DOI® System and Internet identifier specifications
Version 2.3

A standard represents an agreement by a community to do things in a specified way to address a common problem. Whilst the DOI community has developed the DOI System, it has also ensured conformance with relevant generic external formal standards. This factsheet discusses those relevant in the Internet communities (IETF and W3C). There has been considerable debate here on the issue of generic standards for naming objects.

Comparing generic identifier standards

A DOI® name differs from commonly used Internet pointers to material such as the URL, because it identifies an object as a first-class entity, not simply the place where the object is located. A DOI name also differs from identifiers such as the International Standard Book Number (ISBN), International Standard Recording Code (ISRC), etc., because it can be associated with defined services and is immediately actionable on a network.

The comparison of persistent identifier approaches is difficult because they are not all doing the same thing. Imprecisely referring to a set of schemes as 'identifiers' doesn't mean that they can be compared easily. Similarly, when any two technologies (e.g., two web browsers) are compared, the criteria used for comparison must be defined.

URI, URL, and URN

Historically there was ambiguity and confusion in the use of these terms. RFC 3986 (2005) aimed to end this by stating that a URI can be classified as a locator, a name, or both. In this view, the term URL refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource; the term URN has been used historically to refer to both URIs under the "urn" scheme (RFC 2141) which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name.

RFC 3986 requires that the terms URL and URN be deprecated. This brings a uniformity to the technical treatment of all URIs. However the risk of confusion remains, from:

- cited documents which rely on earlier, now superseded, statements of the position;
- the use of one simple top level term (URI) may hide useful distinctions which some users, e.g., librarians, may wish to make between a unique name and a location, for example when a named resource is available at multiple locations;
- considerations of how widely used non-web identifiers (such as ISBNs, RFIDs, social security numbers, etc) relate to URIs, which can lead to:
    - confusions re identifier, representation, and access mechanism;
    - lack of appreciation of identifier usage outside the WWW;
    - use for non-digital referents; and
    - the requirement to perceive the web as only part of the Internet and the Internet as only part of information.

In the view now considered by RFC 3986 to be obsolete, URIs have two subclasses: URN (identifying names) and URL (identifying single locations). In the RFC 3986 view, web-identifier schemes are all URI schemes, as a given URI scheme may define subspaces; some of these may be access mechanisms (e.g., "http:") whilst others may be namespaces (e.g., "urn:").

There are strong arguments against all URIs being expressed forever as http protocol strings: see a good summary on the IETF URI Review mail list at http://www.ietf.org/mail-archive/web/uri-review/current/msg00978.html

URI

Uniform Resource Identifier (RFC 3986) provides an extensible means for identifying a resource within the World Wide Web. Each URI begins with a scheme name that refers to a specification for assigning identifiers within that scheme; each scheme's specification may further restrict the syntax and semantics of identifiers using that scheme.

URI specification defines (1) an implementation to access a location on a file server, commonly accessed using the http protocol though other protocols are allowed; (2) a syntax for referencing, through which e.g., ISBNs can be specified as URIs. The network path of the URI is implicitly DNS based; original URI specifications that assume the URI to be opaque have been overtaken by practical usage which assumes that the initial URI parser will look for meaningful characters (such as dot and slash).

The use of URIs as identifiers that don't actually identify network resources (for example, they identify an abstract object, or a physical object) was recognised as an unanswered problem in RFC 3305. This usage is important in any semantic application. To address this, the info URI scheme (RFC 4452: http://info-uri.info) was developed by library and publishing communities for "URIs of information assets that have identifiers in public namespaces but have no representation within the URI allocation". OpenURL adopts it and was a key the motivation for it. InfoURI registrations can be made by anyone, not necessarily the authority for a particular namespace. DOI is registered in the infoURI scheme.

URN

Uniform Resource Name (RFC 2141) is a specification for defining names (identifiers) of resources for use on the Internet. Locations are assumed to be independent of names. URN resolution is still an active topic of discussion, especially in the library community (e.g. for treatment of National Bibliography Numbers as URN in RFC 3188). RFC 2141 defines (1) a formal registration process as a urn namespace, and (2) accompanying specifications to implement a series of functional requirements for such namespaces. Existing identifiers may thereby be specified as a URN: e.g. an ISBN as urn:isbn:9789521061547; such identifiers may be implemented using a specially written URN plug-in and resolved to URLs: functionally this gives nothing beyond that achieved by coherent management of the corresponding URLs.

URN architecture assumes a DNS-based Resolution Discovery Service (RDS) to find the service appropriate to the given URN scheme. However no such widely deployed RDS schemes currently exist: browsers cannot action URN strings without some additional programming in the form of a "plug-in". These carry no guarantee of ready interoperability with other deployments, which may require a different plug-in for each implementation and may use conflicting data approaches. Therefore most existing URN implementations embed the URN as a http URI which contains the URL of the relevant resolution service (e.g. for the URN form of the ISBN shown above, resolved via the Finnish national URN service http://urn.fi, the actionable form of the URN is http://urn.fi/URN:ISBN:978-952-10-6154-7). There is no global service aware of national and/or regional URN resolution services, but there are some proposals to provide one (e.g. http://www.persid.org.)

The set of URNs, of the form "urn: nid: nnnnnn", is a URN namespace. ("nid" is here a URN namespace identifier, neither a "URN scheme", nor a "URI scheme.") The official IANA list of registered NIDs at http://www.iana.org/assignments/urn-namespaces lists 40 registered NIDs; many of these are not widely used as URNs (e.g., ISSN, ISBN).

DOI is not registered as a URN namespace, despite fulfilling all the functional requirements, since URN registration appears to offer no advantage to the DOI System. It requires an additional layer of administration for defining DOI as a URN namespace (the string urn:doi:10.1000/1 rather than the simpler doi:10.1000/1) and an additional step of unnecessary redirection to access the resolution service, already achieved through either http proxy or native resolution. If RDS mechanisms supporting URN specifications become widely available, DOI will be registered as a URN.

URL

Uniform Resource Locator (RFC 1738) is a location on a file server in the WWW; redefined in RFC 3986 as "a type of URI that identifies a resource via a representation of its primary access mechanism (e.g., its network "location"), rather than by some other attributes it may have". In this view "URL is a useful but informal concept" (RFC 3305). In practice, it identifies a single location, and therefore is widely used incorrectly as a (mutable) identifier of the resource at that location (so the same resource at two URLs would have two URL "identifiers"). This bad practice arose from the failure to distinguish name and location in early WWW development. Adding to the problem, URLs carry semantics of the Domain Name they are based on and are therefore unsuitable as opaque identifiers; they may also be contextually qualified. URLs are pervasive as the foremost mechanism of location specification throughout the WWW, but less useful outside it.

Attempts to circumvent the problem of using URLs as citable identifiers by developing persistent identifier alternatives are well documented (PURL, DOI, ARK, etc.).

A DOI name may be represented as a URL (http string) by prefacing the string http://dx.doi.org/ to the DOI of the document (e.g., to resolve the DOI name 10.1000/182, enter into a browser the address: http://dx.doi.org/10.1000/182). Web pages or other hypertext documents can include hypertext links in this form.

DOI functional requirements
The DOI system is designed to fulfil several additional functional requirements which offer significant advantages in generic naming, notably:

- Neutral as to implementation. DOI allows but does not require http or other protocols. The design principle is that DOIs are not specific to the web or any other implementation (e.g., information may be delivered in non-web platforms such as PDAs). DOI is designed to be applicable in any environment on the Internet (the global information system linked by a globally unique address space based on the Internet Protocol (IP) using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite).
- Flexible as to implementation. The DOI system has been designed around a data and transaction model that can work in a wide variety of environments. The current implementation works well with, but does not require, http or other web protocols, and can be used in any environment on the Internet (the global information system linked by a globally unique address space based on the Internet Protocol (IP) using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite).
- Granularity of naming and administration at the object level. Allows but does not mandate coarser level granularity tools such as domain names. Specifically, DOI resolution in native resolver form does not require the use of the DNS (Domain Name System): the DNS administrative model argues against using it as a general-purpose name system and has well-recognised problems of security and updating.
- Neutral as to language/character set. Compatible with, but not restricted to, the ascii character set. DOI names can use the Unicode capability of the Handle System to develop DOI names in Japanese, Chinese, etc., characters. The current DOI syntax restricts initial implementations to ascii simply for ease of adoption, but is intended to be widened (backward compatibility) to Unicode in a future revision.
- Multiple resolution to typed data offers the possibility of expressing semantic relationships.

- Social infrastructure providing persistence through organizational backup, data integrity measures, etc.

Other internet persistent identifier schemes

The Handle System is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. It is the underlying resolution component for the DOI System. The Handle System is the most appropriate persistent identifier management system for the DOI System: see the related DOI factsheet "DOI System and the Handle System" (http://www.doi.org/factsheets/DOIHandle.html). There are several other Internet persistent identifier mechanisms proposed by individuals or organisations, having various emphases on social infrastructure or technology. There are several studies of persistent identifier management sustainable infrastructure and services available, such as the PILIN project (http://www.pilin.net.au/Closure_Report.pdf).

A persistent uniform resource locator (PURL) is a Uniform Resource Locator (URL) that does not directly describe the location of the resource to be retrieved but instead describes an intermediate (more persistent) location which, when retrieved, results in redirection (e.g. via a 302 HTTP status code) to the current location of the final resource. PURLs are said to be "an interim measure, while Uniform Resource Names (URNs) are being mainstreamed, to solve the problem of transitory URIs in location-based URI schemes like HTTP". (http://en.wikipedia.org/wiki/PURL)

Extensible Resource Identifier (XRI) is a scheme and resolution protocol for abstract identifiers. While the Handle System is focused on the secure administration and resolution of identifiers into handle records, XRI is more concerned with defining properties and semantics of identifiers to allow for extensible namespace resolution, segmenting of identifiers, identifier cross referencing, and semantics for accessing resources. The Handle System uses its own protocol over udp, tcp and http; XRI uses its own XRDS over http or https. Handles could be implemented in XRI as a internal resolution system within the XRI resolver, or as a registered XRI Service End Point (SEP). See also http://en.wikipedia.org/wiki/XRI.

Archival Resource Key (ARK) is a Uniform Resource Locator (URL) that provides a multi-purpose identifier given to information objects of any type. ARKs contain the label ark: in the URL, which sets the expectation that the URL terminated by '?' returns a brief metadata record, and the URL terminated by '??' returns metadata that includes a commitment statement from the current service provider. (http://en.wikipedia.org/wiki/Archival_Resource_Key).

Last revised May 2010

This is one of a series of DOI factsheets. To see the latest version of this factsheet online, and to see the other factsheets, go to: http://www.doi.org/factsheets.html