Adoption of Persistent Identifiers for Biodiversity Informatics

Recommendations of the GBIF LSID GUID Task Group, 6 November 2009

Phil Cryer (Missouri Botanical Garden),
Roger Hyam (Natural History Museum, London, and PESI),
Chuck Miller (Missouri Botanical Garden),
Nicola Nicolson (Royal Botanic Gardens, Kew),
Éamonn Ó Tuama (GBIF),
Rod Page (University of Glasgow),
Jonathan Rees (Science Commons),
Greg Riccardi (co-chair, Florida State University),
Kevin Richards (Landcare Research, New Zealand),
Richard White (co-chair, Cardiff University)

Summary

Effective identification of data objects is essential for linking the world's biodiversity data. If GBIF is to enable the exchange of biodiversity data it must promote identifier adoption. GBIF can do this in three ways:

Leadership: The GBIF data portal is a focal point in the flow of biodiversity data. The feedback and data cleaning tools provided through the portal influence the quality of data being published by providers. GBIF should place the use and re-use of identifiers as a high priority in assessing the quality of data. GBIF should move to a position where it mandates the use of identifiers and well known vocabularies for all data accepted by the portal.

Education, training and outreach: All users must appreciate the importance of issuing identifiers for their data and re-using identifiers from other people's data. Literature and training courses should be offered to those who need assistance with this.

Practical services: For technical and social reasons many data suppliers are not able to provide reliable resolution of the identifiers they issue. GBIF should provide services to support resolution of these identifiers. It should also support the hosting and maintenance of essential vocabularies.

Version History

Version	Date	Modified by	Comments
v1.1	21/01/2010	Éamonn Ó Tuama (eotuama@gbif.org)	added references to GeoSpecies and PubMed in Section 3.3; added GeoSpecies to diagram.

List of recommendations

This is a summary of the full recommendations which appear in boxes later in the document in sections 5 to 7.

GBIF should:

1: take the lead in driving the application and use of identifiers in biodiversity informatics,

2: provide materials such as an executive summary targeted to administrative leaders explaining the costs and benefits of implementing persistent identifiers,

3: educate the community in general persistent identifier principles and practices,

4: encourage, support and advise on the use of appropriate identifier technologies, in particular LSIDs and HTTP URIs, but not impose a requirement for one at the expense of the other, and provide specific advice for the issuing and use of LSIDs and for HTTP URIs,

5: support a promotional programme,

6: demonstrate good practice in its data portal,

7: assist providers that are not currently maintaining their own persistent identifiers to do so: this includes both education and technology,

8: make data more inter-connected,

9: start a programme to become an RDF consumer and encourage data providers to deploy RDF services,

10: provide services to support identifier resolution, redirection, metadata hosting, and caching,

11: provide additional services, including persistent identifier monitoring services,

12: encourage the use of metadata vocabularies and extend the role of its data portal by hosting resources related to the use of identifiers, such as the TDWG vocabularies,

13: assist with the availability of software for data and service providers, and

14: continue to be funded to provide support to data providers for the foreseeable future.

Contents

1 Introduction

GBIF has identified the provision of identifiers for biodiversity objects as one of the central challenges to developing a global bioinformatics infrastructure. One of the stated goals in the GBIF strategic plans document "*GBIF Plans 2007 – 2011 from prototype towards full operation*" (http://www2.gbif.org/strategic_plans.pdf) is to consolidate the underlying enabling infrastructure and standardisation for global connectivity of biodiversity data and information through an activity to "develop a system of globally unique identifiers and encourage their use throughout biodiversity informatics". The GBIF plans envisage using TDWG standards to "allow all data objects to be identified using standard actionable globally unique identifiers" and provision of a GBIF web service and user interface to allow users "to locate and view any data object with a standard globally unique identifier".

GBIF convened a task group, the "LSID GUID Task Group" (LGTG) to explore the issues and offer recommendations on the way forward, with particular reference to the GBIF network, that will enable GBIF to provide architecture leadership and best practices for implementation. The principal objective of the group is to provide recommendations and guidelines on deployment of identifiers on the GBIF network with particular reference to the potential role of GBIF as a stable, long term provider of identifier resolution services. This document is the report of the group.

2 The characteristics of effective identifiers

For our purposes, an identifier is a character string associated with an object. "GBIF" and "http://wiki.gbif.org/guidwiki/" are examples of identifiers. Identifiers are used in informatics to refer to objects in data sets, documents and repositories.

There are two over-arching use cases that make identifiers effective for users:

- **Uniqueness of reference:** An identifier can be used to aggregate information about the identified object. For example, information received from multiple sources associated with a single identifier is assumed to be information about a single object.
- **Action:** An identifier can be used to find further information about the object, concept or data to which it refers. This information might be interpreted directly or used to support services.

Effective identifiers will make a vital contribution to facilitating the use of biodiversity data by software agents, so that data can be used by and become embedded in an unlimited number of future information systems, as the world moves towards Web 2.0, the Semantic Web, Linked Data and the e-Science Grid.

2.1 Persistent actionable identifiers

Identifiers should be *persistent* and *actionable* in order to be effective tools in managing and integrating information.

- **Persistent:** An identifier is persistent if it always refers to a specific object. All information associated with a persistent identifier is about the same object. The properties of the object are subject to change, but once a persistent identifier is assigned to one object, it *cannot* be reused to refer to a different object.

For example, the ITIS (Integrated Taxonomic Information System, http://www.itis.gov/)

TSNs (Taxonomic Serial Numbers) are integers that are persistent identifiers for taxa. Once ITIS assigns a TSN to a taxon, that TSN will never be used for a different taxon.

- **Actionable:** An identifier is actionable if there is a service that, given the identifier, provides information about the object identified (for example, a resolution service).

Actionable identifiers should contain information which locates an appropriate resolution service if presented to a suitable client.

For example, an HTTP URI is actionable. It necessarily begins with "http://" and thus is recognisable by its structure. The HTTP system provides mechanisms for clients to access a data object from its associated identifier. ITIS TSNs, which are simple integers, are potentially actionable because ITIS supports services that provide information for TSNs.

The two identifier systems described below (HTTP URI and LSID) represent different strategies to provide actionable identifiers.

One important type of action is *resolution*, the process in which an identifier is presented to a network service to receive in return a specific output of one or more pieces of current information related to the identifier or its related object or both. For example, the Domain Name System (DNS) resolves domain names meaningful to humans into numerical IP addresses.

GBIF does not currently support persistent actionable identifiers for objects in the data portal. The identifiers attached by GBIF to their occurrence records are based on the *Darwin Core triplet*: the three fields of institution id, collection id and catalogue number provided in Darwin Core records. These identifiers are intended to be unique within the GBIF data cache at a particular time. However, although it is a recommended best practice, not every data provider ensures consistency of identification and thus a Darwin Core triplet may represent different objects at different times, e.g. through reassignment of catalogue numbers when re-indexing a database. Hence not all Darwin Core triplet identifiers are guaranteed to be persistent.

2.2 Identifier terminology

The biodiversity informatics community has been using "globally unique identifier" (GUID) as a generic term for persistent, resolvable identifiers (hence the name of the LSID GUID Task Group). However, outside biodiversity informatics the term "GUID" is most often a synonym of "universally unique identifier" (UUID). To avoid confusion, we have adopted the term "persistent identifier", which is widely used in discussions of unique identifiers in the digital library and publishing communities. For the remainder of the document, the term "identifier" will generally refer to a persistent, actionable identifier. The qualifiers "persistent" and "actionable" are added for emphasis or to refer to an identifier system that must have one but not necessarily both properties.

3 Some benefits of persistent identifiers

A decentralised, or autonomous, informatics architecture is one in which resolution, search and discovery tools interact with distributed providers, and in which each interacting facility is both consumer and producer. Of particular interest are feedback mechanisms in which providers of information receive comments, or annotations, about that information.

GBIF information providers have long wanted a mechanism for consumers to report the usage of the information and to give feedback on data quality. Usage reports would provide

evidence for the impact of providing data items to the wider community. Data quality feedback from consumers would allow for correction and enhancement of data by providers.

The GBIF informatics architecture is currently based on a provider/consumer model in which information flows primarily from provider to consumer. Feedback from consumers about quality of information and about usage of information flows back to providers outside of the information architecture, typically by email.

Support for identifiers is crucial to decentralised architecture in order to allow replication of information (one server keeps an exact copy of another server's object), annotation of information (one server records an assertion about another server's object), and reporting results of searches as collections of identifiers.

This section describes some opportunities for enhancement of the GBIF information services to provide specific feedback mechanisms that are of interest to the biodiversity informatics community.

3.1 Tracking citation and impact

Tracking the usage of identifiers is a special case of creating, managing and distributing associations among digital objects. For example, a collection of occurrence records is used as input to a data analysis activity and presented in a publication. The person, or people, who found the occurrence records, performed the data analysis, and wrote the publication are also represented by digital objects. Each of those objects has an identifier.

The association among these objects might be contained in a blog post:

> *Joe writes "I searched the GBIF repository for all frogs from Cuba. The collection of objects that I found useful is in the collection [ID1]. I plotted the locations of the records [ID2] and reported the results in my paper [ID3]."*

The blog post can be scanned by a search engine and incorporated into rankings and ratings of the associated objects. The blog post has an identifier (HTTP URI) of its own and is associated with the writer (Joe).

In general, associations like the blog post are identified, stored in repositories, scanned by search engines and other aggregators, and enhance the usefulness of the associated objects.

3.2 Management and disambiguation of taxon names

Disambiguation of taxon names requires services that support tests of difference as well as of equality. A persistent identifier always refers to a specific object but different identifiers do not necessarily refer to different objects. A single object may have many identifiers. Tests of inequality for objects must rely on evaluation of metadata or of the objects themselves.

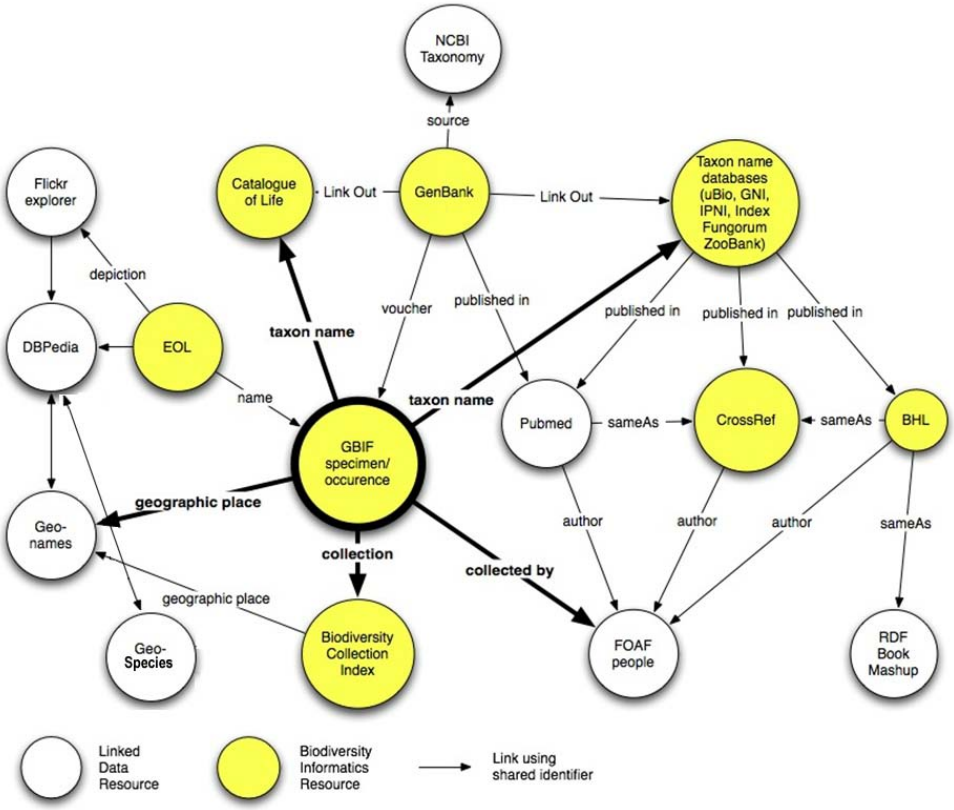
The ambiguities inherent in taxon name usage are widely recognised. For example, biodiversity researchers ask whether two specimens with different name strings are believed to be of the same taxonomic group. The availability of identifiers for name strings, published names and taxon concepts allows the recording of assertions that will help in answering this question.

3.3 Integrating identifiers with the Semantic Web and the Linked Data model

Linked Data is a vision of a web of interconnected data, to be consumed by machines. Typically, HTTP URIs are used as identifiers, and the data is described using RDF. Just as a web page contains links to other web pages, linked data sets contain links to other, related

data. The Linked Data home page (http://linkeddata.org/) displays a graph of the many resources that are being linked together. Many of the resources are clearly relevant to biodiversity, including DBPedia (http://dbpedia.org/), an RDF export of Wikipedia, GeoNames (http://geonames.org/), a resource for geographic places, GeoSpecies, a site helping tie together disparate data about species (http://lod.geospecies.org/), UniProt, a repository of genomics data (http://www.uniprot.org/), PubMed, a citation repository for biomedical and life science publications (http://www.ncbi.nlm.nih.gov/pubmed/), and the World Factbook (https://www.cia.gov/library/publications/the-world-factbook/).

The diagram below illustrates some of the potential linkages between biodiversity resources and the broader linked data cloud that would be enabled if biodiversity data were published following Linked Data recommendations (http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/). White circles represent already existing linked data resources, yellow circles represent biodiversity resources (such as GBIF, nomenclators, EOL, etc.) and related sources such as CrossRef.



The **bold links** between GBIF and other resources represent elements of a GBIF specimen record (for example) that could be represented by an identifier in an external database. For example, a plant specimen record could contain the identifier of the plant name in IPNI, the collection identifier from the Biodiversity Collections Index (http://biocol.org/), a geographic place identifier from GeoNames, and an identifier for the collector. Linking together biodiversity data (yellow circles) enables more sophisticated biodiversity queries, such as "where in the world are most new species being described?", which requires specimens linked to names linked to publication dates. It also facilitates data citation and data cleaning. As an example, see "Biodiversity informatics: the challenge of linking data and the role of shared identifiers" (http://dx.doi.org/10.1093/bib/bbn022) whose identifier is "doi:10.1093/bib/bbn022". But the real power comes from linking biodiversity to other data, for example population, economic, climatological, etc. Following Linked Data recommendations offers additional benefits of economies of scale, including making use of

already existing guidelines, tutorials, and tools such as the linked data validator (http://validator.linkeddata.org/).

3.4 Linked data requirements

The guide "How to Publish Linked Data on the Web" (http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/) states that "Information has to fulfil the following minimal requirements to be considered 'published as Linked Data on the Web':

1. "Things must be identified with actionable HTTP URIs.
2. "If such a URI is dereferenced asking for the MIME-type application/rdf+xml, a data source must return an RDF/XML description of the identified resource.
3. "URIs that identify non-information resources must be set up in one of these ways: [303 redirects or fragment identifiers]
4. "Besides RDF links to resources within the same data source, RDF descriptions should also contain RDF links to resources provided by other data sources, so that clients can navigate the Web of Data as a whole by following RDF links."

In essence, HTTP URIs are the identifiers (1), RDF describes the data (2), and the RDF should have links to other data (4). Requirement (3) is discussed further in Appendix 1. This has implications for biodiversity informatics, in that our identifiers should be capable of being represented as HTTP URIs and our metadata as RDF.

4 Review of identifier technologies for biodiversity informatics

A variety of different types of persistent identifiers have been reviewed by Garrity *et al.* (2009). Below we describe the two persistent identifier technologies the Task Group recommends GBIF should support: HTTP URIs and LSIDs. Other identifier mechanisms such as DOIs are in use in other data domains, and are described in Appendix 3, since they may occur in data to which biodiversity data might be linked.

4.1 HTTP URIs

A Uniform Resource Identifier (URI) consists of a string of characters used to identify or name a resource on the Internet. A URI scheme defines a specific syntax and associated protocols for a collection of URIs.

HTTP URI is a URI scheme whose identifiers are prefixed with "http://". An HTTP URI can be used to locate network resources via the HTTP protocol. An example of a biodiversity data HTTP URI is http://ci.nii.ac.jp/naid/110004021938#article.

4.2 Life Science Identifiers (LSIDs)

An LSID is a particular kind of actionable identifier which is recommended for use by TDWG and which

- enables global uniqueness by including an Internet domain name, which is itself subject to rules and procedures ensuring uniqueness, and
- uses the domain name system to locate a resolution service which enables a user to find out more about the entity to which an LSID refers.

An LSID provides a means to identify and locate a piece of biological data and/or metadata

on the web. An example of an LSID is urn:lsid:ipni.org:names:20012728-1:1.1. For a more detailed description see the LSID Resolution Project Homepage (http://lsids.sourceforge.net).

By themselves LSIDs do not meet the requirements of Linked Data because they are not HTTP URIs. Standard Linked Data clients will not be able to handle them. One solution to this problem is to represent LSIDs as HTTP URIs.

For example, the bioguid.info Web site provides LSID resolution proxy services. Appending the LSID "urn:lsid:ipni.org:names:20012728-1:1.1" to "http://bioguid.info/" yields the HTTP URI http://bioguid.info/urn:lsid:ipni.org:names:20012728-1:1.1. That URI, when presented to a Web browser, produces an HTML document containing the metadata of the referenced name object.

Additionally the bioguid proxy conforms to the Linked Data requirements (as in Section 3.4) by supporting content negotiation (requirement 2) and 303 redirects (requirement 3).

5 Role of GBIF in leadership and education

The GBIF organisation holds a unique role of bringing together various global efforts that aim to produce useful and usable biodiversity information resources. To ensure these efforts work towards a common goal, and that the integration of related data and services is always achievable, it is essential for GBIF to provide assistance to the interested parties of these efforts. The interested parties can range from large institutions with masses of biodiversity data, to small non-funded organisations that wish to have their data made available globally. Clearly, this assistance should include identifier hosting and provisioning services.

Several other goals in the GBIF Work Programme 2009-2010 depend directly or indirectly on the deployment of identifiers.

Recommendation 1: GBIF should take the leadership role in driving the application and use of identifiers in biodiversity informatics.

5.1 Institutional support

An important aspect of persistent identifier implementation at an institution is marshalling the support of the administrative leadership. The leaders of an institution must agree to commit the resources needed both to implement and to sustain a reliable identifier system. Obtaining this agreement will involve some form of persuasive request or presentation that clarifies the benefits and the resources required. Among the benefits to the institution would be improved branding and clearer ownership of the institution's intellectual property through more definitive attribution and citation; also identifiers enable clearer metrics for how the institution's information is being used and therefore how its mission is being accomplished.

Recommendation 2: GBIF should provide materials such as an executive summary targeted to administrative leadership explaining the costs and benefits of implementing persistent identifiers.

5.2 Providing education, training and outreach

5.2.1 Advice to users and providers on persistent identifier principles and metadata

Users need to be informed about the need to adopt good practices in handling persistent identifiers.

Recommendation 3: GBIF should educate the community in general persistent identifier principles and practices, such as:

- the unsuitability of local database identifiers, when to change the identifier if the data changes, not to re-use identifiers, not to embed semantics in the identifier, policies for caching and re-caching, provisions to allow data sets and their identifiers to be transferred to a new custodian (and potentially re-branded), etc.,
- expressing metadata in RDF with preferred vocabularies, and containing other applicable identifiers for the same object,
- citing the correct identifiers, for example when there is a chain of derived objects from the original source and from aggregators, and in the use of taxon concept identifiers where possible instead of just taxon name identifiers.

5.2.2 Advice and support for particular kinds of identifiers

Users (including providers and aggregators) need advice to help them choose, issue and use particular kinds of identifiers, on how providers should present their identifiers, and on what support GBIF can provide:

Recommendation 4: GBIF should encourage, support and advise on the use of appropriate identifier technologies, in particular LSIDs and HTTP URIs, but not impose a requirement for one at the expense of the other. GBIF should provide specific advice for the issuing and use of LSIDs and for HTTP URIs, including points such as:

- identifiers such as LSIDs should include "same as" links in the RDF metadata to HTTP URIs to provide a proxied version as an alternative resolution mechanism (e.g. for Semantic Web clients) for its own identifiers.
- LSIDs should also adopt the Linked Data HTTP URI conventions.

Linked Data conventions allow LSIDs and DOIs to work well with Linked Data. See section 3.4 and Appendix 1 for more information about Linked Data requirements.

Education and support need to be targeted for three types of data providers, described further in Appendix 2, which:

- submit data to GBIF, with no permanent online presence,
- use online wrapper software (such as IPT) with identifiers but provide no resolution or guaranteed reliability,
- have a full online presence (such as IPT) and persistent resolvable identifier support.

5.3 Stimulating growth of the community

Recommendation 5: GBIF should support a promotional programme, including:

- workshops for data providers on awareness of identifiers and choosing and implementing persistent identifiers;
- technical and deployment training programmes;
- maintaining a system of "quality marks" for compliant collaborators (data providers, aggregators, etc.)

The following recommendations will help GBIF to demonstrate good practice and lead the biodiversity informatics community forwards.

Recommendation 6: the GBIF data portal should demonstrate good practice for persistent identifiers by:

- maintaining fields for identifiers including those from data providers,

- assigning GBIF identifiers to cached objects,
- property values in GBIF records should be persistent resolvable identifiers if possible.

Recommendation 7: GBIF should assist providers that are not currently maintaining their own persistent identifiers to do so: this includes both education and technology.

Due to the fact that the recommended response type for resolvable identifiers is RDF, it is important for GBIF to support the use of RDF documents and semantic technologies. This will include helping data providers to identify which vocabularies are applicable for their response when resolving an identifier, consuming RDF from providers and possibly producing RDF documents as output.

Recommendation 8: GBIF should make data more inter-connected by:
- adopting current best practice for interconnected data (Linked Data principles),
- outputing RDF documents,
- using existing vocabularies and identifiers wherever possible (see also Recommendation 12).

Recommendation 9: GBIF should start a programme to become an RDF consumer and encourage data providers to deploy RDF services by:
- allowing data providers to upload RDF as an alternative to current formats,
- promoting the use of resolver services and interconnected data.

6 Role of GBIF in technical support

There can be obstacles to the technical implementation of identifiers such as insufficient IT skills available to do the work, organisational barriers to the network, or server changes needed. To the extent possible, the process to implement identifiers should be simplified to reduce the barrier to adoption. Since IT environments vary between institutions there should be alternative methods for implementation of identifiers for the more commonly occurring situations, for example Linux and Windows. Packaged installations and documented approaches would help lower the technical hurdles.

6.1 Role of GBIF as a persistent identifier service provider

Recommendation 10: GBIF should provide services to support identifier resolution, redirection, metadata hosting, and caching.

Although some data providers can provide resolution for their identifiers, this is frequently not possible for many. Also, many providers lack the IT resources to ensure high availability (i.e. up-time) for their data and metadata.

The simple model where a user seeking biodiversity data just goes to the original data provider for resolution

- does not provide for data providers which are not yet online, or have gone offline, and
- may not ensure reliability (sufficient up-time).

These frequently occurring problems can be mitigated by establishing one or more highly available, high capacity identifier service providers. If a data provider has no resolver it can publish all its data through one or more service providers. Users can go to the original data supplier's resolver, if it has one, or (if that fails) to a service made available by a global

service provider, such as GBIF.

Three such services have been identified:

- **Redirection:** identifiers resolve to the service provider, which just redirects the user to the data provider for the metadata. PURLs are an example of redirection, but it works just as well with LSIDs where the service provider responds to the user with WSDL files but the final location of the metadata (as indicated in the WSDL service file) is with the data provider.
- **Metadata Hosting:** identifier resolution is to the service provider, who holds a copy of the metadata previously received from the data provider. No call is made to the data provider during resolution of the identifier, so they do not need a reliable web presence.
- **FallBack Cache:** identifier resolution is to the **data provider** initially, but if resolution fails the user can call the service provider as a fall back option. The service provider will then supply a cached copy of the metadata along with metadata specifying when they last received it from the data provider.

Suggestions for how these services would operate are given in Appendix 2. It may be that some providers do not want their data cached. This could be achieved by "do not cache" HTTP settings, or by annotation properties that specify this is a cached version, and the consumer therefore needs to go to the original to get the non-cached version.

6.1.1 LSID resolution services

The main technical requirement for an organisation that wishes to issue LSIDs for its data is owning a stable internet domain (for example, ipni.org) which acts as an LSID authority. Using special LSID server software, an LSID authority, e.g. IPNI, designates one or more namespaces for its resources (collections, databases) and issues unique identifiers (within the context of its databases) for individual records. Because a URN is not like a normal URI, it cannot be resolved directly in a web browser. Instead special client software which can be built into applications that make use of LSIDs recognises an LSID when encountered and sends a query to an internet domain name system (DNS) to obtain the network location of the associated LSID authority. An LSID authority thus not only requires a Domain Name Service (DNS) to be registered, but also a Service Record (SRV). An SRV record provides information on available services and, e.g., in the case of the example LSID above, associates ipni.org with a service (a component of the LSID server software) at http://lsid.ipni.org which then returns a Web Service Description (WSDL) on the protocol for submitting the LSID for resolution. In completing the process, the service returns metadata in response to the submitted LSID resolution request.

Several GBIF participants have expressed a commitment in moving ahead with deployment of LSIDs and are looking to the GBIF Secretariat to provide leadership and essential services. It would be therefore be suitable for GBIF to take the role as an LSID hosting/proxy service to reduce the technical threshold of LSID authoring and LSID resolution for GBIF participants. This would help to shield the participants from the necessity of having to deal with SRV records which, while not technically challenging, does require access to a DNS server.

It would be advantageous for GBIF to index and cache identifiers as an alternative point of resolution for biodiversity data, especially for those identifier technologies such as LSIDs that are not resolvable by default over HTTP.

6.1.2 HTTP URI resolution

As with LSIDs, HTTP URI identifiers require best practices and a degree of infrastructural

support. To help with their adoption, the following best practices should be encouraged:

- multiple DNS A records,
- institutional agreement to persistence of URIs.

6.1.3 Other services

In addition to resolution, there are opportunities for services to provide increased functionality including tracking provenance, usage (as in BitLink) and uniqueness, and testing whether the data associated with an identifier has changed.

Recommendation 11: GBIF should provide additional services, including persistent identifier monitoring services.

An essential component to any reliable web service is a monitoring system to ensure that use of that service is always available. In this case, a monitoring service would ensure that any GBIF-hosted identifier service is running, along with any other registered identifier services.

A useful example of this would be where GBIF is hosting identifiers for a set of provider URLs (i.e. the identifiers resolve to the resource at the associated URL). A regular check that the data at these URLs is available would improve the quality of service. For example, DOI "monitoring" services which on detecting a broken DOI present a web form to report the issue, which is then dealt with by DOI support staff.

6.2 Availability of vocabularies and software resources

6.2.1 Vocabularies for metadata returned on identifier resolution

Identifier resolution must always result in metadata. This metadata must be represented as RDF serialized as XML (RDF/XML). The metadata must type the object using a well-known vocabulary such as the TDWG ontology or Dublin Core. Entirely bespoke ontologies should not be used but existing ontologies should be extended where necessary. Machine and human clients that retrieve the metadata associated with an identifier will use the associated typing information to decide how to process the metadata and any associated data. If the type information is novel, processing may be difficult or impossible. Use of well known vocabularies allows the development and integration of applications that exploit the known types.

Recommendation 12: GBIF should take a leadership role in encouraging use of metadata vocabularies for information in the GBIF data portal and extending the role of the data portal by hosting resources related to the use of identifiers, such as the TDWG vocabularies. GBIF should:

- continue to participate with TDWG (and other bodies for non-biodiversity data) in standardising these vocabularies,
- consider organising a "hackathon" (programmers workshop) to merge the new DarwinCore vocabulary with the existing TDWG ontology and the current GBIF vocabularies work. (This workshop should present a model for on-going maintenance and development and should be at a technical level and not consider domain specific issues)
- work with others to identify on-going support mechanisms for essential shared vocabularies.

6.2.2 Software resources

Recommendation 13: GBIF should assist with the availability of software for data and service providers by:

- providing easy-to-deploy server packages for several platforms, to make it easier for institutions to set up their own server nodes as LSID resolvers, HTTP URI proxy services, caches, etc.,
- offering funding to encourage application development and service deployment, such as server deployment packages, semantically aware clients, and validators for RDF, HTTP URIs and vocabularies, etc.

7 A business model for adopting identifier technologies

The costs involved in providing highly available and long-lived global identifier services include the following

- Software design and development (scripting) for redirect and caching services
- Server hardware purchase, setup, housing, maintenance
- Bandwidth
- DNS setup and configuration maintenance
- Curation: fixing broken redirects, populating and tracking replicates
- Help desk
- Outreach – educating the community in how to create and use identifiers

Parties who might be interested in funding these services include data providers, data users, and organisations with a general interest in promoting the activities of both. Each of these parties is a candidate for providing the resources required to run the services, and one finds existing identifier systems using each of these three business models. Some examples:

1. The web itself provides unreliable and non-persistent resolution with costs taken on by data providers, with services often outsourced to various kinds of service providers.
2. The Handle system (including DOIs) has costs assumed by data providers, with some services provided by organisations such as CrossRef. CrossRef itself has a complex pricing model involving membership fees and per-identifier charges; the Handle system has a more limited level of service and lower costs.
3. Digital repositories such as GenBank do not charge the original data provider, but rather take on the cost of provisioning as part of their duty to their community. OCLC's purl.org service has a similar approach.
4. Some systems charge end users for access to resolution services.

Complete reliance on data providers is not a robust solution, because a data provider failure or the withdrawal of a provider from the system leaves users of identifiers high and dry. Reliance on user fees is not really an option in the context of scientific research. This implies that the best business model for a reliable resolution framework shares responsibility between data providers and service providers (such as GBIF) that represent the community of users. Sometimes a data provider will be willing to take on some or all of the costs and commitments; when it is not, the community still needs to be served, and it is best if a service provider picks up the pieces.

In either case, the resources must come from somewhere. Some possible sources might include:

- Informal sources: Some individual in an organisation unilaterally takes on the job of setting up stable and reliable resolution. This may often be feasible and has low overhead, but then resolution is at risk if this individual leaves or gets busy with other things.
- Grants: It may often be possible to obtain funding to set up a resolution system by applying for a grant. Of course the project may be at risk when the grant runs out, so a different strategy has to be used for maintenance.
- Cost of doing business: If an organisation (either data provider or service provider) can be convinced that identifier resolution is in its interest, or is its responsibility, then costs can be written into budgets and responsibilities can be institutionalised in job descriptions.

As a general principle, data providers should do as much as they can manage, but their ability to provide long-term support may be limited, and GBIF should be able to offer support where it is needed to ensure the continued availability of data and services on which scientific research and other public services depend.

Recommendation 14: GBIF should continue to be funded to provide support to data providers for the foreseeable future:

- the biodiversity informatics community needs indispensible support services in order to grow, flourish and provide the answers which society demands,
- the only business model which would work to support these services is the one in which GBIF takes on a significant portion of the provisioning.

Glossary

Actionable persistent identifier	A persistent identifier that can be used (resolved) to obtain metadata about the related object.
DOI	Digital Object Identifier (http://en.wikipedia.org/wiki/Digital_object_identifier)
GUID	Globally Unique Identifier. GUID is often used as a synonym for UUID, as in http://en.wikipedia.org/wiki/Globally_Unique_Identifier
Handle system	The Handle System is a technology specification for managing persistent identifiers for internet resources (http://www.handle.net/; http://en.wikipedia.org/wiki/Handle_System)
HTTP URI	Hypertext Transfer Protocol Uniform Resource Identifier, a URI (q.v.) which uses the HTTP protocol (http://www.rfc-editor.org/rfc/rfc2616.txt)
LSID	Life Sciences Identifier (http://en.wikipedia.org/wiki/LSID), a type of actionable persistent identifier that has been adopted by members of the biodiversity community.
Persistent identifier	An identifier with a unique and stable relationship with an object. A persistent identifier refers to a single object during its lifetime and is never reused as a reference to a different object.

Proxy	An intermediate service which seeks a resource on behalf of a client.
PURL	Persistent Uniform Resource Locator (http://en.wikipedia.org/wiki/Persistent_Uniform_Resource_Locator)
Resolution	The ability and mechanism to obtain the metadata about a specific persistent identifier.
RDF	Resource Description Framework (http://en.wikipedia.org/wiki/Resource_Description_Framework)
RDF triple	A three part piece of data, subject, predicate, object. The subject is the object (conceptual or physical) that the data is about, and must be a persistent identifier. The predicate is the type of data (property, e.g., hasAuthor). The object is the value of the data, which can either be a literal value or another object.
URI	Uniform Resource Identifier (http://www.rfc-editor.org/rfc/rfc3986.txt, http://en.wikipedia.org/wiki/Uniform_Resource_Identifier), consists of a string of characters used to identify or name a resource on the Internet. URLs (web addresses) are an example, as are URNs such as LSIDs.
URL	Uniform Resource Locator (http://en.wikipedia.org/wiki/Uniform_Resource_Locator), a web address. It specifies where to find a resource (e.g., www.google.com) and how to retrieve it (e.g., use the HTTP protocol), hence http://www.google.com
URN	Uniform Resource Name (http://en.wikipedia.org/wiki/Uniform_Resource_Name), a name of a resource. Intended to be a persistent, location-independent identifier. A URN is a kind of URI. Note that a URN is a name, and there is no implication that the resource is digitally available.
UUID	Universally Unique Identifier (http://en.wikipedia.org/wiki/Universally_Unique_Identifier)

References

Berners-Lee, T., “What do HTTP URIs Identify?” http://www.w3.org/DesignIssues/HTTP-URI.html

Berners-Lee, T., “What HTTP URIs Identify” http://www.w3.org/DesignIssues/HTTP-URI2

G.M. Garrity, L.M. Thompson, D.W. Ussery, N. Paskin, D. Baker, P. Desmeth, D.E. Schindel and P.S. Ong, Study on the Identification, Tracking and Monitoring of Genetic Resources, Convention on Biological Diversity, 2 March 2009 (http://www.cbd.int/doc/meetings/abs/abswg-07/information/abswg-07-inf-02-en.pdf).

Appendices

Appendix 1: Implementation and use of HTTP URIs in Linked Data

A HTTP URI can identify either an information resource (such as a web page) or a non-information resource (such as a person, a place, or a concept). For example, if we want to say that the image shown left (from the GBIF web site) depicts the offices of the GBIF secretariat, how do we say this using HTTP URIs as identifiers? In practice, we make this distinction all the time (nobody confuses GBIF the organisation and http://www.gbif.org the web site), but it needs to be made explicit for computers.

Given an HTTP URI for the image, and an HTTP URI for GBIF, we could write:

http://www.gbif.org/uploads/pics/4797_small.jpg *depicts* http://www.gbif.org

But this is ambiguous, because it could be interpreted to mean that the image depicts GBIF's home page (http://www.gbif.org), instead of what we intend, which is that it depicts offices of the secretariat (a building, not a web page).

One solution to this issue is the HTTP 303 redirect. A client sends an HTTP request to a server. If the URI is an information resource (such as an image or a document) the server responds by returning the HTTP 200 response, and the corresponding document. If the URI identifies a non-information resource (such as a person or building), the server returns an HTTP 303 response, along with a URI to where a document describing that non-information resource can be found. In other words, the server is saying "here is where you can get some information about the thing identified by that URI". The nature of the document returned can be specified using HTTP content negotiation. A web browser will ask for HTML, a linked data client will request RDF.

Another solution to distinguishing between a document and the thing described by the document is to use a fragment identifier. For example, the HTTP URI `http://ci.nii.ac.jp/naid/110004021938#article` identifies the article:

T. Okada (1990). New Taxonomic Changes in the Family Drosophilidae (Diptera). *The Entomological Society of Japan*, **58**: 154

A web browser resolving this URI will strip off the fragment identifier (#article) and retrieve the document at `http://ci.nii.ac.jp/naid/110004021938` (which is an HTML page). This HTML document contains a link to the metadata in RDF (`http://ci.nii.ac.jp/naid/110004021938/rdf`), which can be consumed by a linked data client. The RDF contains the statement

```
<rdf:Description rdf:about="http://ci.nii.ac.jp/naid/110004021938#article">
```

which asserts that the RDF document is about the article. In this way the URI of the resource and the description of that resource are kept distinct.

Appendix 2: Services to support data providers and resolution

Three data provider types were noted in section 5.2.2 and three types of service were described in section 6.1. The relationships between data provider and service types can be summarised in a table. (Note that all types of data providers need to have an appreciation of persistent actionable identifiers, even if they don't provide resolution services, because they need to maintain persistent identifiers in their data internally.)

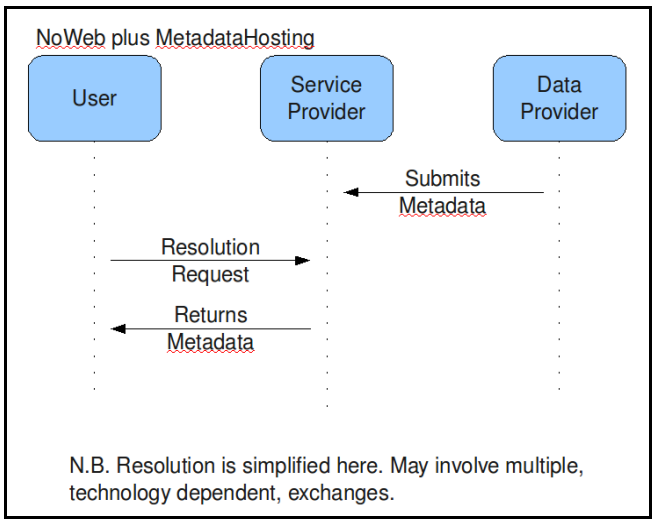
Provider type	**Redirect**	**Metadata Hosting**	**FallBack Cache**
NoWeb: the data provider has no web presence (it can't or doesn't want to host its data on the web)	No – a service provider can't redirect if no web presence	Yes – the only option if metadata isn't available from the data provider	n/a – if provider can't provide resolution then it won't be available for fall-back
SomeWeb: the data provider has the ability to place data on the web but can't guarantee stability of web location (domain permanence) and high availability	Yes – a service provider gives stability to DNS or other part of resolution mechanism	Yes – it may be a choice of the data provider not to be redirected to.	Yes – if the data provider is down for metadata, a service provider could give a cached copy.
WebSavvy: the provider can host its own data and provide its own identifier resolution services including setting up DNS records	No – ID resolution is to the data provider	No – ID resolution is to the data provider	Yes – if the data provider is down and a service provider has a cached version

The matching of Data Providers to appropriate Resolution Service types can also be described by means of an informal decision tree, as follows:

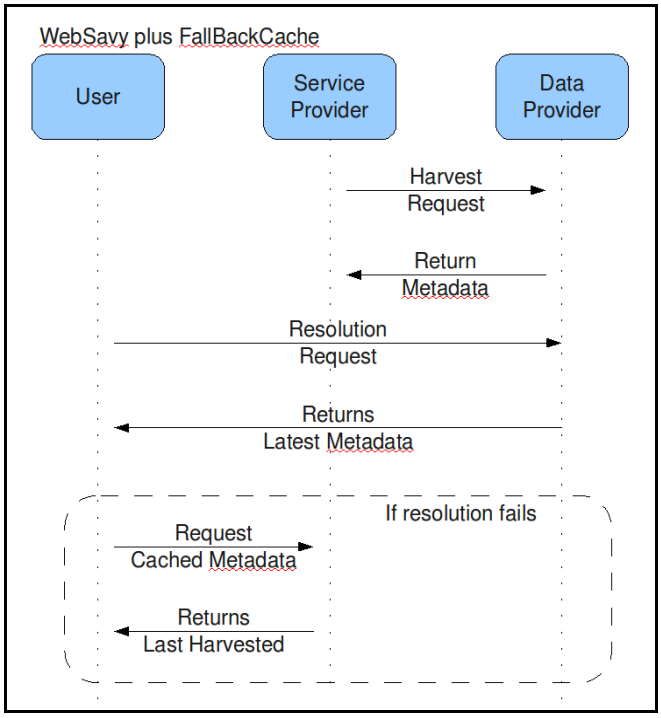
- No internally unique IDs
  - ⇒ Go to data management workshop and come back later
- Managed internally unique IDs that can be mapped to externally unique IDs
  - No web presence for hosting data or no desire to set it up and manage it (**NoWeb**)
    - ⇒ Submit data to a service provider on a regular basis; the service provider will handle all ID resolution: **Metadata Hosting**
  - Have web presence or willing to set it up and manage it
    - Web presence but not willing to maintain high availability at a stable domain location (**SomeWeb**)
      - ⇒ Data harvested by a service provider who then provides all ID services: **Metadata Hosting**
    - Reliable web hosting of data
      - Not able or willing to commit to long term maintenance of DNS; unable to alter DNS entry of LSID or perhaps create subdomain of corporate domain
        - ⇒ A service provider supplies **Redirect** service for IDs but also harvests metadata so it can provide a **FallBack Cache** if the data provider goes down
      - Able to provide long term maintenance of DNS entries and handle full resolution of IDs
        - ⇒ A service provider still harvests metadata so as to provide **FallBack Cache**

The flow of requests and data between a user, a service provider, and the original data provider, can be described using the following diagrams:

Sequence Diagram for 'NoWeb' data provider



Sequence Diagram for other provider types



Sequence diagram for interaction with LSID proxy with optional caching / endpoint monitoring

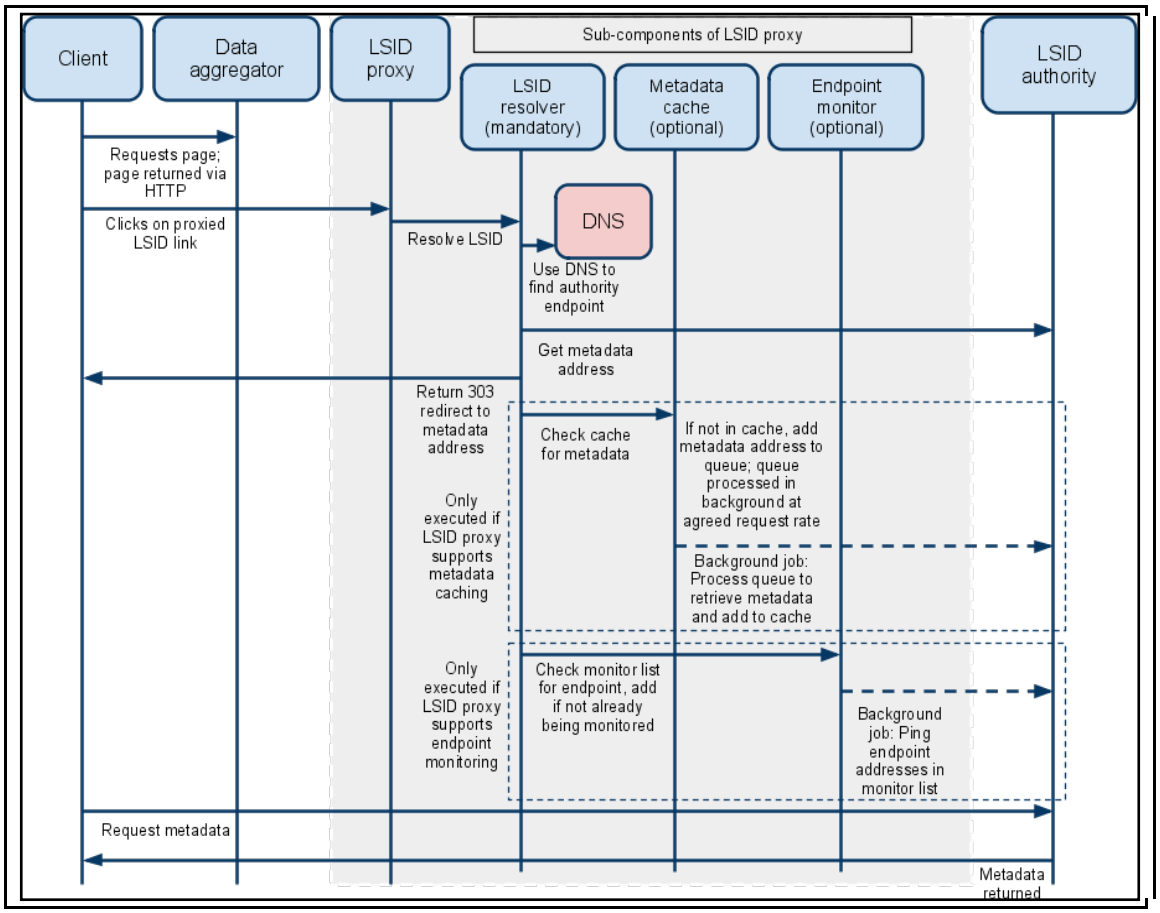
Client: An end user interacting with resources using a web browser (or some other HTTP aware tool).

Data aggregator: A site such as GBIF which displays data from multiple sources labelled with LSIDs. These LSIDs are shown as clickable links using HTTP with the URL of the link specifying the proxy e.g. http://lsidproxy.tdwg.org/[lsid]

LSID proxy: A service which resides at an address such as lsidproxy.tdwg.org accepts LSIDs using the format: http://lsidproxy.tdwg.org/[lsid] (where [lsid] is in the form e.g.: urn:lsid:ipni.org:names:12345-1) and resolves to the address of the metadata location at the authority endpoint. Shown split into sub-components:

- LSID resolver (mandatory)
- Metadata cache (optional)
- Endpoint monitor (optional)

LSID authority: An endpoint which serves metadata for LSIDs



Services provided by LSID proxy and interaction with these

Resolution: Given an LSID (e.g. http://lsidproxy.tdwg.org/[lsid]), the proxy returns the metadata address for the authority endpoint associated with the LSID as an HTTP 30* redirect. *Mandatory* - all LSID proxies must support this.

Caching: On resolving an LSID (see the diagram above), the proxy service checks whether the LSID metadata is in the cache. If not, the LSID and metadata address are added to a cache population queue. This queue is processed as a background job at *an agreed rate*. "Processing" means resolving the metadata address to get structured metadata and adding to a local cache store. *Optional*.

Monitoring: On resolving an LSID (see above) the service checks whether the authority endpoint is being monitored for uptime. If not, the authority endpoint address and sample LSID are added to a list of authority endpoints for monitoring. Monitoring is a background process that periodically pings endpoints using a sample LSID call. *Optional*.

Alternative resolution process if caching is enabled (not shown on diagram): If the proxy service supports caching, and the endpoint is not available, the proxy should check cache for the metadata associated with the LSID. If it is found in the cache, the address of the cached location should be returned as a HTTP 307 redirect (or 302 Moved Temporarily for compatibility with HTTP 1.0 clients) (see http://www.w3c.org/Protocols/rfc2616/rfc2616-sec10.html). The cache address would be something like http://lsidproxy.tdwg.org/cache/[lsid] – note that the use of the cache is made explicit using the HTTP response codes, rather than the cached version being returned from the original request with HTTP status 200.

Statistics on the usage of the cache and endpoint uptime should be made available from the proxy service.

Distributed services

There are several ways in which this service infrastructure can be implemented.

- A centrally located and managed identifier hosting and provision service
  - assigns, stores, and resolves all hosted identifiers and data in a central location
  - requires substantial input and funding from the community
  - has the danger of being the bottle neck of the system, or single point of failure
  - may work well with a strong identifier business model
- A distributed identifier system
  - assigns identifiers at central location,
  - resolves identifiers back to the original data (the provider's specific web resource),
  - requires minimal implementation at a central location, and
  - provides multiple nodes for resolution and caching – improves persistence.

Either way, it is apparent that fail-over is an important issue. The resolution system should therefore use load balancing using round-robin DNS or similar. Multiple synchronised nodes providing resolution could be maintained.

Appendix 3: Other kinds of identifiers

UUIDs

A UUID (Universally Unique IDentifier) is an identifier created by an algorithm which virtually guarantees that no two identical UUIDs will ever be generated, anywhere in the Universe. This avoids the need to check for identical identifier values, and helps to ensure that future search and resolution mechanisms will find the correct instance and will not return multiple alternative interpretations. However, no identifier is proof against someone copying the identifier and then changing the data to which it refers.

Handles

The Handle System is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. The protocols specified enable a distributed computer system to store identifiers (names, or handles), of

digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. That information can be changed as needed to reflect the current state and/or location of the identified resource without changing the handle.

The Handle System enables management of objects as first class entities, rather than as packets of bits with dependency on other attributes such as locations.

The Domain Name System resolves domain names meaningful to humans into numerical IP addresses (locations of file servers). The Handle System is compatible with DNS but does not necessarily require it, unlike persistent identifiers such as PURLs or LSIDs which utilise domain names and are therefore ultimately constrained by them.

DOIs

The Digital Object Identifier (DOI) System is a managed system for persistent identification of entities on digital networks. "DOI" is parsed as "digital identifier of an object", rather than "identifier of a digital object". As well as identifying content items such as digital files and digital media manifestations of intellectual property, DOI names can also identify physical objects, performances and abstract works. For example, they can be used to identify: e-texts; images; audio or video items and software, etc. DOI names can also be assigned to related entities in a content transaction (e.g. licenses, parties, etc.) The DOI name is the identifier string that specifies a unique object (the referent); the DOI System is the functional deployment of DOI names as identifiers in computer sensible form through assignment, resolution, referent description, administration, etc.

DOI names resolve to data specified by the registrant, and use an extensible metadata model to associate descriptive and other elements of data with the DOI Name. The DOI System is an implementation of the Handle System and of the indecs Content Model, and so inherits the design principles and features of each.

The DOI System is implemented through a federation of DOI Registration Agencies, under policies and common infrastructure provided by the International DOI Foundation, which developed and controls the system.

Major applications currently include persistent citations in scholarly materials (journal articles, books, etc.) through CrossRef, scientific data sets, through a consortium of leading research libraries and technical information providers, building on work by the German National Library of Science and Technology (TIB), and European Union official publications, through the EU publications office

PURLs

A persistent uniform resource locator (PURL) is an HTTP Uniform Resource Identifier (URI) (i.e. location-based Uniform Resource Identifier or URI) with a redirect mechanism. It 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 (for example by means of a 302 HTTP status code) to the current location of the final resource.

PURLs are being promoted to solve the problem of transitory URIs in location-based URI schemes like HTTP. Persistence problems are caused by the practical impossibility of every user having their own domain name, and the inconvenience and money involved in re-registering domain names, that results in web authors putting their documents in rather arbitrary locations of questionable persistence (i.e. wherever they can get the web space).