and has a high probability of adoption by ccTLD operators as well?

Responses from Working group members:

1. Yes provided we delineate the problems and solutions appropriately and consider the ccTLD perspective as well as the gTLD perspective. I suspect many gTLD people know far less about ccTLDs than vice versa and so some education might be appropriate. For example could we get a sample of WHOIS outputs for different countries? (Jay)
2. Jiankang: ICANN policy is useful.

Q1d) What do we require from internationalized registration data that registration data be collected and displayed (or returned via a whois/port 43 or successor protocol) *uniformely*, in manners that would allow applications to process the data efficiently? In particular, should applications that do not involve humans (automation) not be complicated by variations across display/collection practices/policies by registries and registrars?

Responses from Working group members:

1. I think this is way out of scope for this group. It would be useful to tackle standardised flagging of encoding in port 43 input/output but no more than that and even then I would want that to go through the IETF (Jay)
2. Of course, easy one is better.(responding to IN particular …) (Jiankang)
3. There is no technical restrictions for web representation and few problems with traditional command string (pipe, scripts, etc) for various *ix platforms. These problems must be resolved due to IDN era.” (Andrea)

Q1e) What do we require from internationalized registration data that some filtering or processing be considered to reduce the opportunity for deception/misuse when characters from multiple scripts are used in the composition of certain registration data? While perhaps not a perfect fit, are the prohibitions of mixed-scripts in the IDN guidelines that applied to labels of a domain name representative of the opportunities we might attempt to define?

>> Jay: I would go one step further and expect whole sets of data to be in one script or another, not mixed.

>> Jiankang: if all data are requested to use UTF8, that should be no problem.
>> Jiankang: prohibitions of mixed-scripts is necessary.

2) What should the registration data look like?

a) some form of tagging of the data to identify the piece (object), i.e., "this is a contact address"
Using XML, for example, such a tag and the data might look like
<contact-address>3 Myrtle Bank Lane</contact>

>> Jiankang: XML is good.

b) alternatively, tagging of blocks of data where a group of objects such as an administrative contact would be tagged
Using XML, for example, such a tag and the data might look like

```
<admin-contact>
   David Piscitello
   3 Myrtle Bank Lane
   Hilton Head SC 29926
   ...
 </contact>
```

>> Jay: Both of these are now heavily into protocol design, which is what we have the IETF for, not ICANN. I think it entirely inappropriate for us to be discussing XML (or other) field identifiers.

If we were to identify a list of requirements for a new/amended protocol then that could be pushed over to the IETF, but we should not design the solution here. This is not the right place to do it.

>> Jiankang: I prefer the following format

```
<organization> CNNIC</organization>
<address>
   <postal>
      <street>No.4 South 4th Street, Zhongguancun</street>
      <city>Beijing</city>
    </postal>
   <phone>+86 10 58813007 </phone>
   <email>yaojk@cnnic.cn</email>
</address>
```

c) some means to identify the language or script that the characters in the data "belong to"

>> Jay: Yes, I agree, see above.

>> Jiankang: that will be better since it will allow the user to easily identify the whois data language used.

d) should some elements of registration data always be represented in US-ASCII7 (e.g., sponsoring registrar)?

>> Jay: see above.

>> Jiankang: agree. such as telephone number.

but the email address is not in this category since there will have an internationalized email address soon.

Other issues:

This begs the question of whether, in the future, ICANN would accredit a company whose entity name makes use of extended character sets as a registrar, but I felt it useful to share the concern with the group.

> (Jay responding to Dave P)
> Very good question. It probably requires a dual code/name display for registrars, one in a standard character set and one in the local language set.

Another issue for internationalized data is security issues.If all whois data are uniformed, it not only convient to the normal internet users but also to the misusers who may easily get the information they want to do some offensive things.

Is security issue in the scope of our WG?