Delivering "Cool URIs" that do not change

Nick Juty[1], Nicolas Le Novère[1,2], Henning Hermjakob[1], and Camille Laibe[1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[2] The Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT, United Kingdom

Abstract. Uniform Resource Identifiers (URIs) are an accepted and standard way to identify data records and entities, providing a necessary element in data access, query and integration. Identifiers.org is an infrastructure that enables the generation (and resolution) of perennial and globally unique URIs for life science data, independently of where the data record is actually stored. This system allows the composition of 'cool URIs', based on the providers' accession numbers, which can be used for efficient and effective cross-referencing. We describe Identifiers.org, and compare it with other pre-existing identification systems, highlighting the additional benefits it offers.

1 Introduction

Identifiers are lexical tokens that can be used to 'name' entities or data. The primary function of an identifier is served through its unique association with a specific piece of information, and the longevity of that association.

In 2006, Tim Berners-Lee proposed a set of basic "rules"[1] to promote the interconnection of data on the World Wide Web. One of these rules encouraged the use of HTTP URIs, and further suggested that these URIs should not change[2]. The unambiguous and perennial identification of data entities on the World Wide Web is becoming increasingly important, especially with the semantic web effort[3] to build a Linked Data[4] cloud. The inter-linking of entities in this cloud is reliant upon the use of appropriately constructed canonical identifiers, which can effectively act as semantic anchors spanning the whole web of information. In order to facilitate the incorporation of data into this web of information, it is necessary to address entities through a single unique identifier.

Even before the advent of the semantic web, the need for identifiers already existed: all resources providing access to data sets required the use of identifiers to specify individual records. Over the years, several efforts were launched in the field of Life Sciences in order to uniquely identify pieces of information on a more global scale. In this review, we present the current landscape of such systems, and discuss their use and limitations.

2 Identification schemes

We describe below the six main identification schemes used to date within the field of the Life Sciences: Identifiers.org[5], DOI (Digital Object Identifiers[6], LSID (Life Science IDentifiers[7], LSRN (Life Science Record Names[8]), PURL (Persistent Uniform Resource Locators[9]), and Shared Names[10].

Identifiers.org

Identifiers.org[5] provides a means to identify data records through URIs. In addition to providing unique and perennial identifiers, the data underlying the URI itself can be accessed on the web, since these URIs are HTTP URIs (or URLs). This system directly uses the information contained in the MIRIAM Registry[11], a free-to-use, curated repository of data collections and namespaces. The primary information stored in the Registry are "data collections", which are defined as a set or pool of data concerning comparable entities, and referenced through a common identification scheme defined by the data provider responsible for that collection. Data collections are, by definition, location independent, since a particular set of data could be distributed by any number of physical resources. The Registry also uses the concept of "namespace", a string of characters that can be used to refer to a specific data collection.

Identifiers.org provides a generic URI generation framework which is based upon a combination of the data record identifier (assigned by the original data provider) and the collection (using its namespace) from which the record is drawn. For example, `http://identifiers.org/uniprot/P12345` identifies the record `P12345` from the UniProt Knowledgebase. Although the resulting URI is actually a URL, it is a location-independent one as this system effectively decouples the identification of the data itself from its location on the web, allowing the mapping of

a single identifier to any number of registered physical resolving locations (recorded in the Registry as "resources"). In the simplest cases, this allows the selection of alternative mirrors where entity information can be retrieved.

The manner in which information is captured in the registry underlying Identifiers.org allows the creation of URLs at varying levels of granularity: references can be built to data sets, individual records (location independent), and to specific resolving locations. Identifiers.org also handles content negotiation, allowing the user to retrieve resolving information about the identified entities (currently in HTML and RDF/XML, with plans to accommodate additional formats).

The incorporation of further data collections can be freely submitted online or requested in bulk (for example, incorporated from public cross-reference lists) to our curation team. Once submitted, the information is actively maintained to ensure that physical locations remain current and resolvable. To achieve this, processes are in place to facilitate the identification and correction of downstream errors, such as dead links, relocation of resources or infrastructure changes.

2.1 Life Science Identifiers

The Life Science Identifiers (LSIDs) specification[13] was published by OMG[3] in 2004 to provide location-independent resource identifiers in a URN form. The basis of this system relies upon the registration of a namespace with the 'LSID authority'[14]. Once assigned, identifiers can be constructed of the form 'urn:lsid:ubio.org:namebank:11815', and are composed of five parts[16]. These specify the scheme (*LSID*), Authority (*ubio.org*), namespace (*namebank*) and the record identifier (*11815*). There is also a trailing, optional revision component. These URIs are not directly resolvable. Instead, it is suggested that adopters of this system make use of the LSID HTTP proxy[15], which uses web services to resolve the LSID. Essentially, this process involves the suffixing of the LSID to the HTTP proxy stem, for example generating the following resolvable URL `http://lsid.tdwg.org/urn:lsid:ubio.org:namebank:11815`. LSIDs are therefore not directly resolvable themselves, but may be modified to be so. The early adopters of this system included IBM[17]. This scheme is still used within the Biodiversity community[18].

2.2 Life Science Record Names

Life Science Record Names[8] (LSRN) is based on a centralized repository of database abbreviations, and provides mappings to existing URL location records. This centralized web-based resource, while accessible and editable by anyone, is not accessible nor queryable through programmatic means. LSRNs are constructed simply by concatenating a *database abbreviation*, with a *record identifier*, conjoined through the use of a colon. This in itself does not constitute a URI; while an LSRN identifier, for example UniProtKB_Swiss-Prot:P12345 (which refers to a protein record listed in the UniProt Knowledgebase[19]), is not resolvable, it is relatively simple to transform it into a URL: `http://lsrn.org/UniProtKB_Swiss-Prot:P12345`.

2.3 Digital Object Identifiers

Digital Object Identifiers[6] (DOIs) were first used in applications around the year 2000, originating from a proposal in 1996 to develop an infrastructure for digital publishing. Administered by The International Digital Object Identifier Foundation (IDF), the DOI system offers a simple redirection service for those data providers who register with an appointed agency. The standard is specified by ISO 26324[22], and describes general principles for the creation, registration and administration of DOI names. The registration process requires the payment of a fee and each identifier created is charged additionally.

A DOI name is composed of a prefix that identifies the registrar of the name, and a suffix (chosen by the registrar) which identifies the data object to be associated with the DOI. To resolve DOIs, for example when embedded in a HTML document, they must be associated with a resolver. For instance, within `http://dx.doi.org/10.1093/nar/gkr1097`, *10.1093* specifies the registrar in the DOI registry, *nar/gkr1097* the record, and *http://dx.doi.org/* the resolver. DOIs are commonly used to identify published scientific articles, but can be employed to identify other types of data.

2.4 Persistent URLs

Persistent Uniform Resource Locators (PURLs) were introduced in 1995 by the OCLC[4] to provide URLs to permanently identify resources on the World Wide Web. They function by simple redirection through an intermediate

[3] `http://www.omg.org/`
[4] `http://www.oclc.org/`

resolution service, which associates the PURL with a target URL endpoint, returning the target to the client. If the URL endpoint changes over time, the PURL can be modified to reflect the new target. Hence, the URL location to which a PURL resolves may be, and most likely will be, modified over time. An example of a PURL is: http://purl.obofoundry.org/obo/OBI_0000225.

The creation or modification of PURLs requires registration with a PURL server. Most available PURL servers allow anybody to register.

2.5 SharedNames

The Shared Names[10] initiative is a more recent effort to assign URIs to publicly available information records. The initial population of this Shared Names list is intended to be derived from cross-reference lists provided by databases such as the Gene Ontology[20] and Enzyme[21] as well as from the LSRN repository.

Shared Names URIs, agnostic of encoding format, should allow the retrieval of an RDF document stating the original source of the record, its identifier, and providing links to various encodings of the record (HTML, XML, ASN, *etc*) as available from the data provider. In addition, this RDF document should contain links to resources belonging to other semantic web projects participating in the Shared Names initiative, as well as links to related external resources that build upon the primary record.

While there is no public infrastructure to judge this effort at present, it is likely that any redirection would be accomplished using a PURL server. Little more information is available about this fledgling effort at this time.

3 Discussion

Identification schemes described above differ in some key characteristics, such as uniqueness or resolvability. Moreover, though not a characteristic *per se*, it is important to understand the barriers that need to be overcome in the adoption of a particular scheme. These include costs that may be incurred, the openness of the system, and the appropriateness of the scheme under consideration. These aspects are discussed below, and a summary is presented in the Table 1.

3.1 Unique URIs

We define uniqueness in this context as the characteristic of a given identification system to strictly associate only one URI to an individual data record. Unique URIs, or at least the ability to limit the number of URIs for individual records, is desirable to minimize the amount of URI mapping activities that need to be performed during data gathering and consolidation/harmonisation tasks. In the case of PURLs, which can be created *ad hoc* by interested parties, this is clearly not the case. For instance, the PURLs http://purl.obolibrary.org/obo/GO_0015976, http://purl.uniprot.org/go/0015976 and http://purl.org/obo/owl/GO#GO_0015976 all identify the same Gene Ontology term. Since there is no queryable centralized registry, it makes the identification and reuse of existing PURLs difficult, at best, and thereby encourages the creation of even more PURLs.

3.2 Resolvable URIs and resolving cardinality

The resolvable URI characteristic represents the capacity for an identifier (employing a particular scheme) to be resolved to physical location(s) on the World Wide Web, in order to allow data access. The main advantage of Shared Names, PURLs and Identifiers.org URIs lies in their URL form, which means they are directly usable on the web. All three require the use of a resolver which provides the necessary redirection between the persistent URL and the non-persistent one actually used by a specific resource.

An intermediate position also exists with regard to resolvability: some schemes require the manipulation of an identifier, usually through its association with a separate resolver, to generate resolvable URIs. In the case of DOI names, this is often done for embedding the manipulated DOI name within, for example, a HTML page. A similar situation exists for LSRN identifiers, where the identifier (database name:record identifier) must be associated with a given stem URL. For instance, a web page may display an entity identifier such as P12345, with the underlying hyperlink pre-combined with the resolver (http://lsrn.org/UniProt:P12345 for LSRN).

Solutions based on PURLs (so potentially true for Shared Names too) have a drawback: there can be only a one to one mapping between the identifier (the persistent URL in this case) and the physical location (the non-persistent URL) where the information can be obtained. This is one criticism that is often levelled at PURLs: should the PURL fail to resolve, one would find it difficult to ascertain whether the PURL itself had failed, or the resolving location

endpoint. It should be noted, however, that the purpose of PURLs is subtly different to that of Identifiers.org URIs. PURLs are meant to address final resolving locations, which can be viewed as instances of records, while location-independent URIs, such as Identifiers.org and DOI, are meant to identify the records themselves.

Solutions relying on a resolver can theoretically (and in the case of Identifiers.org, actually do) handle a one to many mapping between the identifier (a URI) and the physical locations (URLs) which serve the data. The key difference between these two methodologies lies in whether the identifier for the information is separated from the resource that actually provides that information. Schemes such as PURL lock the two pieces of information together, hence are only able to provide a one-to-one mapping. At its registry level, Identifiers.org decouples the identifier from the physical resource, facilitating a more flexible one-to-many mapping and reflecting the true nature of data provision on the web.

A high resolving cardinality, exemplified by Identifiers.org, is desirable since the reliance on a single physical location through which to retrieve data is prone to single point of failure issues. Providing multiple potential resolving locations avoids these potential dead ends, and may provide further auxiliary information around the specified data record: for example, instances of records hosted on diverse databases may have different cross-references.

3.3 Cost and registration

Both cost and the need for registration form barriers to the adoption of a given scheme. Financial constraints may preclude the use of fee-based identification schemes, particularly for large data sets. Registration in the schemes described varies between the different efforts, sometimes requiring an institutional-level commitment, which may be hard to attain. Additionally, the need for data providers to register to the various schemes necessitates that they are aware of the reasons why this would be beneficial to the community at large, and are aware of the mechanisms by which it can be accomplished. This may well not be the case, since a particular data provider may not be involved in semantic web activities, and may only serve a smaller niche community of users.

Cost is a major concern for adoption of the DOI system. In addition to an annual fee for the registration of an organisation, a per-identifier fee of a few pence, negligible in the total production cost of a scientific publication, may be prohibitively expensive for large life science data collections.

The other schemes listed are all open to public submission to varying extents and by different mechanisms. LSID creation requires the data provider to register a namespace with an LSID Authority, placing the onus firmly with the data providers. While LSID may nominally be an existing standard, there appears to be little activity in recent years. Moreover, once an institute has registered and been assigned an *Authority*, it is then responsible for ensuring that there are no namespace collisions, and that each LSID uniquely identifies a single record. Since the authorisation to issue LSIDs is given at the institutional level, the lexical string specifying the Authority ID may be liable to change. Finally, since LSID Authorities (necessary to mint LSIDs) are created at the request of an institution, it would not be possible to generate official (data provider approved) LSIDs until the data provider themselves desired to make their data available through that system.

PURLs are free to use, but require registration to both create or modify. Registration is done at the level of an individual PURL resolver, with there being no requirement to furnish equivalent registration information between different resolvers. In addition, the approval of a newly minted PURL for use may vary between different resolvers, since the administration is autonomous for each.

Both LSRN and the registry underlying Identifiers.org are free to use and open to submissions by anyone. Shared Names at present has no central repository, though it is envisaged it will be free to use and accessible by anyone.

3.4 Query and export

A queryable list of namespaces already existing within an identification scheme is highly desirable, allowing appropriate community awareness and therefore reuse. Of all the identification schemes described, only the registries of LSRN and Identifiers.org are public-facing, with dedicated web access. LSRN operates an open, searchable registry of namespaces to which anyone can contribute. This is not, however, accessible through programmatic means. The PURL system allows anyone to create a PURL, but these are not available centrally to search by others. DOI also operates a closed repository, which is neither available to view nor to access by programmatic means. The MIRIAM Registry underlying Identifiers.org has an open mechanism allowing public submissions, is free to use, and is queryable and downloadable through computational means, such as web services and XML export. The ability to download or export a list of registered namespaces allows query and information processing in the absence of a web connection, and removes the need for constant web service calls.

3.5 Curation

Once URIs from a particular system come into common use, it is crucial that they retain their ability to uniquely identify the underlying data concept, and in the case of resolvable URIs, to also retain their function to resolve to a correct physical location. These both require a continuous curation process to refine existing, and incorporate new data collections as required, and to update final resolving locations as needed.

While the LSRN repository contains a reasonable number of database entries (around 200), it is inconsistently curated and sporadically maintained. In addition, this scheme is now largely deprecated, with no update recorded for the past half year. In contrast, the MIRIAM Registry contains around 380 data collection entries, associated with around 500 resolving locations. A further 80 collections are under curation, and will be published to the live registry in the near future. The content of the MIRIAM Registry has been continuously updated since its launch. Furthermore, the information stored in the LSRN repository has also been incorporated into Identifiers.org.

In the case of PURLs, since there is no central, queryable repository of existing assigned PURLs, this could result in the creation of a plethora of PURLs, originating from multiple users, all redirecting to the same URL endpoint. In addition, since the update and modification of each PURL is the responsibility of its creator, it is likely that many PURLs would, over time, become vestigial artefacts leading to "dead ends" in the web of data - diametrically opposed to the original intent of the PURL itself, and indeed to the objectives of Linked Data. The fact that Identifiers.org URIs provides all registered physical resolving locations illustrates the point that these URIs could encompass existing PURLs.

Of the identification schemes listed, only the registry underlying Identifiers.org has any documented curation process. In addition, the information stored in the MIRIAM Registry is used to perform a variety of consistency checks: regular expression patterns for data collections identifiers are used to confirm that URIs specifying individual data records are valid, and physical resolving locations are checked daily to ensure their accuracy. For interested parties, these daily health checks can also provide a history of uptime for individual resources, and are accessible to the users through a calendar view.

For the majority of the other schemes, since they are curated or administered to at least some extent, the uniqueness of URIs will depend upon the accuracy of the curation process; the duplication of namespaces through which records can be identified could, for example, result in potential duplication of URIs for certain records. This situation exists currently in the LSRN repository, where for example both *UniProt* and *UniProtKB_Swiss-Prot* database abbreviations can be use to specify the same record (`http://lsrn.org/UniProt:P12345` and `http://lsrn.org/UniProtKB_Swiss-Prot:P12345`). It is difficult to judge the extent of such namespace collisions in the closed or unsearchable repositories of DOI and LSID. Identifiers.org, however, is publicly available, and any such errors can be reported for amelioration.

3.6 Miscellaneous

Identifiers.org provides a means to refine the behaviour of the resolving system itself through the use of a variety of parameters. These can be used to specify resolving locations or to use a 'profile', which can store predefined resolving behaviours for URIs. One such default profile called 'most_reliable' is already available to use, and automatically resolves to the resource with the greatest uptime, as recorded over its lifetime in the MIRIAM Registry. Moreover, malformed queries through Identifiers.org result in human-readable error messages, with appropriate HTTP status codes. For example, when an invalid identifier is provided in a given URI, a '400 Bad Request' message is returned, together with the regular expression recorded against the collection that resulted in that failure; for URIs based on collections that are not recorded in the Registry, a '404 Not Found' message is returned, stating the unknown namespace.

4 Conclusion

With the distributed nature of biological data comes a need to provide a unified and standard means to identify and access it in a consistent and persistent manner. Over time, with the advent of Semantic Web efforts, it has also become highly desirable that any solution integrates well with the web of linked data. Several efforts have arisen over the years, being mainly designed and implemented to address a single specific issue from the wider set of problems. This restricted view, which at the time may have been sufficient to address a specific goal, has now been found wanting.

A preferred identification scheme should meet the basic requirements of unique (one record, one identifier), stable (immutable) and perennial (longevity) identification of data. Additionally, such an identification scheme

Table 1. Comparison of identifier schemes characteristics. The 'Unique identifier' column specifies whether the URI generated by a given scheme can be regarded as a unique identifier (as defined earlier). The 'Resolvable identifier' column refers to whether the URI can resolves to a web page providing information (for example when pasted into a web browser). The 'Standard compliant identifier' column describes whether the presented scheme adheres to a publicly available standard. The 'Resolving cardinality' column describes whether the identifier generated from a given scheme is able to resolve either 1 (1:1) or to many (1:many) physical locations. The 'Cost' column describes whether fees are required to use the service. The 'Registration' column describes the level at which the scheme accepts new submissions. The 'Queryable and/or downloadable' column describes whether or not existing namespaces may be identified, and how the information is made available to users. The 'Curation' column describes whether or not any maintenance is performed on the stored data subsequent to its incorporation into said scheme. In the case of 'Shared Names' much of the information is based upon the documentation stating how the system will behave once implemented. Where insufficient detail is provided, the relevant cells are marked 'unknown' (grey). The colour coding of the cells depicts the ranked desirability of the specified characteristic with reference to a particular identification scheme: most desirable characteristics are coloured green, less desirable amber, and limited desirability, salmon. Further details are provided in the 'Discussion' section.

	Unique identifier	**Resolvable identifier**	**Standard compliant identifier**	**Resolving cardinality**	**Cost**	**Registration**	**Queryable and/or downloadable**	**Curation**
DOI	yes	yes	yes	1:1	yes	fee-based	no	no
LSID	yes	no	yes	1:1	no	institutional	no	no
LSRN	yes	yes	no	1:1	no	individual submissions	web page and RDF/XML export	no
PURL	no	yes	yes	1:1	no	individual submissions	no	no
Shared Names	yes	yes	yes	1:1	no	unknown	unknown	unknown
Identifiers.org	yes	yes	yes	1:many	no	individual and bulk submissions	web page, web services, XML export	yes

should also use resolvable HTTP URIs to facilitate their integration into the Linked Data cloud. Taken together, these requirements translate into a set of desired traits, and can be used to gather qualitative metrics on existing identification schemes

As summarized in Table 1, most identification schemes are at least suboptimal in one characteristic. For example, LSID and PURL provide no centralized or searchable public-facing repository, thereby encouraging the proliferation of identifiers through the creation of new URIs for data entities. Only LSRN and Identifiers.org operate an open repository, which does not require registration, and provides at least some way to view existing and submit new information. Of the schemes discussed, only DOI operates a fee-based model. This would seemingly be at odds with the ideal of open sharing of data, and indeed the costs associated with assignment of DOI names to huge data sets could be prohibitively large.

To date, Identifiers.org is the only identification scheme which addresses all the necessary and desired characteristics stated for perennial identification and location of data entities on the web. It uses a simple and generic design, which is extendable should any additional requirements be raised.

As a standardisation effort, support for the use of resolvable Identifiers.org URIs is growing. LSRN has recently decided that it will be transitioning its information into this scheme, with their existing repository moving into a "maintenance only" mode to support legacy system dependencies.

Identifiers.org can be considered as an interoperable and universal cross-referencing framework for the Life Sciences, which uses resolvable URIs (HTTP URLs), in line with Linked Data requirements, and is 'free' in all senses.

Acknowledgements Identifiers.org and the MIRIAM Registry have been developed with funding from Open PHACTS, the Biotechnology and Biological Sciences Research Council (BBSRC), the European Molecular Biology Laboratory (EMBL) and ELIXIR (Preparatory Phase).

References

1. http://www.w3.org/DesignIssues/LinkedData

2. http://www.w3.org/Provider/Style/URI
3. http://www.w3.org/2001/sw/
4. http://linkeddata.org/
5. Juty, N., Le Novère, N., Laibe, C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. **40** (2012) D580–86
6. http://www.doi.org/
7. http://lsids.sourceforge.net/
8. http://lsrn.org/
9. http://www.purl.org/
10. http://sharedname.org/
11. Laibe, C., Le Novère, N. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. BMC Systems Biol. **1** (2007) 58
12. Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., *et al* Minimum Information Requested In the Annotation of biochemical Models (MIRIAM). Nat. Biotech. **23** (2005) 1509–15
13. http://www.omg.org/cgi-bin/doc?dtc/04-05-01
14. http://www.tdwg.org/activities/online-services/lsid-authority-ids/
15. http://wiki.tdwg.org/twiki/bin/view/GUID/LsidHttpProxyUsageRecommendation
16. Clark, T., Martin, S., Liefeld, T. Object Identification for Biological Knowledgebases. Brief Bioinform. **5** (2004) 59–70
17. Martin, S., Hohman, M., Liefeld, T. The impact of Life Science Identifier on informatics data. Drug Discov. Today **10** (2005) 1566–72
18. http://wiki.tdwg.org/twiki/bin/view/GUID/LSID
19. The Uniprot Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt) Nucleic Acids Res. **40** (2012) D71–75
20. http://www.geneontology.org/cgi-bin/xrefs.cgi
21. http://www.expasy.org/links.html
22. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=43506