Regid Technical Considerations and Specification Proposal

Norbert Bollow <nb@bollow.ch>

2012-05-04

Abstract: In this document I comment on a few points about regids regarding which the WG21 internal draft of 2012-04-24 for 19770-3 has a number of shortcomings, and I propose text to address them in a way that is hopefully precise and clear.

1. Overview of issues addressed in this document

Issues of internationalization:

- Internationalized Domain Names (IDNs)
- Internationalized email addresses (RFC 5890)
- Appropriate character encoding from the usability perspective

Considerations related to the use of regids as part of filenames:

- Portability in avoiding characters that are not allowed in filenames
- Handling the case that such characters occur in an email address
- Explaining which this part of ISO/IEC 19770 does not require regid length limits

Further practical concerns

- Interoperability with implementations of the ISO/IEC 19770-2:2009 regid format
- Avoiding the need to know the precise month when the domain name was acquired
- Email address based regids also need date information.

Issues of precision

- Precise references to required characters as Unicode code points.
- Is an individual who does not own a domain name considered to be “an entity”?
- Requirements on domain names from which a regid can be contructed
- Precise normative reference to the syntax specification for email addresses in RFC

2. Discussion of some of the Issues

This section discusses some of the various issues listed above, giving the rationale for the proposals in section 3.

2.1 Issues of internationalization

It is not appropriate for International Standards to be based on the assumption that the ASCII character set (which is suitable for the English language) is the only one that matters. It is true that (unfortunately) that the basic protocols of the Internet and the corresponding naming and addressing systems were to a large part based on this assumption. Nowadays all people everywhere insist on being able to use their own language and its corresponding writing system for all aspects of interacting with the Internet. In many contexts it is very inconvenient for people whose working language uses a non-Latin character set to enter even very short text that uses the Latin alphabet, such as a domain name or an email address, due to the inconvenience of having to switch their software to using a keyboard layout for Latin characters and their reduced familiarity with this keyboard layout.

Consequently, the Internet community is working on the internationalization of domain names and email addresses, and it is no longer appropriate to assume that domain names and email addresses are always based on the ASCII character set.

2.2 The precise month of domain name acquisition may be difficult to know

The current draft states: “A date coded in YYYY-MM format representing the date when the naming entity initially acquired the domain. This should be the first month they owned it on the first day at 00:01 GMT. ”

However the person who implements the regid creation process may not have the information on

when the naming entity initially acquired the domain name readily available. Furthermore, it is conceivable that it may be highly undesirable to publish a date which “should be the first month they owned it” is situation where that date is subject to dispute. Imagine for example a situation when a domain name was first “possibly transferred” by a contract of which the validity or precise meaning is being disputed. If domain name ownership has been clarified later by a second, clear, contract, but the original disputed contract also covers other points that are still disputed, it would be very undesirable for the naming entity to publish a date which “should be the first month they owned it”.

I propose: “A month of the Gregorian calender ISO 8601 represented in YYYY-MM format, representing a month when the naming entity owned the domain on the first day of the month at 00:01 GMT. The naming entity should consistently use the same month for all regids that are based on the same domain name.”

2.3 Interoperability with implementations of the 19770-2:2009 regid format

Probably the biggest change that I propose is a pretty radical change to email based regids to make them fit into the structure of 19770-2 regids, that that they can be handled by implementations of the ISO/IEC 19770-2:2009 regid format without causing problems.

2.4 Issues of precision

2.4.1 Precise references to required characters as Unicode code points

The reference to a specific character as “a dot” is not sufficiently precise, as in the Unicode standard there is a variety of characters that can be described as “a dot”. This issue can be addressed by including a precise Unicode code point reference. For the sake of consistency, and because precision never hurts, I propose to also add similar Unicode code point references for the other specific characters required by the specification.

2.4.2 Is an individual without a domain name considered to be “an entity”?

The current draft is inconsistent regarding this point: The second paragraph of clause 7.4 speaks explicitly of the need “to support entities which do not own a domain name ”, giving “an individual who purchases a product and receives an entitlement ” as an example. Also, definition 4.1.56 uses the term “entity” for whoever creates a tag. However the text of the lower half of page 38 is based on the implied assumption that an individual who does not own a domain name is not considered “an entity”.

I propose to leave definition 4.1.56 which comes from the already published standard ISO/IEC 19770-2 unchanged, and write the text regarding regid tags so that it is not inconsistent with the understanding that anyone who creates a tag (individuals as well as organizations of any kind) can be included in the meaning of the term “entity”.

2.4.3 Requirements on domain names from which a regid can be constructed

ISO/IEC 19770-2 and the current draft are unclear on what exactly is meant with “a domain name”. I understand that the intended meaning is the kind of registration that you can get from a top-level domain name registry (like “example.com”) or from a country domain name registry (like “example.ch” or “example.co.ke”) but not the kind of third, fourth, etc level domain name that anyone can create without formal registration and without any formal accountability regarding record-keeping of who owns the domain name (like “co.example.com”) and also not a domain name registration under one of the “alternative roots”, and I propose corresponding text. This has a security aspect, since if domain name registrations from unaccountable informal registries are accepted for regid use, such registries may fail to have policies that disallow the registration to different parties of domain names which look alike even if they are different technically.

3. Proposed text for the standard

The following text is intended to replace the 24 lines in the draft of 2012-04-24 starting with the nineth line of clause 7.4 (“The regid for an entity...”) and ending with the third line on the following page (“RegIds shall not contain characters that are inconsistent with filename use such as '/', '\', '[', etc. ”).

BEGINNING OF PROPOSED TEXT FOR THE STANDARD

7.4.1 What domain names are suitable for use in a unique registration identifier (regid)

All formally registered Internet domain names can be used as the basis for a regid. This includes in particular all second level domain names such as domain names of the form “example.org”, “example.info”, “example.com” etc., and it includes all domain names that can be registered at country-code top level domain (ccTLD) registries which may include third-level domain names of the form “example.or.ke” or “example.ac.uk”.

The formal definition of what is a *formally registered Internet domain* name is based on IANA's Root Zone Database, see http://www.iana.org/domains/root/db/ : A *formally registered Internet domain name* is one which has been registered in the the registry that is operated or commissioned by the Sponsoring Organization listed in the Root Zone Database for the respective top-level domain.

NOTE in most cases, individuals or organizations wishing to buy an Internet domain name do not interact directly with the organization that operates the registry. Rather, for most top-level domains (TLDs) such as “.org”, “.info”, “.com”, etc., registrations of second-level domain names are executed by specialized ocmpanies, so called registrars, on behalf of the customer.

NOTE domain names by themselves do not constitute a unique identifier since domains because they can expire or be sold to another entity. However in the regid they are combined with a date so that uniqueness of the regid is assured.

NOTE it is not required for the domain name to be in active use on the Internet, or to resolve to an address.

NOTE although anyone who has registered a domain name can create subdomains such as e.g. “subunit.example.com” and use them on the Internet, such subdomain domain names are not formally registered domain names, and they are not used in regids. In the context of regids, a different mechanism is used when it is desired for subunits of a naming entities to be able to issue tags autonomously, see 7.4.2.1(f) below.

7.4.2 Structure of regids

7.4.2.1 Regids for naming entities that own a formally registered domain name

Any naming entity which owns a formally registered domain name (see 7.4.1 above) shall create a regid string by concatenating the following:

(a) The string “regid” , consisting of the five characters with Unicode code points U+0072, U+0065, U+0067, U+0069, U+0064.

(b) A dot “.”, i.e. the character with Unicode code point U+002E.

(c) A month of the Gregorian calender (as defined in ISO 8601) represented in YYYY-MM format, representing a month when the naming entity owned the domain on the first day of the month at 00:01 GMT. The naming entity should consistently use the same month for all regids that are based on the same domain name.

(d) A dot “.”, i.e. the character with Unicode code point U+002E.

(e) The reversed domain name of the naming entity . For example, if the domain name of the

naming entity is “example.com”, then the reversed domain name is “com.example”.

(f) Optionally a further string introduced by a comma “,”, the character with Unicode code point U+002C.

NOTE if the naming entity owns multiple domain names, regids based on multiple different domain names may be used. For example, a company which has customers in Western Europe and customers in China might use a regid constructed from the company's domain name in Chinese characters when dealing with customers in China, while a regid constructed from the company's domain name in Latin is used when dealing with customers in Western Europe.

7.4.2.2 Regids for naming entities without formally registered domain name

In the case of entities which do not own a formally registered domain name (see 7.4.1 above), or when a regid is generated for an individual by means of an automated process during which do information is available regarding any domain names that the individual may own, a regid with the same basic structure shall be created from an email address for the individual or a representative of the naming entity, by concatenating the following:

(a) The string “regid” , consisting of the five characters with Unicode code points U+0072, U+0065, U+0067, U+0069, U+0064.

(b) A dot “.”, i.e. the character with Unicode code point U+002E.

(c) A month of the Gregorian calender ISO 8601 represented in YYYY-MM format, during which the individual or representative of the entity used the email address.

(d) A dot “.”, i.e. the character with Unicode code point U+002E.

(e) The string “invalid.unavailable”.

NOTE “.invalid” is the IANA standard for non-existing domains.

(f) A comma “,”, the character with Unicode code point U+002C.

(g) The email address for the individual or the representative of the naming entity.

(h) A comma “,”, the character with Unicode code point U+002C.

(i) The name of the individual or other naming entity, after applying the algorithm specified in 7.4.4.2 below to remove or replace any problematic characters. For example, if the naming entity's formal name is “Example, Inc.”, it is the string “Example_Inc.” that would be included in the regid.

(j) Optionally a further string introduced by a comma “,”, the character with Unicode code point U+002C, and conforming to the requirements of 7.4.4.1 below.

NOTE regids which are constructed from email addresses conform to the structure of regids that are constructed from domain names.

NOTE regids can be recognized as having been constructed from email addresses by the presence of the string “invalid.unavailable” where otherwise there would be a formally registered domain name.

7.4.3 Considerations regarding length limits

This part of ISO/IEC 19770 only specifies the use of regids within the XML data content of tags. Therefore it is not necessary here to impose length limits on regids.

NOTE in ISO/IEC 19770-2, regids are used for filenames of files that hold software identification tags. That is however in a context where the naming entity creates not only its own regid but also has control over the other components of the filename, and furthermore the naming entity has information about the software platform on which the software is being installed. Therefore, in that context, the naming entity has the necessary information for being able to avoid filenames that

violate the platform's length limits. In particular, the need to include an arbitrary entity name and email address in a regid does not occur in that context.

7.4.4 Avoiding problematic characters

Even though this part of ISO/IEC 19770 does not foresee the direct use of regids as parts of filenames, for consistency with ISO/IEC 19770-2, characters that are problematic in filenames shall be avoided.

7.4.4.1 Characters to avoid

Regids shall not contain any of the following characters:

(a) U+0000, NULL.

NOTE on many platforms, this character is disallowed in filenames.

(b) Any of the control characters in the ranges from U+0001 to U+001F and U+0080 to U+009F.

NOTE filenames including these characters are problematic from a usability perspective on many computer system platforms.

(c) U+0020 SPACE.

NOTE filenames including SPACE characters are problematic from a usability perspective on some computer system platforms.

NOTE the algorithm of 7.4.4.2 below replaces SPACE characters with “_” characters.

(d) U+0022 QUOTATION MARK.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(e) U+002A ASTERISK, “*”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(f) U+002F SOLIDUS (slash), “/”.

NOTE on many platforms, this character is disallowed in filenames.

(g) U+003A COLON, “:”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(h) U+003C LESS-THAN SIGN, “<”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(i) U+003E GREATER-THAN SIGN, “>”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(j) U+003F QUESTION SIGN, “?”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(k) U+005C REVERSE SOLIDUS (backslash), “\”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

(l) U+007C VERTICAL LINE (vertical bar), “|”.

NOTE on Microsoft platforms, this character is disallowed in filenames.

7.4.4.2 Sanitizing the “further string” and “name” fields

Software tools that are used to generate regids shall process input data for the “further string” fields 7.4.2.1(f) and 7.4.2.2(j) as well as for the “name” field 7.4.2.2(i) as follows:

- All SPACE characters (U+0020, “ ”) are replaced with LOW LINE characters (U+005F, “_”).
- All other problematic characters listed in 7.4.3.1 are simply removed.

NOTE the character U+0026 AMPERSAND, “&” is *not* considered a “problematic character”. It shall not be removed. In XML contexts, this character needs to be encoded as “&”.

7.4.4.3 Sanitizing the “email address” field

It is unusual but allowed by the relevant Internet standards IETF RFC 5322, IETF RFC 5890 for email addresses to contain all of the problematic characters of 7.4.4.1 above. (NOTE although the inclusion of NULL characters and control characters in email addresses is declared “obsolete”, it is still allowed.)

Software tools that are used to generate regids from email addresses shall encode these problematic characters by means of the following algorithm:

- All instances of the character U+0025 PERCENT SIGN, “%”, are replaced by “%25”.
- All instances of the problematic characters listed in 7.4.4.1 above are likewise encoded by the percent encoding mechanism specified in section 2.2 of IETF RFC 1738.

No instances of other characters shall be encoded by means of this percent encoding mechanism.

7.4.4.4 Internationalized domain names (IDNs) and internationalized email addresses

This part of ISO/IEC 19770 supports internationalized domain names (IDNs) and internationalized email addresses.

Internationalized domain names (IDNs) are specified in IETF RFC 5890.

The domain name of the naming entity that is used in 7.4.2.1(e) may be an IDN. Software tools that are used to generate regids shall support IDNs. A part of the process of creating the regid, IDNs shall, if necessary, be converted into the representation that does not include "Punycode strings". This can be stated precisely as follows: In an internationalized domain name, each of the labels that together constitute the domain name is either an NR-LDH label or an A-label or a U-label, as these terms are defined in sections 2.3.1 and 2.3.2.1 of IETF RFC 5890. As part of the process for generating a regid from an IDN, all A-labels shall be converted to the equivalent U-labels.

Internationalized email addresses are specified in section 3.3 of IETF RFC 6532. Software tools that are used to generate regids from email addresses shall support internationalized email addresses. As part of the process of creating the regid, the <Domain> part of the email address in 7.4.2.2(g) shall if necessary be converted into the representation that does not include any A-labels.

7.4.5 Examples of regids

(a) A typical regid, derived from the domain name “example.com”

regid.2012-07.com.example

(b) A regid derived from the domain name “example.org” and the further string “General Agency for EMEA”

regid.2012-12.org.example,General_Agency_for_EMEA

(c) A regid, derived from the IDN παράδειγμα.δοκιμή

regid.2012-04.δοκιμή.παράδειγμα

(d) A regid for an individual named “John Doe” with email address <john.doe@example.com>.

regid.2012-05.invalid.unavailable,john.doe@example.com ,John_Doe

(e) A regid for an entity named “Exam & ple” which has no domain name but which has the email

address <”Exam & ple”@example.com>. (NOTE this email address is highly unusual in that it contains space characters, and therefore it must use quoting.)

regid.2012-06.invalid.unavailable,%22Exam%20&%20ple%22@example.com,Exam_&_ple

END OF PROPOSED TEXT FOR THE STANDARD