"duri" and "tdb": URN Namespaces based on dated URIs

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

This document defines two persistent namespaces of URNs based on prepending a date to an (encoded) URI. The results are namespaces in which names are readily assigned but which offer the persistence of reference that is required by URNs. The first namespace (duri) is used to refer to URI-identified resources themselves, while the second namespace (tdb) is used to refer to abstractions that are not themselves networked resources but are "described by" them. This idea and things like it have been discussed for several years, but recent discussion about use of URIs and URNs for identifiers in XML-based constructs has inspired writing this up more completely.

The purpose of this document is to help focus the discussion of the role of URIs and URNs as names within non-Web applications. This document is not a product of any working group, but may be discussed on the mailing list <uri@w3.org>. (Discussion of related topics has occurred on urn-ietf@lists.netsol.com and www-rdf-interest@w3.org and w3c-uri-ig@w3.org).

Table of Contents

1. Overview
2. Encoding URIs
   2.1 Characters that must be encoded
   2.2 No need to encode "/"
3. Dates
4. Additional considerations
   4.1 URI schemes
   4.2 Date ranges
   4.3 Free assignment
   4.4 Resolution
4.5 Why Names with Semantics?

In some cases, the guarantee of persistence comes through (a promise of) good management practice; a promise that "Cool URIs don't change" [COOL]. However, a promise of good management practice is different from a design that insures reliability.

The primary principle of "Uniform" URIs is that they are intended to mean the same thing, no matter in what context they appear; thus URIs are a Uniform (in meaning) way to Identify a Resource. However, even when URIs have Uniform meaning from the point of view of the source of the reference, they don’t implicitly guarantee stability over time. Despite best efforts and intentions, identifying information can change in unpredictable ways, be it domain names, name assigning organizational structure or identity.

It is traditional in convention references and citations in printed works to include the date of publication; this practice serves the important purpose that the context of the naming can be determined.

The "duri" URN namespace takes the form:

\[
\text{urn:duri:<date>::<encoded-URI>}
\]

where <date> is a digit string corresponding to a date (Section 3), and an <encoded-URI> is an absolute URI-reference [RFC 2396] in which any character excuded from URN syntax has been escaped (Section 2).

The meaning of a duri is "the resource (or fragment) that was identified by the <encoded-URI> (after hex decoding) at the very first instant of the date given".

For example, urn:duri:2001:http://www.ietf.org is a persistent identifier to 'http://www.ietf.org' as of the very first moment of the year 2001. A duri may not be a resource locator in a practical sense, because the time of location has passed. However, is an acceptable resource identifier, and fulfills all of the requirements for URNs.[RFC 1737].

The second URN namespace defined is a parallel space which is useful for describing entities, concepts, abstractions, and other items which are not themselves network accessible resources, but have been described by network accessible resources.
uses for URIs are for objects or concepts that don’t actually correspond to networked resources, but for which the URI space is used as the identifier. To fill some of the need for such identifiers, a second namespace is defined which designates the "thing described by" the resource at the given URI at the given date and time. This URN namespace is described by 'tdb', e.g.,

urn:tdb:<date>:<encoded-URI>

with the same syntactic rules as duris.

So "urn:duri:2001:http://www.ietf.org" can be used to designate the Internet Engineering Task Force organization, at least as it was described by or referenced by its home page at the first instant of 2001.

There are various other proposals for URN name spaces for abstract entities that don’t make reference to a concrete networked resource for the purpose of identification; in much the same way that ASN.1 object identifiers don’t contain any particular semantics of the object identified. The "tdb" URN namespace satisfies a different set of needs, since the designation of what is actually identified by the tdb is clear and determinable without reference to the context of its use.

2. Encoding URIs

Both "duri" and "tdb" URN namespaces require that some characters in the URI references be encoded.

2.1 Characters that must be encoded

The characters that must be encoded are:

* All characters marked <excluded> in RFC 2141, section 2.4
  These are excluded because they are not allowed in URNs.
  \"&<\>

* The character "#"
  Note that the <encoded-URI> of a "duri" or "tdb" can include a fragment identifier, but the "#" character used to delimit it must be encoded.

* The character "%"
  The encoded-URI can itself contain encoded characters, which are encoded with the same method. To insure that decoding happens at the right level of processing, the "%" itself must be encoded.
  Unfortunately, this results in a confusing double encoding, but this is difficult to avoid.

2.2 No need to encode "/*"

The URN recommendation discourages the use of "/*" in URNs because, in general, there is no good interpretation of hierarchy and relative URIs for assigned names. However, for the particular case of duris (at least), there seems to be no good reason to avoid the "/*" because it corresponds fairly naturally (in many cases) to the hierarchy of the original space.

3. Dates
A <date> is a simple expression of date, optional time, with arbitrary precision. The goal is to allow relatively short expressions of dates with no ambiguity, and with arbitrary precision. (The idea for this syntax came from [RFC 2550].)

\[
date = \text{year } [ \text{ month } [ \text{ day } [ \text{ hour } [ \text{ minute } [ \text{ second } [ \text{ fraction } ] ] ] ] ] ]
\]

year = 4digit
month = 2digit
day = 2digit
hour = 2digit
minute = 2digit
second = 2digit
fraction = *digit

The representation of a date or time refers to the very first instant of the given date, so that, for example, 1999 and 199901010000 are equivalent. If necessary, dates can include times and even fractional times, so that a generator of duris can be arbitrarily precise.

Dates are interpreted relative to International Atomic Time [TAI], so that there is no ambiguity about time zone.

4. Additional Considerations

4.1 URI schemes

Many URI schemes are appropriate for use inside duris and tdb URNs.

Of course, a common usage would be use a "http" URI to refer to a web page or the subject of a web site at a given time. This can be a way of referring to a web site at some date in the past, or an organization that has changed or merged.

Local systems that have unique host names can use "file" URIs in their duris, for example,

\[
\text{urn:tdb:20010814142327:file://this.example.com/c\|/temp/test.txt}
\]

can uniquely and unambiguously refer to a concept whose description is contained in a system’s local disk. While file URIs are difficult to use for global resolution because of ambiguities of file system and access methods, in this case, because the instant is fixed, the naming mechanism of the host can prevail.

Even the "data" URI scheme might be used with "tdb" to designate concepts that can be described briefly inline. For example,

\[
\text{urn:tdb:2001:data:,The\%2520US\%2520president}
\]

names the concept described by the (text/plain) string "The US president" at the very first instant of 2001. (Note the awkward double quoting of space as "%20" and then the "\%" as "%25".)

Even urns might appear within a duri in unusual circumstances. For example, there are circumstances where the assignment of names a URN namespace are not in practice be permanent, or that one might want to refer to the assignment as of a given date. In this case, it is possible to use a "urn" within a "duri", e.g.,

\[
\text{urn:duri:2000:urn:ietf:std:50}
\]
might be used to refer to "the document that was STD 50 that was in
effect as of the first instant of 2000". [RFC 2648]

4.2 Date ranges

Dates in the future SHOULD NOT be used, because the meaning of the
duri or tdb cannot readily be determined in advance reliably. Dates
far in the past or merely prior to the actual assignment of the
resource to the URI SHOULD NOT be used, because the meaning of the
reference is left in question. For example, using http URIs before a
web service was available at the given URI doesn’t make much sense.

However, although these practices are not recommended, there is no
assurance that they have followed; by itself, a duri/tdb does not
constitute an assertion that the encoded-URI was available or assigned
at the date specified.

Note that the use of the "very first instant" means that a duri/tdb
using only a year must give a year greater than the first year in
which the corresponding URI was published; if a web page is published
in the middle of 2001, then "duri:2001:..." would be inappropriate.

4.3 Free assignment

Because of the many possible schemes that can be used in the
<encoded-URI> portion, there should be no difficulty in almost any
computational process being able to assign duris or tdbs at will. Of
course, it is necessary for there to be some resource which is
available at some point in time, and to have a clock which is
accurate to the granularity of the frequency of assignment.

4.4 Resolution

There are no accurate resolution servers for duri or tdb URNs. A duri
might be "resolvable" in the sense that a resource that was accessed
at a point in time might have the result of that access cached or
archived in an Internet archive service. A "tdb" is only resolvable in
the sense that if the corresponding duri can be resolved, the result
can be accessed and interpreted.

Clients without access to an Internet archive service might take the
decoded <encoded-URI> of a duri and attempt resolution of *that*
identifier. This will give an approximation whose reliability depends
on the amount of time elapsed since the date indicated.

4.5 Why Names with Semantics?

There are a number of proposals for URN schemes that create otherwise
unbound "names", where the URN scheme only provides for uniqueness.
Neither "duri" nor "tdb" intrinsically have the property that the
names assigned are without any resolution semantics. This is
intentional; it’s difficult to create names that carry no semantics
whateooever about the authority that assigned the name and the
intention of the authority for what the name should designate.

4.5 Avoiding MetaData

One might consider the date in a duri/tdb to be just one piece of
additional metadata about the encoded-URI, and consider adding other
pieces of metadata as annotation.
However, the use of the date in a duri/tdb is intended primarily as a mechanism of accomplishing uniqueness over time. No other bit of metadata or description readily fills that purpose. Further, the date is not descriptive (an assertion about the encoded-URI) but merely refining.

4.6 Avoiding duri and tdb

Many applications of URIs already provide a context of date. For example, one could imagine a hypertext system where the URIs contained within a document were intended to refer to the resources as of the date of the enclosing document. This would be a reasonable interpretation of URIs within an Internet archive system, for example.

And some applications of URIs arguably already contain the level of interpretive indirection that is explicit with "tdb". For example, one might consider the use of URIs as namespace names within XML [XMLNAME] as a reference to the "thing described by" the URI used.

The Resource Description Framework [RDF] is an XML-based framework for describing assertions. RDF uses URIs to identify the objects being described and XML-based tags to describe the relationships between them. The relations in RDF, however, may already provide for the "thing described by" indirection. For example, the example in Section 3.2.1 of RDF claims the model for the sentence

"The students in course 6.001 are Amy, Tim and Mary"

would be written in RDF/XML as

```xml
<rdf:RDF>
  <rdf:Description about="http://mycollege.edu/courses/6.001">
    <s:students>
      <rdf:Bag>
        <rdf:li resource="http://mycollege.edu/students/Amy"/>
        <rdf:li resource="http://mycollege.edu/students/Tim"/>
        <rdf:li resource="http://mycollege.edu/students/Mary"/>
      </rdf:Bag>
    </s:students>
  </rdf:Description>
</rdf:RDF>
```

but the resources listed are web pages (served by HTTP) and the class and students are the "things described by" those web pages.

Other resource description frameworks may require using "tdb" to distinguish between assertions about classes or students and the web pages that describe them.

5. URN Specification Templates

5.1 "duri" Specification Template

Namespace ID: "duri" requested.

Registration Information:
  Registration Version: 1
  Registration Date: 2001-08-19

Declared registrant of the namespace:
Larry Masinter (see Section 10 of this document.)

Declaration of syntactic structure:
Briefly, the syntax is
urnduri:<date>:<encoded-URI>
The syntax is described in Sections 1-3 of this document.

Relevant ancillary documentation:
(See Section 10, References, of this document)

Identifier uniqueness considerations:
Uniqueness is guaranteed by the structure of adding
a designation of a specific instant to a URI. However,
URIs with ambiguous interpretation at any given
instant (e.g., "file" URIs without a given host name)
will not be unique.

Identifier persistence considerations:
The designation of a dated URI is completely persistent
for all time.

Process of identifier assignment:
Any date can be used with any URI independently
by anyone.

Process of identifier resolution:
Identifiers can only be resolved approximately. See
Section 4.3.

Conformance with URN Syntax:
Note that the use of "/" for hierarchy, while discouraged
in the URN specification, is allowed in duris.

Rules for Lexical Equivalent:
For dates, YYYY is equivalent to YYYY01, YYYYMM is equivalent to
YYYYMM01, while YYYYMMDD is equivalent to YYYYMMDD0... followed
by any number of 0’s.

In considering equivalence of the encoded URI, if two duris with
equivalent dates contain lexically equivalent URIs, the duris
are equivalent.

Validation mechanism:
Dates should be reasonable and meet the syntactic requirements.
The URI encoded within should meet the syntactic requirements of
the URI scheme used.

Scope:
Global.

5.2 "tdb" Specification Template

Namespace ID:
"tdb" requested.

Registration Information:
Registration Version: 1
Registration Date: 2001-08-19

Declared registrant of the namespace:
Larry Masinter (see Section 10 of this document.)
Declaration of syntactic structure:
Briefly, the syntax is
urn:tdb:<date>:<encoded-URI>
The syntax is described in Sections 1-3 of this document.

Relevant ancillary documentation:
(See Section 10, References, of this document)

Identifier uniqueness considerations:
Uniqueness is guaranteed by the structure of adding
a designation of a specific instant to a URI. However,
URIs with ambiguous interpretation at any given
instant (e.g., "file" URIs without a given host name)
will not be unique.

Identifier persistence considerations:
The designation of a dated URI is completely persistent
for all time, although the intent of a resource that
is no longer available will be hard to discern.

Process of identifier assignment:
Any date can be used with any URI independently
by anyone.

Process of identifier resolution:
Resolution of "tdb" identifiers requires interpreting
the resource identified by the corresponding "duri".
See Section 4.3 of this document.

Rules for Lexical Equivalent:
As with "duri", see section 5.1.

Conformance with URN Syntax:
As with "duri", see section 5.1.

Validation mechanism:
As with "duri", see section 5.1.

Scope:
Global.

6. IANA considerations

This document includes two URN NID registrations (sections 5.1 and 5.2) that should be entered into the IANA registry of URN NIDs.

7. Security Considerations

duris and tdb$s are not any more reliable because they are dated.
URIs don’t contain enough information to supply the authority for
deciding what was or wasn’t at a given URI at a given date.

8. Acknowledgements

Many thanks to the many discussions on the relationship of URLs, URNs,
URIs and resource identifiers, as well as similar ideas, that have
been floated over the last many years.
This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

10. Author’s address

Larry Masinter
Adobe Systems Incorporated
345 Park Ave
San Jose, CA 95110
mailto: LMM@acm.org
http://larry.masinter.net
Tel: +1 408 536-3024

11. References


<http://www.w3.org/Provider/Style/URI>.


[TAI] "International Atomic Time", 
<http://www.bipm.fr/enus/5_Scientific/c_time/time_1.html>.
