Adoption of Persistent Identifiers for Biodiversity Informatics

Draft recommendations of the GBIF LGTG, 18 August 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)

This document is a draft prepared for review in advance of its submission to the GBIF Board of Governors. Please review and send any comments to Éamonn Ó Tuama, eotuama@gbif.org, by 7 September 2009, after which a final version will be produced.

Contents

1 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 peoples' data. Literature and training courses should be offered to those who don't understand 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.

2 List of recommendations

This is a summary of the full recommendations which appear in boxes later in the document in sections 7 and 8.

GBIF should:

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

2: provide materials such as an executive summary targeted to administrative leadership 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: 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.

3 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 draft report of the group.

4 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 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.

4.1 Persistent actionable identifiers

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

- **Persistence:** The property that an identifier 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 (e.g., 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 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 and/or its related object. 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.

4.2 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.

5 Some benefits of 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 flow 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.

The primary goal of this section is to describe some examples of opportunities for enhancement of the GBIF information services to provide specific feedback mechanisms that are of interest to the biodiversity informatics community.

5.1 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 are 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.

5.2 Management and disambiguation of taxon names

Disambiguation of taxon names requires services that support tests of difference as well as of

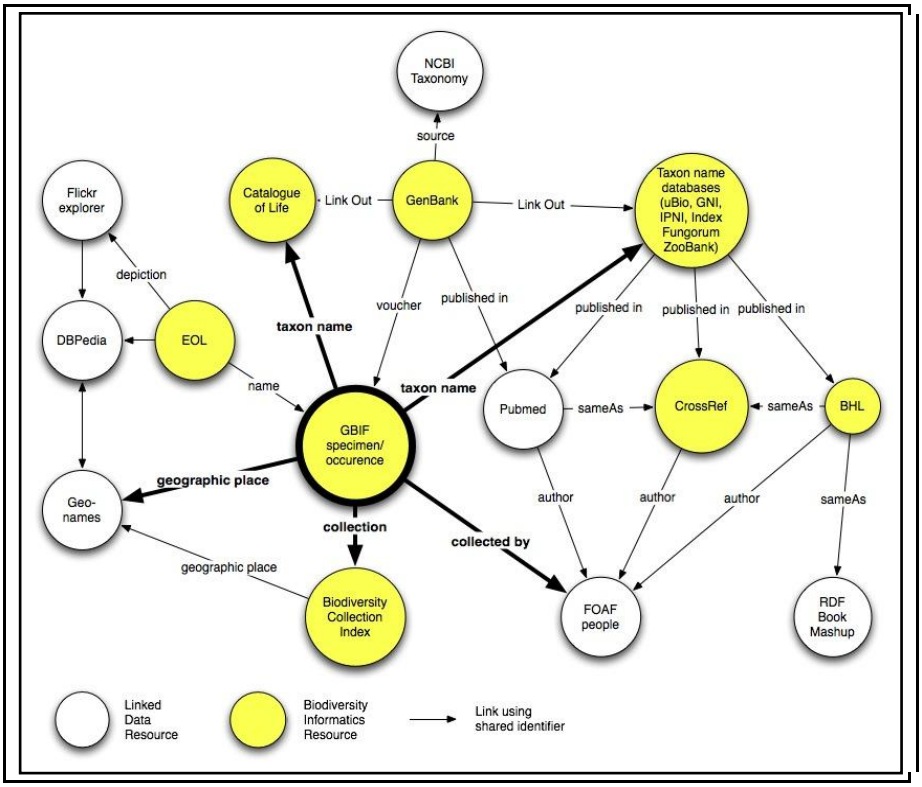
equality. A persistent identifier always refers to a specific object but different identifiers do not necessarily refer to different objects. A single object may have many identifiers. Tests of inequality for objects must rely on evaluation of metadata or of the objects themselves.

The ambiguities inherent in taxon name usage are described elsewhere. 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.

5.3 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, UniProt, a repository of genomics data (http://www.uniprot.org/), 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 was 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/).

5.4 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). This has implications for biodiversity informatics, in that our identifiers should be representable as HTTP URIs and our metadata representable as RDF.

6 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 described the kinds of identifiers which we recommend that GBIF should support, namely LSIDs and HTTP URIs. Other identifier mechanisms such as DOIs are in use in other data domains, and are described in an Appendix, since they may occur in data to which biodiversity data might be linked.

6.1 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.

6.2 Life Science Identifiers (LSIDs)

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

1. enables global uniqueness by including an Internet domain name, which is itself subject to rules and procedures ensuring uniqueness, and
2. 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. 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.

The bioguid proxy conforms to the Linked Data requirements (as in Section 4.2) by supporting content negotiation (requirement 2) and 303 redirects (requirement 3).

7 Role of GBIF in leadership and education

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

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.

7.1 Institutional support

An important aspect of 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.

7.2 Providing education, training and outreach

7.2.1 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.

7.2.2 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 GUIDs.
- 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 5.4 for more information about Linked Data requirements.

Education and support need to be targeted for three types of data providers, those 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.

7.3 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 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 graphs
- using existing vocabularies and GUIDs wherever possible.

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.

8 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.

8.1 Role of GBIF as a service provider

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

Why is it not sufficient to have a simple model where a user just goes to the original data provider for resolution? This simple model

- does not provide for data provider not yet online, and
- does not ensure reliability (sufficient up-time).

A service provider can be added to the model. If the 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 provider (e.g. 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.

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.

8.1.1 LSID resolution services

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.

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.

8.1.2 HTTP URI resolution

As with LSIDs, HTTP URI identifiers require best practices and a degree of infrastructural support. To help with HTTP URI identifier adoption, the following best practices should be encouraged:

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

8.1.3 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 are 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.

8.2 Availability of software resources

Recommendation 12: GBIF should extend the role of its data portal by hosting resources related to the use of identifiers, such as the TDWG vocabularies.

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, including such as server deployment packages, semantically aware clients, and validators for RDF, HTTP URIs and vocabularies, etc.

9 A business model for adopting identifier technologies

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

1. Software design and development (scripting) for redirect and caching services
2. Server hardware purchase, setup, housing, maintenance
3. Bandwidth
4. DNS setup and configuration maintenance
5. Curation: fixing broken redirects, populating and tracking replicates
6. Help desk
7. 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:

- The web itself provides unreliable and non-persistent resolution with costs happily taken on by data providers, with services often outsourced to various kinds of service providers.
- 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.
- 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 character.

- 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 availablility 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.

10 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://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)

11 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*).

12 Appendices

Appendices are not included in this draft document. They may be viewed in draft form at http://biodiversity.cs.cf.ac.uk/gbif/PersistentIdentifiers.html