RESOURCE IDENTIFIER

Greg Janée
Institute for Computational Earth System Science
University of California at Santa Barbara
Santa Barbara, CA 93106-3060
gjanee@alexandria.ucsb.edu

SYNONYMS

document identifier; GUID; URI

DEFINITION

In a networked information system, a *resource identifier* is a compact surrogate for a resource that can be used to identify, retrieve, and otherwise operate on the resource. An identifier typically takes the form of a short textual string. An identifier must be *resolved* to yield the associated resource.

SCIENTIFIC FUNDAMENTALS

Resource identifiers can be broadly characterized as being *locations*, which identify resources by where they reside, or as being *names*, which identify resources by intrinsic properties of the resources. This distinction is not absolute, and identifiers can exhibit characteristics of both classes. Furthermore, whether name or location, every identifier requires a resolution mechanism to retrieve the associated resource. Nevertheless, the distinction between locations and names is useful in characterizing the properties of identifiers and the relationships between identifiers and resources. Consider the answers that might be given to the following questions:

> *Can two distinct, yet identical resources have the same identifier?*
> *If a resource changes, must its identifier change?*

If the answer to these questions is yes, then we would probably consider the identifiers to be names; if no, locations. To take two examples, International Standard Book Numbers (ISBNs) are names: all copies of a book are identified by the same ISBN, and a revision of a book causes a new ISBN to be issued. By contrast, HTTP URLs on the World Wide Web are locations: two identical but distinct Web resources must necessarily have different URLs, and a resource's URL is independent of its potentially changing content.

The distinction between names and locations is codified in the World Wide Web architecture [2], which roughly partitions the universe of Uniform Resource Identifiers (URIs) into Uniform Resource Locators (URLs) and Uniform Resource Names (URNs) [3].

Two desirable characteristics of resource identifiers are uniqueness and persistence.

Uniqueness

Uniqueness is the property that an identifier resolves to a single resource. The converse property—that every resource is identified by a single identifier—is generally desirable, but is often not enforced (or even enforceable) by systems that allow free generation of identifiers. A corollary of Metcalfe's Law on the "network effect" states that the value of a given resource can be measured by the number and value of the other resources that reference it; duplicate identifiers (or identifier "aliases") cause resources to be undervalued [2].

Schemes for generating universally unique identifiers generally fall into two categories. Schemes in the first category guarantee uniqueness by incorporating into each identifier unique characteristics of the identified resource, for

example a content-based signature such as an MD5 digest, or characteristics of the context in which the resource and/or identifier system reside, for example a network address and timestamp. UUIDs [4] incorporate both content-based and contextual characteristics.

Schemes in the second category guarantee uniqueness by acquiring identifiers from an "authority" that maintains a centralized store of previously-generated identifiers (identifier–resource associations are often stored as well). For scalability such systems are often arranged hierarchically so that a root authority, located at a well-known address, may delegate identifier generation and resolution requests to distributed sub-authorities. DNS and the Handle system [5] are two well-known examples of this approach.

Persistence

Persistence is the property that an identifier continues to reference the associated resource over time. Persistence is most visible when it fails, for example, when a URL on the World Wide Web is "broken" and no longer refers to the original resource (or to any resource at all).

Strictly speaking, persistence is not a property of an identifier; it's an outcome of the commitment of the operator of the identifier resolution system to maintain the association between the identifier and its resource. But different identifier schemes mitigate to different degrees known risks to persistence, and in this limited respect persistence can be thought of as a property of an identifier.

All persistence mechanisms fundamentally employ indirection: an identifier does not directly identify a resource, but instead identifies an intermediate quantity which is maintained by the resource owner to identify the resource even as the latter moves and changes over time. In principle this indirection mechanism may be hidden from the users of resource identifiers, but for scalability reasons it is typically exposed. For example, the Persistent Uniform Resource Locator (PURL) system [6] employs HTTP's redirection mechanism.

One of the risks to identifier persistence is the inevitable change in resource ownership and naming, which can induce concomitant changes to identifiers. One approach to mitigating this risk is issuing so-called "semantics-free" identifiers. For example, Digital Object Identifiers (DOIs) [7] are sequences of numeric digits having no external referent. However, the benefit of mitigating this risk to persistence must be balanced by the inscrutability of such identifiers to humans. Taking the opposite approach, OpenURLs [8] identify resources purely by semantics, specifically, by constraints placed on resource bibliographic metadata. For example, an identifier for an article in a scholarly journal might contain the name of the journal, the issue number, and the article title, all encoded into a URI. Such identifiers can be seen as being names as opposed to locations.

Robust hyperlinks [9] provide a different approach to persistence by combining a location (a URL) with a content-based signature (a set of words that distinguishes the resource); the signature can be used to search for the resource should the URL break. The Archival Resource Key (ARK) scheme [10] integrates into its identifier resolver system a protocol for obtaining persistence guarantees, policies, and other preservation-related metadata.

Other characteristics

Additional desirable characteristics of resource identifiers include global scope, global uniqueness, scalability, extensibility, machine readability, recognizability in text, and human transcribability [11]. There are substantial benefits to expressing identifiers as URIs, including hyperlinking, bookmarking, and indexing by Web search engines [2]. Identifiers that are subject to transcription errors (credit card numbers, to take one example) benefit from having error-correcting codes incorporated into them [12].

RECOMMENDED READING

[1] Hans-Werner Hilse and Jochen Kothe (2006). *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations*. London/Amsterdam: Consortium of European Libraries and European Commission on Preservation and Access. ISBN 90-6984-508-3.
`http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8`

[2] Ian Jacobs and Norman Walsh, eds. (2004). *Architecture of the World Wide Web, Volume One*. http://www.w3.org/TR/webarch/

[3] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter (2005). *Uniform Resource Identifier (URI): Generic Syntax*. IETF RFC 3986. http://www.ietf.org/rfc/rfc3986.txt

[4] Paul J. Leach, Michael Mealling, and Rich Salz (2005). *A Universally Unique IDentifier (UUID) URN Namespace*. IETF RFC 4122. http://www.ietf.org/rfc/rfc4122.txt

[5] Sam X. Sun, Larry Lannom, and Brian Boesch (2003). *Handle System Overview*. IETF RFC 3650. http://www.ietf.org/rfc/rfc3650.txt

[6] Keith Shafer, Stuart Weibel, Erik Jul, and Jon Fausey (1996). "Introduction to Persistent Uniform Resource Locators." *INET96 Proceedings* (Montreal, Canada; June 24–28, 1996). http://www.isoc.org/inet96/proceedings/a4/a4_1.htm

[7] Digital Object Identifier System. http://doi.org/

[8] Herbert Van de Sompel and Oren Beit-Arie (2001). "Open Linking in the Scholarly Information Environment Using the OpenURL Framework." *D-Lib Magazine* **7**(3). http://dx.doi.org/10.1045/march2001-vandesompel

[9] Thomas A. Phelps and Robert Wilensky (2000). *Robust Hyperlinks Cost Just Five Words Each*. University of California, Berkeley Technical Report No. UCB/CSD-00-1091. http://www.eecs.berkeley.edu/Pubs/TechRpts/2000/5442.html

[10] John A. Kunze and R.P.C. Rodgers (2007). *The ARK Persistent Identifier Scheme*. http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

[11] Karen Sollins and Larry Masinter (1994). *Functional Requirements for Uniform Resource Names*. IETF RFC 1737. http://www.ietf.org/rfc/rfc1737.txt

[12] Joseph A. Gallian (1996). "Error Detection Methods." *ACM Computing Surveys* **28**(3):504–517. http://doi.acm.org/10.1145/243439.243457