Initial Notes

Jens:

= Recommended Practice for the IGSN Syntax =

The IGSN is organised by namespaces to ensure globally unique allocation of IGSNs to samples. The standard format of the IGSN is a nine character string consisting of a three character namespace identifier and a six character sample number.

Example: PRR001407 (a sample from the US Polar Rock Repository)

The Handle value is concatenated by the following rule:

prefix/<igsn>

where prefix is 10273 and <igsn> is the value (IGSN) assigned by an Allocating Agent.

Example: 10273/PRR001407

This identifier can be resolved to the URL of the metadata page of the sample through any handle resolver, e.g. http://hdl.handle.net

Example: http://hdl.handle.net/10273/PRR001407

IGSN user communities may adopt their own IGSN formats, if necessary.

Use cases:

(1) shorter collection prefix to allow for a larger collection namespace or sub-namespaces.

Example: ICABR1133 (hypothetical example), ICDP/IODP mission specific platform (proposed namespace IC), sample AB-R-1133

(2) more than 9-character IGSNs to allow the use of well established legacy formats, e.g. in core repositories.

Example GEOB12802-13 (Gravity Core from the Bremen Core Repository, data published as:

Lin, Y-S et al. (2012): Primary geochemical analysis of a transect during Meteor cruise M76/1. <http://dx.doi.org/10.1594/PANGAEA.778018>,

Supplement to: Lin, Yu-Shih; Heuer, Verena B; Goldhammer, Tobias; Kellermann, Matthias Y; Zabel, Matthias; Hinrichs, Kai-Uwe (2012): Towards constraining H2 concentration in subseafloor sediment: a proposal for combined analysis by two distinct approaches. Geochimica et Cosmochimica Acta, 77(15), 186-201, <http://dx.doi.org/10.1016/j.gca.2011.11.008>

The above example illustrates the value of keeping the samples names used in the Bremen Core Repository and in the PANGAEA database.

= Recommended Practice for Using IGSN in the Literature =

To aid the identification of IGSNs for text-mining in the literature it is recommended to precede the IGSN with igsn:

Example: igsn:PRR001407

Example: igsn:GeoB12802-13 (note: IGSN are not case sensitive)

If the IGSNs are listed in a table column a column header IGSN should be sufficient for identification in text mining.

= Namespace Allocation =

In the current technical system the authoritative list is the list of namespaces and allocators in the IGSN registry and metadata store. At the moment I do not see an explosion of name space applications to IGSN e.V., maintaining a list of authorised namespaces in the wiki should be possible. I guess, Damian can also extract views from the underlying IGSN metadata store database.

The question now is, how far shall we formalise the procedure of namespace allocation? Should there be a period in which IGSN members can comment on applications for namespaces? How should an application for a namespace be entered? I would prefer if applications for namespaces were entered through the trac ticket system, as this would provide us with the best transparency and formalised workflow without having to set up a new system.

= AGU 2012 =

- Find a date for the IGSN General Assembly at AGU (suggested breakfast meeting)
- Other action items in preparation for AGU 2012

= Accepting for-profit members =

I have not heard back yet from Jan Brase on how they accepted Microsoft

Research as an Associated Member in DataCite. Frauke Ziedorn of TIB Hannover said that according to her information it was a matter of the members of DataCite simply voting on it. As long as accepting a for-profit organisation as a member is not in conflict with §2.2 <http://dokuwiki.gfz-potsdam.de/datawiki/doku.php?id=igsn:statutes#purpose_objectives_charitable_status> of the statutes, simply voting on it should be fine. Changing the statues produces quite a bit of work (with legal departments) and paperwork.

Comment and Reply

From Jens

Kerstin,

In our discussion we should clearly distinguish between IGSN syntax rules and namespace governance, and SESAR syntax and namespace governance.

In general, I agree with the rules set for SESAR as they are well thought out and make a lot of sense in that context. In a broader context of a more general application of IGSN, legacy data, and text mining I would allow exceptions to some degree, and as long as they do not break the system.

Am Sonntag, den 16.09.2012, 15:40 -0400 schrieb Kerstin Lehnert:
> Couple of comments on the syntax:
>
>
> I would strongly recommend that we not allow dashes, slashes, etc..
> People will start to mess this up, they will put slashes instead of
> dashes, leave them out all along, etc. There are a lot of issues with
> that.

I would still allow dashes as an exception, because core repositories use them and you will find these names used in the literature.

Example GEOB12802-13 (Gravity Core from the Bremen Core Repository, data published as:

Lin, Y-S et al. (2012): Primary geochemical analysis of a transect during Meteor cruise M76/1. <http://dx.doi.org/10.1594/PANGAEA.778018>, Supplement to: Lin, Yu-Shih; Heuer, Verena B; Goldhammer, Tobias; Kellermann, Matthias Y; Zabel, Matthias; Hinrichs, Kai-Uwe (2012): Towards constraining H2 concentration in subseafloor sediment: a proposal for combined analysis by two distinct approaches. Geochimica et

Cosmochimica Acta, 77(15), 186-201,
<http://dx.doi.org/10.1016/j.gca.2011.11.008>

For any new IGSN I agree that these should not contain dashes, etc.

Example: ICABR1133 (hypothetical example), ICDP/IODP mission specific platform (proposed namespace IC), sample AB-R-1133

However, we did agree in the past to follow the CrossRef recommended PID syntax:
<http://www.crossref.org/help/Content/02_Getting_started/Establishing_a_DOI_suffix_pattern.htm>.

> I think that we had already agreed that name spaces will need to be
> managed in a separate system. This system should allow clients to
> request name spaces, and the system will check for validity and
> availability of the requested name space and register the name space
> if it is valid and available. I think that 2-digit name spaces should
> only be handed out rarely to big facilities with a special approval
> process as each 2-digit name space uses up 36 3-digit name spaces (or
> am I missing something?) Should this system run at the Managing Agent
> or at the registry?

Sorry, I was a bit imprecise in what I had written. I meant to say that we should find a way to start issuing namespaces right away. Since there is no system in place to administer namespaces, we need to find a way to start.

Next, we need to detail the procedure for requesting and registering namespaces. Having defined the process, we can specify the requirements for a namespace registry and find/build a suitable technical system.

The IGSN metadata store already has a namespace administration that takes namespaces and ownership into account. Maybe there is a way to modify this feature.

Yes, we all agree that 2-character namespaces should be rare exceptions with a special approval process. Yet, we need to define this process.
>
> Maintaing a list of namespaces in the wiki does not work for SESAR
> because users can currently register new name spaces without any
> interference of SESAR personnel. Our experience has been that it is
> not well appreciated by users if they have to wait for approval of a
> requested name space. They need to be able to start registration of
> samples right away. That's why we need a system that allows users to
> register a desired name space automatically. Only 2-digit name spaces
> should require approval by the IGSN eV.
>

In general I agree. One thing I understood quite differently from our conversations. I thought that, in analogy to European telephone area codes, Assigning Agents would register a number of top-level namespaces and the assign these to certain applications or clients, or create sub-namespaces by appending further characters to the top-level namespace. We discussed this procedure to accommodate small collections without using up to much of the IGSN namespace and keep the 9-character syntax.

> SESAR will have be modified to give up control of the name spaces.
> Until a new system is in place, we will have to continue operating
> SESAR.
>
With the approach outlined above, SESAR still has a role in assigning and administering namespaces - its own namespaces. SESAR would register its top-level namespaces with IGSN.

Example: a hypothetical collection of samples from an ongoing project on carbonatites. The top-level namespace registered by SESAR with IGSN and administered by SESAR may be CAR.

Sub-namespaces are negociated between SESAR and the SESAR community and administered by SESAR (e.g. CARK - Kaiserstuhl, CARO - Ol Doinyo Lengai, CARP - Phalaborwa). In this case all sub-namespaces below CAR are reserved for SESAR, no conflicting registration from outside SESAR is possible. This would eliminate the need for an instantaneous registration at IGSN level which would make life much easier in the starting phase. SESAR could continue the way users obtain their namespace, it would only be limited to namespaces owned by SESAR.

> Regarding IGSNs in table columns: If we want the tables to be included
> in text mining, we could not delete the prefix.
>
Tables in hypertext used by the publishers always have tagged column headers. If one of the columns is tagged "IGSN" it should be clearly identifiable as a column on IGSNs. This would spare us of using igsn: prefixes in tables and save a lot of space.

More

From Jens:

I found some literature on Friday, which I forgot to mention in my previous e-mail.

Bechtold, S. (2003), , Loyola of Los Angeles Law Review, 36(3), 1239–

1320, doi:10.2139/ssrn.413681. [online] Available from: http://digitalcommons.lmu.edu/llr/vol36/iss3/6 (Accessed 14 September 2012)

After ten years the examples may seem a bit dated (MS Passport, Napster, ...) but the principles of namespace governance and control are well explained.

From Simon:
Re: PID Structure
Wed, Jul 11, 2012 at 11:50 PM

OK - Now I think I understand the issue better, and the precedents we have to follow.

Using ABNF notation [1] the syntax for an IGSN used by SESAR is currently

```
Igsn            = AllocatingAgent Code
AllocatingAgent = 3UPPER          ; a 3 character code denoting the allocating agent
Code            = 6DIGIT          ; a 6 digit code
UPPER           = %x41-5A           ; A-Z
DIGIT           = %x30-39           ; 0-9
```

The Allocating Agent ensures that the 6-digit code is unique within their namespace.
(does this have to be digits, or are other characters allowed? 1M samples per namespace does not seem quite enough to me!)

The request from Anthony, Anders, etc is to relax the rule for 'Code'.
I agree with this - it makes sense to delegate the design of the code to the allocating agent, in the way that the design of DOI suffix is delegated to the DOI provider.
Since we intend the IGSN to be combined into a URI, n order to retain maximum compatibility with URI production rules [2] I suggest we limit the gamut of characters that can be used in the code to the so-called 'unreserved' + 'reserved' set, but not allow any other or percent-encoded characters which may exist on the keyboard or other character sets (e.g. no accented characters or non-latin alphabets, no space, CR, LF characters, leading to .

```
Igsn            = AllocatingAgent Code
AllocatingAgent = 3UPPER          ; a 3 character code denoting the allocating agent
UPPER           = %x41-5A           ; A-Z
Code            = 6*CHAR          ; minimum of 6 characters, maximum undefined
CHAR            = unreserved / reserved
unreserved      = ALPHA / DIGIT / "-" / "." / "_" / "~"
reserved        = gen-delims / sub-delims
gen-delims      = ":" / "/" / "?" / "#" / "[" / "]" / "@"
sub-delims      = "!" / "$" / "&" / "'" / "(" / ")" / "*" / "+" / "," / ";" / "="
```

[1] http://tools.ietf.org/html/rfc5234
[2] http://tools.ietf.org/html/rfc3986

Simon Cox
Research Scientist
CSIRO Earth Science & Resource Engineering

Phone: +61 8 6436 8639 | Fax: +61 8 6436 8555 | Mobile: 0403 302 672
simon.cox@csiro.au | www.csiro.au
Address: ARRC, PO Box 1130, Bentley, WA 6102, Australia

From: IGSN Members [igsn-members@gfz-potsdam.de] On Behalf Of Anthony Koppers [akoppers@coas.oregonstate.edu]
Sent: Friday, 22 June 2012 12:12 AM
To: IGSN Members
Subject: [IGSN Members] Re: European Persistent Identifier Consortium (EPIC) - PID Structure

I agree with Simon, each organization will do the hierarchy differently for very good reasons. That is why I suggested to have no structure to the optional part, except for barring some strange characters and maybe limiting the string length. The examples I gave were just examples. I view the IGSN service as a registry of unique sample identifiers, nothing more. I just want to add some flexibility so that a large variety of organizations can buy into the concept. My main inspiration is the DOI, where the second part of the string is up to publishers to fill in and they all have different solutions. Also, in my mind it should be up to the repositories to decide what will be their parent IGSN number, which might be different for each organization. In short, I would strive to keep it flexible.

Cheers. Anthony.

==========