Experience has shown that identifiers associated with persistent names have properties and requirements that may be somewhat different from identifiers associated with the locations of objects. This is especially true when such names are expected to be stable for a very long time or when they identify large and complex entities. In order to allow Uniform Resource Names (URNs) to evolve to meet the needs of the Library, Museum, Publisher, and Informational Sciences communities and other users, this specification separates URNs from the semantic constraints that many people believe are part of the specification for Uniform Resource Identifiers (URIs) specified in RFC 3986, updating that document accordingly. The syntax of URNs is still constrained to that of RFC 3986, so generic URI parsers are unaffected by this change.
1. Introduction

The Generic URI Syntax specification [RFC3986] covers both locators
and names and mixtures of the two (See its Section 1.1.3) and
describes Uniform Resource Locators (URLs) -- first documented in the
IETF in RFC 1738 [RFC1738] -- as an embodiment of the locator concept and Uniform Resource Names (URNs), specifically those using the "urn" scheme [RFC2141], as an embodiment of the names that do not directly provide for resource location. This specification is concerned only about URNs of the variety described in RFC 2141 [RFC2141] (i.e., those that use the "urn" scheme). URLs, other types of names, and any URI types that may not fall into one of the above categories are out of its scope and unaffected by it.

Experience with URNs since the publication of RFC 3986 has identified several ways in which their inclusion under its scope has hampered understanding, adoption, and especially extension in ways that were anticipated in RFC 2141. The need for extensions to the URN concept is now being felt in some communities, especially those that include libraries, museums, publishers, and other information scientists.

In particular, the Generic URI Syntax specification goes beyond syntax to specify the meaning and interpretation of various fields, especially the "query" and "fragment" ones. This specification excludes URNs from those definitions of meaning and interpretation so that RFC 3986 applies to their syntax only. The meaning --and any more specific syntax rules-- for those fields for URNs are now defined in a URN-specific document [RFC2141bis].

In its present form, it is intended primarily to focus WG discussions.

This draft does not discuss issues about DDDS resolution or conversion to (and interpretation of) URCs or URN "resolution" more generally. If any of those topics need to be addressed, it should be in other documents. Because URCs (as specified in RFC 2483 [RFC2483] or elsewhere) have not been significantly implemented or deployed, discussion of them is probably out of scope for the WG at this point. The document also does not discuss alternatives to URNs, either those that might use a different scheme name within the RFC 3986 URI framework or those that might use a different framework entirely.

2. Pragmatic Goals

Despite the important background and rationale in the sections that follow, the change made by this specification is driven by a desire to avoid philosophical debates about terminology or ultimate truths. Instead, it is motivated by three very pragmatic principles:
1. Try to accommodate all of those who think URNs are necessary, i.e., that they can and should be usefully distinguished in certain respects from other URIs, at least those that have been defined prior to this document. In particular, provide a foundation for extensions to the URN syntax allowed by and defined in RFC 2141 to support requirements encountered by some of those communities.

2. Try to avoid getting bogged down in declarative statements about definitions and debates about what is and is not correct in the abstract.

3. Avoid a fork in the standard that leads to multiple, conflicting, definitions or criteria for URNs.

In addition, this document is intended to move past debates about whether or not URNs are intended to be parsed at all (i.e., whether a "urn"-scheme URI is simply opaque to a URI parser once the scheme name is identified and, if not, how much of it is actually expected to be understood and broken into identifiable parts by such a parser. The assumption here is that parsing into the components identified in RFC 3986 will be performed but that any meanings or interpretation assigned to those components (including that applicability of the normal English meanings of such terms as "query" or "fragment" are a matter for URN-specific specifications.

3. The role of queries and fragments in URNs

Part of the concern that led to this document was a desire to accommodate URN components that would be analogous to the query and fragment components of generalized URNs. For many cases, the analogy cannot be exact. For example, RFC 3986 ties the interpretation of fragments to media types. Since media type is a function of specific content, URNs that are never resolved to particular content cannot have an associated media type. Similarly, while the syntax for queries (and fragments) may be entirely appropriate for URN use, terminology like "Service Request" (see Appendix B to the "URNs are not..." draft [Appendix B"ServiceRequests] for additional discussion) may be more suitable to the URN context than "query" (if, indeed, the query portion of the URN is where those requests belong).

These issues are discussed as questions facing the WG in Section 6 below.
4. Changes to RFC 3986

This specification removes URN semantics from the scope of RFC 3986. It makes no changes to the generic URI syntax. That syntax still applies to URNs as well as to other URI types. Even as regard to semantics, it has no practical effect for URNs defined in strict conformance to the prior URN specification [RFC2141] or the associated registration specification [RFC3406].

In particular, the generic URI syntax for "queries" (strings starting with "?" and continuing to the end of the URI or to a "#") and "fragments" (strings starting with "#" and continuing to the end of the URI) is unchanged, but the terms "query" and "fragment" become, for URNs, terms of convenience that are defined in URN-specific ways.

5. Other Required Actions

The basic URN syntax specification [RFC2141] was published well before RFC 3986 and therefore does not depend on it. Successors to that specification will need to fully spell out, or reference documents that spell out, the semantics and any required within-field syntax of URNs, using great care about generic or implicit reference to any URI specification.

6. Alternatives and comparison

[[Note in draft: temporary section to facilitate WG discussion]]

If this draft is approved, the WG will then have a number of other choices to make. They include:

6.1. Terminology and Information Location.

RFC 3986 syntax appears to allow three components of a URI in which we could put information for extending URNs past the "urn:nid:nss" syntax of RFC 2141. The syntax that introduces each of these is reserved for future use by RFC 2141 (Section 2.3.2). They are as follows:

path segment(s). The NSS string could be extended to allow one or more "path segments", introduced by "/" and terminating with the next "/", a "?", a "#", or the end of the URI. These path segment elements have been referred to as "facets" on the mailing list. If they are to be used, the WG will need to settle on what they should be called.
query. The URN syntax could be extended by use of what 3986 refers to as a "query", represented as a string that starts with "?" and extends to the first "#" or the end of the URI.

fragment. The URN syntax could be extended by use of what 3986 refers to as a "fragment", represented as a string that starts with "#" and extends to the end of the URI.

The WG will need to determine which of these fields to use (it could allow or require more than one of them, see below) and what to call them. The terms "path segment", "query", and "fragment" have the advantage of being traditional and associated in many people's minds with the corresponding delimiter. On the other hand, the normal conception of what those terms mean (including any semantics associated with them in 3986) may not be a good match for the needs of URNs. In particular, if a string starting in "?" were going to be treated as a collection of "Service Requests", calling that a "query" may strike some people as odd.

Allowing more than one of these components will probably require that the WG understand and document the semantic relationship among them (see below).

6.2. Comparison and "Part of the URN"

There has been fairly extensive discussion of what is compared when one compares URNs for equality. There has been a separate, but possibly equivalent, discussion about what elements associated with a URN identify things. The discussions have particularly emphasized, whether any of path segments, queries, or identifiers that are allowed participate in such comparisons or identification. As with other topics, some WG participants believe the answers are obvious, but don't agree on what they are. Others make a distinction about terminology (e.g., what is "part of the NSS") and assume that it answers the questions. The WG will need to figure out whether these discussions are the same and how to resolve the questions they imply.

6.3. Applicability of components.

The WG will need to decide whether whatever components are allowed are allowed on a per-NID basis or, at least syntactically, across the entire collection of URNs, remembering that, as far as 3986 is concerned, some things have traditionally been associated with schemes and all URNs are formally part of the same scheme. As noted above, RFC 3986 ties the interpretation of fragments to media types, but that is probably not meaningful for URNs, especially URNs that are never resolves to objects. Part of this requires deciding what should happen when a component is specified that is not applicable to
the particular NID-identified namespace. At least part of the web tradition has been to simply ignore such fields but that may not be the right answer for URNs, especially if one or more of them participates in comparisons (see above).

6.4. Internal syntax.

As long as the conditions for terminating substrings were not violated, the WG could decide on syntax within the components that are to be allowed, possibly including defining syntax for identifying keywords and defining or reserving some or all such keywords. Put differently, it may be important to decide whether "a query" is a series of related terms or components, possibly to be applied serially or whether it has components that are assumed to be independent and unordered. The latter choice may or may not interact with considering query components (or some of them) as "Service Requests".

6.5. Extended, embedded, base, and derived URNs

There has been discussion on the mailing list of different types of URNs or near-URNs using at least the above terms. It is not clear whether, once the issues above are resolved, those terminology distinctions will be trivial or whether they represent additional issues that the WG will need to resolve.

Note that this may interact with a discussion on the mailing list (off-topic for this document) about embedding URNs in HTTP or other URLs that locate a particular resolution or information-obtaining system. It may also interact with potential revised registration templates for ISSNs, ISBNs, and other existing URN namespaces and hence with the transition discussion [URN-transition].

7. Acknowledgments

This specification was inspired by a search in the IETF URNBIS WG for other alternatives that would both satisfy the needs of persistent name-type identifiers and still fully conform to the specifications and intent of RFC 3986. That search lasted several years and considered many alternatives. Discussions with Leslie Daigle, Juha Hakala, Barry Leiba, Keith Moore, Andrew Newton, and Peter Saint-Andre during the last quarter of 2013 and the first quarter of 2014 were particularly helpful in getting to the conclusion that a conceptual separation of notions of location-based identifiers (e.g., URLs) and the types of persistent identifiers represented by URNs was necessary. As noted below, Juha Hakala provided much of the text on which Appendix B.1 was based. Peter Saint-Andre provided significant text in a pre-publication review. The author also appreciates the
efforts of several people, notably Tim Berners-Lee, Larry Masinter, Keith Moore, Juha Hakala, Julian Reschke, Lars Svensson, Henry S. Thompson, and Dale Worely, to challenge text and ideas and demand answers to hard questions. Whether they agree with the results or not, their insights have contributed significantly to whatever clarity and precision appears in the text.

The specification was changed considerably and its focus narrowed after an extended discussion at the WG meeting during IETF 90 in July 2014.

8. Contributors

Juha Hakala contributed most of the text of Appendix B.1.

Contact Information:
Juha Hakala
The National Library of Finland
P.O. Box 15, Helsinki University
Helsinki, MA FIN-00014
Finland
Email: juha.hakala@helsinki.fi

9. IANA Considerations

This memo is not believed to require any action on IANA’s part. In particular, we note that there are a collection of "Uniform Resource Identifier (URI) Schemes" that does not include URNs and a series of URN-specific registries that do not rely on the URI specifications.

10. Security Considerations

This specification changes the semantics of URNs to make them self-contained (as specified in other documents), relying on the generic URI syntax specification for syntax only. It should have no effect on Internet security unless the use of a definition, syntax, and semantics that are more clear reduces the potential for confusion and consequent vulnerabilities.

11. References

11.1. Normative References

11.2. Informative References

[ DeterministicURI ]

[IETF90-URNBISWG]


[RFC2141bis]


[ServiceRequests]

[URN-transition]

[URNBIS-MailingList]
Appendix A. Background on the URN - URI relationship

The Internet community now has many years of experience with both name-type identifiers and location-based identifiers (or "references" for those who are sensitive to the term "identifier" -- see Appendix B.1). The primary examples of these two categories are Uniform Resource Names (URNs [RFC2141] [RFC2141bis]) and Uniform Resource Locators (URLs) [RFC1738]). That experience leads to the conclusion that it is impractical to constrain URNs to the high-level semantics of URLs. The generic syntax for URIs [RFC3986] is adequately flexible to accommodate the perceived needs of URNs, but the specific semantics associated with the URI syntax definition -- what particular constructions "mean" and how and where they are interpreted -- appear to not be. Generalization from URLs to generic Uniform Resource Identifiers (URIs) [RFC3986], especially to name-based, high-stability, long-persistence, identifiers such as many URNs, has failed because the assumed similarities do not adequately extend to all forms of URNs. Ultimately, locators, which typically depend on particular accessing protocols and a specification relative to some physical space or network topology, are simply different creatures from long-persistence, location-independent, object identifiers. The syntax and semantic constraints that are appropriate for locators are either irrelevant to or interfere with the needs of resource names as a class. That was tolerable as long as the URN system didn’t need additional capabilities (over those specified in RFC 2141) but experience since RFC 2141 was published has shown that they are, in fact, needed.

Appendix B. Three views of locator-identifier separation

Beginning in the 1990s with the first discussions of generalizing HTTP-style URLs to more general, "URI" forms with more or less different properties, there have been controversies between people and communities with strongly-held views about whether the differences between "locators" and "identifiers" are real, whether the categories are actually disjoint (RFC 3986 says that they are not), and, if real differences exist, how they are manifested and what their interests are. The subsections below are intended to at least partially capture different views of those issues. They are included here in the hope that they will assist with focusing discussion and reduce the frequency with which arguments are repeated. It is almost certain that the community does not have consensus on all of the points made below and that these blocks of text should be moved into other documents if they should be retained at all.
B.1. A Perspective on Locations and Names

Content industries (e.g., publishers) and memory organizations (e.g., libraries, archives, and museums) invest a lot of resources on naming things and the topics of naming and classification are important information science issues. Tens, if not hundreds, of millions of persistent identifiers have been assigned during the last decade.

Several identifier systems have been developed for persistent and unique identification of resources. When there is a real need to preserve something important (such as scientific publications, research data, government publications, etc.) for the long term, URNs or other persistent identifiers are used; URLs (or other generic URIs) are not being used for identification or even linking purposes.

Naming and locating, e.g., for library resources, are both complex activities which have different aims. Traditionally, naming and locating resources have been separate activities, and the rules for the former are much more stringent than for the latter. The same principles are being applied to digital materials as well as more traditional ones. In a library, any book, be it printed or digital, has both unique and persistent International Standard Book Number (ISBN) and non-unique (each copy has its own location information) and short-lived location information which cannot be trusted in the long run. ISBN never changes, but both shelf locations and Web addresses usually do, many times during the book’s life span.

Giving location information a role in identification would not only force libraries to adopt different policies for printed and digital content, it would also undermine the value of existing identifier systems. Let us assume that ten people independently upload a copy of an electronic book into different locations in the Web. Are all these ten URLs valid identifiers of the book? And what is their relation to the ISBN or other identification information of the book such as its title?

From the perspective of the communities who depend on persistent identifiers, critical issues include:

1. Resource identification has to be a managed process. Assigning URIs generally is not. Although it may be possible to introduce some level of control to URI assignment, a user cannot determine whether some URI is reliable or not.

2. Anyone may assign new URIs to resources even if these resources already have proper identifiers assigned to them. Claiming that these URIs actually identify something undermines the value of proper identifiers.
3. There is no 1:1 relation between the resource identified and URIs. An e-book in the Web may be represented as 1-n files (URIs), and a single file may contain several books. And books are simple, we need to name very complex objects such as research data sets, or some component parts within these complex data sets.

4. One resource such as a scientific article is typically available from multiple locations, including (for instance) the publisher's document supply service, a university's open repositories and other cooperative repository systems, legal deposit collections and the Internet archive. A resource should have one and only one identifier of a given type; URIs do not meet this requirement.

5. URIs relate to instances (copies) of resources, whereas traditionally identification has much broader scope. Identifiers may be assigned to, e.g., an immaterial work (such as Hamlet), its expressions (e.g. Finnish translation of Hamlet), and manifestations of works and expressions (e.g. PDF version of Finnish translation of Hamlet).

6. Over time, different resources (or different versions of the same resource) may be found from the same non-URN URI. A user has no way of knowing whether the resource has changed. One of the basic principles for proper identifier systems is that the same identifier is never assigned to another resource. In general, URIs do not meet this requirement.

7. Persistent identification must be available for resources which are available only in databases and other environments that are often identified today as "deep web". URIs for these resources tend to be very complicated and it will be difficult to keep them alive even with the help of DNS redirection when e.g. the underlying database management system changes.

8. The role URI fragment and query could or should have in identification is unclear and the statements in RFC 3986 are definitely problematic from the points of view of existing identifier systems and management of naming.

Does "fragment" identify a location or a certain section of a resource? In the evolving set of URN Internet standards, fragment will not be a part of the Namespace Specific String. Then fragment only indicates a place / segment within the identified resource, but does not identify it. If fragment had a role in identification, fragments would extend the scope of existing standard identifiers to component parts of
resources. For instance, anyone could use URN based on ISBN + fragment to identify chapters of electronic books.

Things get even more complicated with "query" since what the combination of an identifier and a query resolves to may not have anything to do with the original resource. For instance, a URN based in ISBN + query may resolve to the metadata record describing the book. These records have their own identifiers which are not based on ISBNs.

9. For many organizations, persistence means decades or centuries. Anything that is protocol dependent will eventually fail. URLs do not change by themselves, but in the long run it is very difficult for people to not change them or the objects to which they point.

The mention of centuries is intentional. Content industries, memory organizations (such as national and repository libraries and national archives) and universities and other research organizations, need identifiers that will persist for hundreds of years. Such identifiers might even need to outlast the institutions themselves, and definitely should be usable even if current technologies such as the Web and the Internet cease to exist or are supplanted by something new (as unlikely as that might seem today).

In addition, operations on, or additional specifications about, names and the associated objects must be possible, as stable as the names themselves, and reasonably efficient. For example, if a URN were assigned to an encyclopedia that consisted of many volumes, it should be feasible to identify (and locate and retrieve if that were desired) a particular volume or even a particular article without accessing or retrieving the entire set.

B.2. A More Pragmatic Perspective

The subsection above provides an explanation of the reasons for this change and actually for a more radical separation of URNs from generic URIs. That explanation is not without controversy, especially from those who make different assumptions about the future, or even interpretations of the present, than many members of the community (and especially members of the communities described in that section). Some of those who do not accept the explanation above simply do not recognize and accept the distinctions on which it, and URNs more generally, are based, including the name-locator distinction. In some cases, opposition to that explanation is quite
pronounced, involving fundamental differences in philosophy that move beyond mere differences of opinion.

Like most controversies in which one group does not accept the definitions, facts, or logic of another, the differences are unlikely to be resolved by further discussion, no matter how sensible and patient. The material in this appendix is provided for the benefit of those who cannot accept Appendix B.1 or consider the discussion there to be meaningless.

Put differently, the issue is ultimately not whether the perspective that Appendix B.1 reflects is, in some universal epistemology, correct or incorrect or even whether the consequences and implications of the introduction of the web and/or digital media renders it hopelessly obsolete. If only in their manifestation through national repository libraries and archives and setters of standards for them -- activities that have far more formal authority than the IETF or even W3C -- The community involved is relevant and legitimate. If the IETF wishes to maintain authority over things that are called URNs, then those perceived needs probably need to be accommodated in some reasonable way... where "reasonable" is defined as much or more by those communities as by the IETF one.

Independent of the details of the discussion above, in the case of URNs, the IETF is faced with a pair of problems that are ultimately faced sooner or later by all voluntary standards bodies: nothing except quality and broad community consensus prevents a standard from being ignored in the marketplace and nothing prevents another body from creating a competing standard. The effort required to create a competing standard can be increased and its potential for confusion can be reduced somewhat by various measures -- measures the IETF has rarely tried to actually use -- but those measures are rarely effective when the other body is convinced that they have legitimate and significant needs that differ from the original standard. Because of those problems, the key question for the URN effort is ultimately not whether a clear enough distinction exists between names and locator or location-based information, nor whether "persistent" can be defined clearly enough, nor even whether the communities and requirements described in Appendix B.1 are valid or will be judged valid in retrospect in a few decades or centuries. Instead, the question is whether the IETF is willing to evolve and adapt the URN definition to accommodate those perceived needs or whether if prefers to have that work done elsewhere, either by adoption in the broader community and marketplace of a different approach or, potentially, even a competing URN standard. If, in the long run, those other communities and perspectives turn out to be wrong, the additional features will atrophy. But that would be true whether they are specified and standardized in the IETF or elsewhere.
B.3. A more radical (or most conservative) view of URNs and their role

[[CREF3: The text in this subsection was derived from an on-list discussion. I believe it represents an even stronger position than RFC 3986 takes although I think similar positions have come up in other discussions. Because of its origins, the writing style is somewhat different from the rest of this document. Again, this text is provided for convenience and is not expected to survive into RFC publication--JcK ]]

The essence of this position is that URNs are "just" names and that, insofar as one can talk about location or resolution services of various types, they are data associated with the URN (or underlying name) and are not only not part of the URN but they are useful only for constructing locator-type URIs to which the URN (name) is an argument.

Suppose we have a URN that looks like

urn:isbn:1-4012-9876-1

It is really just a name. Associations with objects are someone else’s problem. There is actually no requirement that an object exist, only that the publisher/registrant have sufficient intention to create an object to assign the code. Now a query about metadata associated with that name makes perfect sense although there are questions about how far it should go (see below). For example, one could invoke

urn:isbn:1-4012-9876-1?publisher

and, modulo some issues about queries being defined by the resource, have a more than reasonable expectation of getting back "DC Comics". But, since that is a name and not an object or the location of an object, I don’t know what a fragment is. One could certainly write

urn:isbn:1-4012-9876-1#publisher

or

urn:isbn:1-4012-9876-1#1

but, assuming one knows how ISBNs are constructed, the result would presumably be

1-4012
and not anything useful, since there is no object to retrieve and evaluate with regard to either media type or content.

If we are going to maintain a strong name - object distinction, this approach makes a certain amount of sense.

An extreme version of the argument that we can’t have fragments on URNs because they are just names, not objects, might lead to the claim that the only way one gets "Section 2" of that book is with something like

http://school-library.psl1234.k12.ma.example/?urn="isbn:1-4012-9876-1"&Section=2

or, in two more general cases:

myFavoriteLibraryRetrievalScheme://library-domain.example/?urn="isbn:1-4012-9876-1"&Section=2

or maybe

http://www.generic-bookseller.example/?urn="isbn:1-4012-9876-1"#Section=2

In all three of those cases and some other variations we can thinks of, the URN is, itself, stable and persistent. Neither the two schemes nor the domain parts associated with them need be. If the fragment that refers to a section is valid, it is too (that doesn’t make it part of the name -- that is a separate question). The retrieval/ resolution system is not a property of the URN. Instead, the URN is a name-type argument --an object identifier-- used as input to the retrieval system.

Appendix C. Change Log

[[CREF4: RFC Editor: Please remove this appendix before publication.]]

C.1. Changes from draft-ietf-urnbis-urns-are-not-uris-00 to -01

- Revised Section 1 slightly and added some new material to try to address questions raised on the mailing list.
- Added Section 2, reflecting an email exchange.
- Added a Security Considerations section, replacing the placeholder in the previous version.
o Added Appendix B.2 and inserted a note in the material titled "A Perspective on Locations and Names" pointing to it (that material is in Appendix B.1 in the current version, but was Section 2 and then Section 3 in earlier versions).

o Added temporary Appendix B for this version only.

o Enhanced and updated the Acknowledgments section.

o The usual small clarifications and editorial changes.

C.2. Changes from draft-ietf-urnbis-urns-are-not-uris-01 to draft-ietf-urnbis-semantics-clarif-00

o Changed title and file name to better reflect changes summarized below. Note that the predecessor of this document was draft-ietf-urnbis-urns-are-not-uris-01.

o Revised considerably as discussed on the mailing list and at IETF 90. In particular, the document has been narrowed to change semantics only without affecting the relationship to URI syntax and the document title and other details changed to match.

o Dropped much of the original Introduction (moving it temporarily to an appendix) and trimmed the abstract to be consistent with the new, more limited scope.

o Revised Appendix B.2 to make "perceived requirement" more clear.

o Removed the former Appendix B, as promised in the previous draft, moved considerably more text into appendices, and added some new appendix text. Note that the earlier text is temporarily referenced in Appendix B.3 above. If we intend to keep that appendix material, we will have to drag at least part of the text back in from the earlier draft.

o Added new Section 6 to discuss the next round of decisions the WG will have to make, assuming this provisions of this specification are approved.

Author’s Address
John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA  02140
USA

Phone: +1 617 245 1457
Email: john-ietf@jck.com