A Beginner's Guide to Persistent Identifiers

Version 1.0

February 2011

Suggested citation:
GBIF (2011). A Beginner's Guide to Persistent Identifiers, version 1.0. Released on 9 February 2011. Authors Kevin Richards, Richard White, Nicola Nicolson, Richard Pyle, Copenhagen: Global Biodiversity Information Facility, 33 pp, accessible online at http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf.

ISBN: 87-92020-14-3

Persistent URI: http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf

Language: English

License:

Document Control:

Version	Description	Date of release	Author(s)
1.0		9 February 2011	

Table of Contents

Table of Contents 3
List of Figures 4
List of Tables 4
Executive Summary: Persistent Identifiers for the Life Sciences - In Summary 1
What is a "Persistent Identifier"? 1
Why are Persistent Identifiers useful? 1
Is a Persistent Identifier the same thing as a "GUID"? 1
What should Persistent Identifiers be applied to? 1
Is there only one kind of Persistent Identifier? 2
What good are Persistent Identifiers if they are "unfriendly" to read? 2
What is an "Actionable" Persistent Identifier, and why is this important? 2
How do I make my Persistent Identifiers "Actionable"? 2
Should I create a new Persistent Identifier for something that already has one? 2
Why should I support Persistent Identifiers for biological databases? 2
1. Introduction to Persistent Identifiers 3
a. What is a Persistent Identifier? 3
b. What do you apply an identifier to? 3
c. Data and Metadata 7
d. Use of internally unique, or database, identifiers 7
e. Choosing a Persistent Identifier scheme 8
f. URI Domain Names 9
2. Types of Persistent Identifier 11
a. URI 11
b. PURL 11
c. DOI 12
d. LSID 12
e. What might happen when things change 12
3. Managing Persistent Identifiers 15
a. Classes of data for identifier assignment 15
b. When to change the identifier if the data changes 16
c. Versioning of identifiers 17
4. Publishing Persistent Identifiers and their data 19
a. Working Groups and Recommendations 19
b. Vocabularies 19
c. Reuse of Persistent Identifiers 22
d. Scope of your data 23
e. Transferring datasets and their identifiers 23
f. Linked Data 24
g. Web services 25
5. Checklist for implementing Persistent Identifiers 27
6. References 28
7. Acronyms 29

List of Figures

Figure 1. Physical Specimen (Label)..... 3
Figure 2. Database record of specimen EC 12921. 4
Figure 3. Illustrating the various perspectives of a specimen. 5
Figure 4. Granularity of provided data. A specimen record providing data about the specimen itself and closely related resources such as identifications and collection events. 7
Figure 5. Specimen record with UUID identifier and resolvable URI identifier. 14
Figure 6. Our data broken down into the resources we want to specifically identify. ...16
Figure 7. Example taxon data. 16
Figure 8. Example of a publication record referencing an original taxon record. 18
Figure 9. Example of a version change requiring a reference to the original. 18
Figure 10. Example marked up using some available vocabularies (Darwin Core and Dublin Core) 22
Figure 11. An example showing how to refer to a strongly related resource. 23
Figure 12. An example showing how to refer to a source resource. 23
Figure 13. Example of various URI identifiers denoting different representations of a resource. 25

List of Tables

Table 1. Potential events impacting an existing Persistent Identifier scheme. 13

Executive Summary: Persistent Identifiers for the Life Sciences - In Summary

Note: This is a two-page summary/FAQ of the main points covered in this document.

What is a "Persistent Identifier"?

An identifier is a unique identification code that is applied to "something", so that the "something" can be unambiguously referenced. For example, a catalogue number is an identifier for a particular specimen, and an ISBN number is an identifier for a particular book. In the United States, each citizen is issued a Social Security Number, which is an identifier for each particular person. A *Persistent* Identifier is an identifier that is effectively permanently assigned to an object. For example, once an ISBN number is assigned to a particular book, that number is forever associated with that book, and no other book will ever receive that same number. Persistent Identifiers have their greatest value in the context of computer databases.

Why are Persistent Identifiers useful?

Persistent Identifiers are useful because they are unambiguous. In biology, we deal with information about many different things – specimens, taxonomic names, publications, people, localities, morphological characters, DNA sequences, and so on – and we have many different ways of referring to those things. A specimen, for example, might be referred to by its catalogue number (e.g., "BPBM 37615"), or a publication by a citation (e.g., "Baldwin & Smith, 1998"). But these identifiers leave room for ambiguity: there are two specimens with the catalogue number "BPBM 37615" (one is a fish, the other a mollusc), and there may be several articles published in 1998 by two authors with the names "Baldwin" and "Smith". Whereas a human can often discern the correct specimen and publication based on context, it is difficult for a computer to correctly interpret the context.

Is a Persistent Identifier the same thing as a "GUID"?

A Globally Unique IDentifier (GUID; rhymes with "squid") is effectively the same thing as a Persistent Identifier. However, the term "Persistent Identifier" is preferred for several reasons: 1) in recent years, the term "GUID" has increasingly been associated with a particular kind of identifier, called a Universally Unique Identifier ("Persistent Identifier" is more general); 2) The "GU" part of GUID emphasises global uniqueness, but doesn't emphasize *persistence* of that identifier (in the context of biodiversity informatics, persistence is a key factor); and 3) GUIDs are not necessarily "actionable" (that is, it is not always given that one can retrieve information about the object that is being identified), but in the biodiversity informatics community, the term "Persistent Identifier" is assumed to be an identifier for which the relevant associated information can be easily retrieved through the internet.

What should Persistent Identifiers be applied to?

Perhaps the most important question that needs to be very clearly answered before assigning Persistent Identifiers is: "What, exactly, does this identifier represent? The answer is not always obvious. Some of the objects we manage data for only exist in electronic form (e.g., a digital image or a PDF file). Others are actual physical objects that you can hold in your hand – like a museum specimen. Still others exist only in abstract form – such as a collecting event or a taxon concept. One perspective is that the identifier is not actually assigned to *any* of these things, but rather is assigned to the *database record* that was generated to represent these things. There is some debate about the best approach to follow, but the important thing is to be sure it is unambiguous what, exactly, the identifier is applied to.

Is there only one kind of Persistent Identifier?

There are several kinds of Persistent Identifiers in common use within biodiversity informatics, and each has its own set of advantages and disadvantages. Universal Resource Identifiers (URIs) are the type of identifier promoted by the World Wide Web Consortium. They look like a normal URL (Uniform Resource Locator) in that they rely on the HTTP protocol for retrieving the associated information. Another kind of Persistent Identifier that relies on HTTP is a Persistent Uniform Resource Locator (PURL), which is a kind of URI. Digital Object Identifiers (DOIs) are managed and maintained by the DOI Foundation, and are often used to refer to published articles, but they cost money. Life Science IDentifiers (LSIDs) were developed by IBM specifically for the life-sciences community, and have many features conducive for use with biological datasets. LSIDs follow the template: **urn:lsid:<authority>:<namespace>:<ObjectID>:[version]**. There are several factors to consider when deciding on which of these kinds of Persistent Identifiers should be used. Moreover, they are not mutually exclusive. For example the LSID urn:lsid:zoobank.org:act:8BDC0735-FEA4-4298-83FA-D04F67C3FBEC can be formed as a URI: http://zoobank.org/urn:lsid:zoobank.org:act:8BDC0735-FEA4-4298-83FA-D04F67C3FBEC.

What good are Persistent Identifiers if they are "unfriendly" to read?

It is important to understand that Persistent Identifiers are intended for computers to communicate with other computers. As such, they should be invisible to most users. In fact one of the important qualities of a good Persistent Identifier is *opacity*. That is, the identifier itself should not contain any readily identifiable information.

What is an "Actionable" Persistent Identifier, and why is this important?

Uniqueness and persistence are important, but an identifier is only useful if information about the data object that it represents can be easily retrieved. This is similar to a web address URL. The text "http://www.gbif.org" by itself isn't very useful. But if you enter that text into the address bar of your web browser, you gain access to vast information. Good Persistent Identifiers should always be exposed in a form that is "self-resolving" and, hence, actionable.

How do I make my Persistent Identifiers "Actionable"?

The mechanism for making Persistent Identifiers actionable depends on the kind of Persistent Identifier. There are good sources of information available (including this *Beginner's Guide*) to address the technical details of how to serve information related to identifiers.

Should I create a new Persistent Identifier for something that already has one?

If a Persistent Identifier already exists for an object you want to reference, it is generally better to re-use that existing identifier, rather than generate a new one. However, whenever a service adds new information, or in some way changes the content or meaning of the object represented by an existing identifier, it is sometimes useful to assign a new identifier (in such cases, it is best to refer back to the existing identifier within the information associated with the new identifier).

Why should I support Persistent Identifiers for biological databases?

If you have ever used the internet to locate information relevant to biology, then you should support the implementation of Persistent Identifiers for biological data. Although the Internet has provided unprecedented access to biological data, including images, specimen data, DNA sequences, publications, taxonomy, and more…broad implementation of actionable Persistent Identifiers would dramatically improve the ease of accessing this information. Such identifiers would allow universal cross-linking of relevant information. Search for a species and see all publications citing that species, as well as all images and specimens identified to that species (or one of its synonyms); click a specimen and see who identified it; click on the person's name and see all other specimens identified by that person; and so on – a universe of information at your fingertips.

1. Introduction to Persistent Identifiers

This guide is intended to cover the essential principles of Persistent Identifiers and demonstrate the requirements to start issuing and delivering Persistent Identifiers for biodiversity informatics datasets. The guide is aimed at the Persistent Identifier novice, highlighting pitfalls and suggesting useful tips to get up and running with identifiers. There is an assumption that standard best-practice data management principles have been followed regarding any applicable dataset.

a. What is a Persistent Identifier?

Identifiers for electronic data have been around since computers started to become widely used and the need arose for multiple parties to refer to the same digital resource. Identifiers are a way of giving digital resources, such as documents, images and database records, a unique reference number, in the same way that ISBN [ISBN] numbers work for books and social security numbers work for people.

The other part to **Persistent Identifiers** is the **Persistent** part. This notion is intended to encourage the ongoing support of an identifier. An identifier is of little use if it is short-lived.

- ✓ Persistent Identifiers must be globally unique
- ✓ Persistent Identifiers must exist indefinitely

b. What do you apply an identifier to?

This is often an area of confusion and debate – what sorts of things do we apply identifiers to? Do we apply them to real-world resources such as herbarium specimens, or to conceptual resources such as identifications of specimens? Do we apply them to digital resources (database records) or to the actual physical resources? Or perhaps to all of these? The confusion normally arises because of a lack of clarity within a given community about what to apply identifiers to, or even how to define the resources that identifiers are applied to.

Physical vs. Digital Resources

A good example to demonstrate this issue is that of a specimen in a natural history collection. The label for a physical specimen is depicted in Figure 1.

Collection :	Example Collection (EC)
Accession Number :	EC 12921
Date Collected :	1 August 2001
Collector :	B. Smith
Identified To :	Amanita alba Lam.
Determiner:	B. Smith
Date Identified :	14 August 2001
Locality :	Wards Rd, Canterbury, New Zealand

Figure 1. Physical Specimen (Label)

As we can see from the label, this specimen has an Accession Number that could be used to identify it. This may well be suitable for identifying the physical specimen itself, but it will quickly become inadequate outside this context.

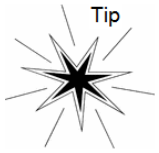
As this guide is primarily concerned with sharing digital data over the Internet, the next step is to consider the digitisation of this specimen. (N.B., in this document the term "digitisation" refers to the creation of digital data from a physical resource, and does not refer to the more specific meaning of creating images of physical resources such as specimens). Due to the fact that the specimen Accession Number quickly becomes inadequate in the wider context, specimens are often assigned another identifier when entered into a database. An example of this is shown in Figure 2 (SpecimenId).

SpecimenId	AccessionNo	DateCollected	Collector	IdentifiedTo	...
...					
3422	EC 12921	2001-08-01	Smith, B.	Amanita alba Lam.	
...					

Figure 2. Database record of specimen EC 12921.

We now have "representations" of the physical resource, including the physical specimen (identified by the Accession Number EC 12921), and the database record (identified by the SpecimenId 3422). When someone refers to the "specimen", to which of the two representations are they referring? The same question arises when you decide to apply a Persistent Identifier to this specimen – which of the two representations does the Persistent Identifier refer to? Do we stamp the physical specimen with the Persistent Identifier (printing the identifier on the label of the specimen perhaps) or do we apply the Persistent Identifier to the database record?

Another way to think about this is to consider the context in which we want to use Persistent Identifiers. Due to the fact that we want to provide digital information about particular resources over the Internet, our primary referent is the database record. Perhaps, therefore, we should apply the Persistent Identifier to the database record, leaving the responsibility of the synchronisation of the database record and the physical specimen up to the data custodian. This decision may vary depending on the type of resource that is being identified. Another approach to this issue is discussed later in the Linked Data section.



Decide on an approach to assigning your identifiers early on and stick to this decision, whether that decision is to apply identifiers to physical resources or conceptual resources.

To illustrate these aspects of resources and their identifiers, Figure 3 shows three different perspectives.

To summarise how these perspectives interact and fit together we can use a well known method called CRUD (Create, Read, Update, Delete), which describes when each type of resource is created, retrieved, modified, and disposed of.

CREATE - Database records are created from physical resources via the digitisation process. Digital resources are created from database records via assignment of a Persistent Identifier, and the mapping of the data into a specific document format.

READ - "What" is read is defined by the data format used to present the values of the digital resource. "How" reads are done is defined by the identifier resolution process.

UPDATE - Updates will be made via curation, either of the "physical resource" itself, or of the "database record" (held by the curating institute), or proposed by users of the "digital resource".

Physical specimen

Table: Specimen
ID: 3422
AccNum: EC 12921
Collector: B. Smith
DateCollected: 1/8/2001
IdentifiedTo: Amanita alba Lam.
SecurityLevel: public

Database record of the specimen

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF>
      <rdf:Description
rdf:about="http://example.org/specimen/ec12921">
            <rdfs:type
rdf:resource="http://rs.tdwg.org/dwc/terms/Occurrence"/>
            <dwc:basisOfRecord
rdf:resource="http://rs.tdwg.org/dwc/dwctype/PreservedSpeci
men"/>
            <dwc:catalogNumber>EC 12921</dwc:catalogNumber>
            <dwc:recordedBy>B. Smith</dwc:recordedBy>
            <dwc:eventDate>2001-01-08</dwc:eventDate>
            <dwc:institutionCode>EC</dwc:institutionCode>
      </rdf:Description>
</rdf:RDF>
```

Digital Object - RDF document representing the specimen, for data transfer

Figure 3. Illustrating the various perspectives of a specimen.

DELETE - The curating institute must be explicit about how deletes are managed. It is crucial that a record is kept of which digital resources have had Persistent Identifiers applied to them, even after the physical resource and/or database record has been deleted and/or destroyed. This issue comes to the fore when a consumer of the resource that has been deleted has no way to obtain the information for that resource, or even to know for sure if that resource no longer exists.

Because of the complexity of the relation between identifiers and the resources they refer to, it is useful to break down the various types of resource into more precise, explicit parts. For example, the following descriptions have been proposed [Baskauf2010]:

resource - a physical, digital or conceptual entity which can be identified by a Uniform Resource Identifier (URI) [Berners-Lee].

information resource - a resource for which all essential characteristics can be transmitted in a message [JacobsWalsh], i.e. a digital resource. Examples: text and digital images.

data - the content of the message that provides the representation of the information resource.

non-information resource - a resource that cannot be transmitted electronically. Technically, a resource is defined as a non-information resource when an HTTP GET request for the resource does not result in a 2xx "Success" response, i.e. no data are returned [W3CTech]. Examples: persons, specimens.

metadata - data about data. All resources can have metadata that describe their properties.

physical resource - a non-information resource that is a material thing. Examples: living organisms, specimens, and 35mm slide images.

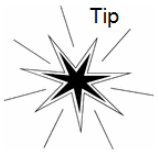
abstract resource - a non-information resource that does not represent a particular material thing. Examples: observations, protein structures, state boundaries, and concepts.

conceptual resource - an abstract resource that is subject to varying interpretation. Examples: relationships, taxonomic concepts, and properties.

defined abstract resource - a resource that represents a defined circumstance or abstract resource, and which therefore is not subject to interpretation. Examples: observations, determinations, and mathematical concepts.

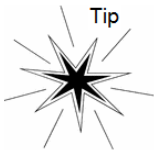
Granularity of Resources for Identification

An important decision when applying Persistent Identifiers is the granularity of the data that is given for a particular identified resource. Granularity is how much detail the data for the identifier cover: whether it covers just the few essential properties of the resource, the entire detail of the resource, or perhaps the resource plus any data that are possibly related to it. The importance of granularity can be shown by again considering the specimen EC 12921. When a Persistent Identifier is given to this specimen, should the data for this identifier be about the specimen record only, such as the Accession Number, Collector and Date Collected, or should they include identifications, history, loan information, images, derived specimens and other related information? This decision is normally dependent on the reason for providing the data and their intended use. But there are several general rules that it pays to follow:



Apply identifiers to the primary resources in your dataset, i.e., the things that consumers of your data are likely to ask for individually.

Restrict the data that are returned for an identifier to those data immediately related to the resource, and closely related sub-resources.



When assessing what forms a "dataset", consider which set of records could be transferred to another curating institution.

Consider whether resources you are identifying can be independently curated.

For example, an identifier for a specimen resource would likely provide data for the specimen itself, its identifications and collection events, as shown in Figure 4. Other related data can be indicated with additional Persistent Identifiers. Referring to related data is an important part of providing your data over the web as interlinking makes them more useful to users (see the section on interlinked data, 4f).

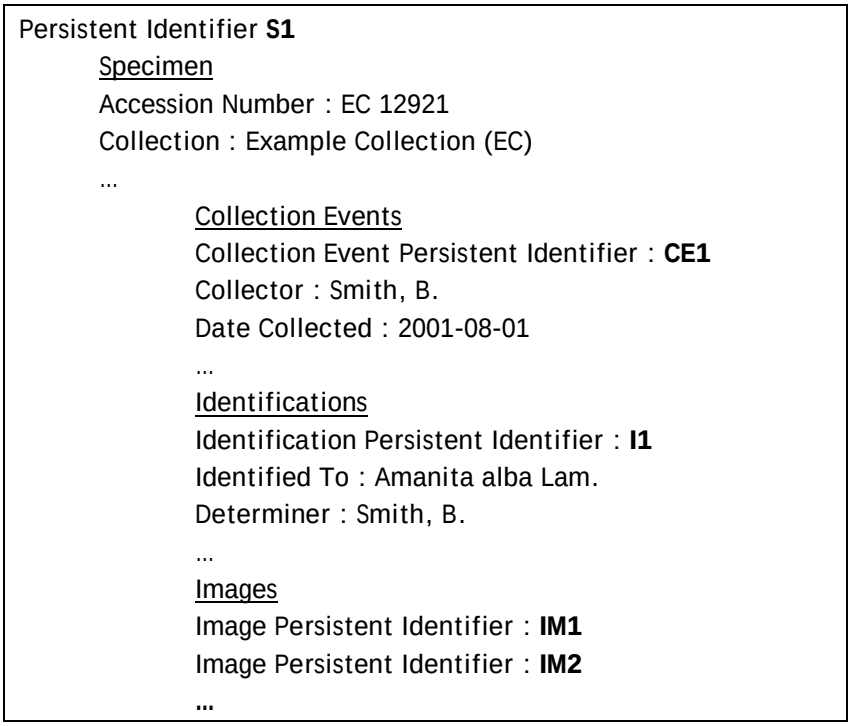


Figure 4. Granularity of provided data. A specimen record providing data about the specimen itself and closely related resources such as identifications and collection events.

c. Data and Metadata

Another source of confusion when assigning Persistent Identifiers arises because the terms "data" and "metadata" can often be used interchangeably (and sometimes the difference is not clear).

Generally "data" have been described as the information about the resource in question, and "metadata" are "data about data". Sometimes the boundaries of these two are blurred and some people even believe that all information in the biodiversity informatics domain is metadata because we are discussing physical resources and events. Data are also believed to be more enduring, whereas metadata may often change.

Overall it does not matter too much whether data or metadata are the subject of a specific Persistent Identifier so long as the identifier always refers to the same resource.

d. Use of internally unique, or database, identifiers

The course of the development of particular datasets and digital information is often piecemeal and this can lead to inadequate identifiers for those digital resources.

A common case is to start with a personal dataset, often maintained in a spreadsheet, which then becomes more broadly valuable, at which point the dataset is transferred into a database giving the data records the standard automatic numeric identifiers. When the dataset then becomes valuable to a still wider audience and provision of this dataset over the Internet is desirable, the data are provided using the numeric identifier, probably on the end of a website URL. This is not a good solution for several reasons. It is extremely difficult to guarantee the uniqueness of numeric identifiers, considering the fact that most datasets and lists of records are numerically numbered and begin at number 1. When merging multiple datasets of the same information it will be necessary to resolve the conflicts of the numeric identifiers, inevitably resulting in new identifiers being assigned

to those conflicting records. These issues make numeric identifiers difficult to maintain as globally unique identifiers, or even as part of a global unique identifier format.
Even if a database identifier is theoretically unique, such as the UUID (Universally Unique Identifier) [UUID], [IEFTUUID], it requires more context and scoping to make it effective as a Persistent Identifier. For example, the UUID {10FC9784-B30F-48ED-8DB5-FF65A2A9934E} is practically guaranteed to be unique globally, but is of little use to someone who comes across it on the Internet. An ID by itself is of little use without directions to the data to which it is assigned.

{10FC9784-B30F-48ED-8DB5-FF65A2A9934E}

Another issue with using internal database identifiers is the likelihood that the database record ID will change. If, for some reason, the database is restructured or refreshed, then particular effort is required to ensure the Persistent Identifiers for this dataset are not broken. This process is made simpler by keeping the Persistent Identifier mechanism separate from the database record identification, and by ensuring management processes are in place to handle the synchronisation of Persistent Identifiers to their associated data. For more discussion on versioning of identifiers, see section 3c.

One attribute of a good Persistent Identifier scheme is support to maintain the identifier in perpetuity. Some schemes enforce management practices for registration and resolution of identifiers, thereby cementing the identifier permanently.

e. Choosing a Persistent Identifier scheme

It will be useful here to identify three elements relevant to the dataset that you want to expose using Persistent Identifiers:

1. Authority
2. Context
3. Object

These terms assist with defining the format of an identifier and refer to the broad to narrow contexts in which the identifiers exist, Authority being the broadest to Object being the most specific.

Authority

This generally correlates to the institute or organisation which has responsibility for the dataset (also known as the information resource). Important considerations when choosing an authority include:

- Stability – organisations that endure little change to their name and the names of their dependencies are a better choice for long term authority names.
- Longevity – organisations that have more community support are more likely to persist into the future.

Context

This correlates to the dataset (which might be a database or a clearly identified subset of a database). Considerations when choosing a context name include:

- Contexts must be unique within an authority.
- What subset of the information resource can be independently curated? If you are applying Persistent Identifiers to a large, heterogeneous dataset (perhaps actively curated in some areas, and fairly static in others) or to a dataset which has been derived from multiple sources, consider splitting the dataset into subsets. Also

consider what might happen to information resources if the authority were to wind down, split or merge with another organisation.

Object
This correlates to the specific resource (possibly a database row). Object names (identifiers) must be unique within a context.

Opacity is of particular importance when considering an object name.

The aim that an object name is opaque means that when using the Persistent Identifier and its associated metadata, the user should not assume anything from the format of the identifier itself. In other words, relationships between resources can only be seen by resolving the Persistent Identifier and examining its metadata.

Consider a dataset of observations, created over a time sequence. These observations are exposed using the following Persistent Identifier format:
http://dataset.xyz.org/observations/100001

Opacity means that the user of the Persistent Identifiers should not assume that the next observation in the time sequence is http://dataset.xyz.org/observations/100002

The identity of the "next observation" can only be determined by accessing the metadata for the observation http://dataset.xyz.org/observations/100001 and looking for a statement which indicates this, i.e.:

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix obs: <http://xyz.org/obs/elements/1.1/>.

<http://dataset.xyz.org/observations/100001>
  obs:precedes http://dataset.xyz.org/observations/100002
```

This principle is based on a frequently occurring design goal for relational databases – that the primary key of a record is opaque and does not indicate anything about the content of the record itself. This example shows that sequential numeric keys are not very opaque (and can lead to users forming misguided assumptions about the content of the data resource), but an object identifier such as a UUID (formed of 32 hexadecimal numbers) would be very opaque and prevent assumptions.

f. URI Domain Names

URI : http://www.example.org/specimen/12921

By using a domain name under the control of the owner of the identifiers, URIs allow the creation of unique identifiers by defining a global scope that can be controlled by that owner. In the example given above, the authority-defining portion of the URI is the www.example.org component.

The "dataset" or "context" portion of the URI further limits the scope of the set of resources we are identifying. In this case the context portion is "specimen", indicating that the identifiers refer to specimen-type resources.

This approach is often seen as breaking the rule of opacity with identifiers, i.e., by using domain names and common English words for the context, it is easy to infer details and relationships about the data the identifier refers to. The pros and cons of using the URI approach need to be examined and a decision made whether this is of concern to the data

provider. By taking the perspective that Persistent Identifiers are primarily used for computer to computer communication, and therefore not involving interpretation by humans, the issue of opacity is trivialised.

The use of URIs has undergone much debate over recent years, especially with the confusion between URIs and URLs (Uniform Resource Locator). URLs are location oriented and therefore describe the "physical" location of a resource on the Internet. URIs are identifier oriented and therefore, theoretically, refer to the resource itself, not its location [URI]. For this reason, URIs should be much more resilient, but because of the similarity between URIs and URLs, they can easily be treated the same way. In this way, URIs can be changed mistakenly in the same way that a URL is changed when the location of a resource moves and can result in broken links. When issuing Persistent Identifiers:

- ✓ Do not simply use database record identifiers as Persistent Identifiers.
- ✓ Do define a context name under your control to ensure the global uniqueness of your identifiers.
- ✓ Do use your context name to define the scope of your identifiers.
- ✓ Do put management processes in place to ensure Persistent Identifiers and their associated data are always synchronised, so that when database records change the Persistent Identifiers for those records are reviewed promptly.
- ✓ Do choose context names that are institution independent, i.e. project names rather than organisation names. Larger, more vital projects will obviously be a more likely permanent context name due to the fact that the community is more likely to ensure the survival of the name.

2. Types of Persistent Identifier

There have been many attempts to define Persistent Identifier mechanisms for identifying digital resources on the Internet. Some of the more common options are listed below. Most identifier mechanisms follow some or all of the principles that define an ideal identifier. Most of these principles can be summarised as:

- Universally unique – a system to ensure each generated identifier is unique worldwide. Generation normally involves a defined algorithm or process that creates a new unique identifier.
- Independent generation – it should be possible to generate an identifier without the need for a centralised system. This requires a defined process and format for defining the identifier structure.
- Unchanging – the identifier should never change (both the identifier itself and the resource the identifier is applied to).
- Opaque – it should not be possible to determine any detail about the identified resource by looking at the identifier alone.
- Actionable (sometimes) – the identifier can be de-referenced so that the data about the resource can be retrieved. This is also called resolution of an identifier.

The following types of Persistent Identifier are commonly in use:

a. URI

Uniform **R**esource **I**dentifier

- E.g., http://www.example.org/specimen/12921
- Web-based identifier – URIs therefore rely on the DNS (Domain Name System).
- Promoted by the W3C [W3C] and the IETF [IEFTURI].
- Independent generation is enabled by the use of domain names. A domain name can be used as the authority component, then the context and identifier part of the Persistent Identifier is determined by the provider.
- May not be opaque due to the use of domain names, context names and sometimes descriptive object identifiers.
- Resolution is achieved via standard web HTTP resolution.

b. PURL

Persistent **U**niform **R**esource **L**ocator

PURL identifiers are based on URIs and use the HTTP redirect mechanism to avoid broken links.

- E.g., http://purl.oclc.org/example/specimen/12921
  http://purl.org/dc/terms/contributor
- Web-based identifier using standard HTTP and HTTP redirect. Can be resolved through use of a common PURL resolver.
- Promoted by the OCLC (Online Computer Library Center) [OCLC]
- Independent generation is enabled by the use of domain names.
- May not be opaque due to the use of domain names, context names and sometimes descriptive object identifiers.
- Authority, context and object identifier components can be defined using the path portion of the PURL (as shown in the example above).

c. DOI

Digital **O**bject **I**dentifier

DOI identifiers are a managed identifier system, maintained and controlled by the DOI Foundation [DOI]. The DOI Foundation manages a commercial infrastructure for the assignment and use of DOI identifiers.

- E.g., 10.1000/186
  http://dx.doi.org/10.1000/186
- DOI identifiers must be bought at a cost per identifier from the DOI Foundation.
- Generated on demand by the DOI Foundation.
- Registration, support, persistence control and policy making is provided by the DOI Foundation, ensuring a robust system for maintaining the identifiers.
- Resolved through the online DOI resolver by appending the DOI to the URL http://dx.doi.org/
- DOIs are very opaque.
- Authority, context and object identifier components are obscured with use of DOIs (which can be seen as a positive aspect if opacity is deemed important).

d. LSID

Life **S**cience **Id**entifier

Life Science Identifiers were initially created to provide a protocol-independent identifier mechanism for the life sciences community. By using a URN (Uniform Resource Name) instead of a URL, a specific protocol (HTTP) is avoided. This is seen as an advantage because the identifier is much less likely to be invalidated by broken web links, but is also seen as a disadvantage because LSIDs cannot be resolved using basic HTTP resolution, and therefore cannot be resolved by most web browsers.

- E.g., urn:lsid:example.org:specimen:12921
  urn:lsid:[authority]:[namespace]:[object id]
- LSIDs have a specification for assignment, structure and resolution.
- LSIDs can be generated and adopted by anyone willing to set up an LSID resolution service.
- Resolution requires a three-step process, including a DNS lookup, a call to retrieve a SOAP [SOAP] web service description document (WSDL [WSDL]), and a standard web resolution.
- LSIDs have been recommended by the biodiversity informatics community, but not to the exclusion of any other Persistent Identifier type.
- LSIDs may not be opaque due to the domain name and context name components and descriptive object identifiers.
- The authority portion is provided by the LSID domain name; the context is provided by the LSID namespace portion and the object identifier is provided by the object ID portion of the LSID.
- LSIDs provide built-in versioning abilities. See section 3b and 3c for more discussion on this topic.
- It is worth noting that these different types of Persistent Identifiers are not necessarily mutually exclusive. For example, an LSID could be represented as a URI using what is known as an "HTTP Proxy", whereby the LSID is embedded within a URI, e.g., http://example.org/urn:lsid:example.org:specimen:12921.

e. What might happen when things change

This section aims to identify what might happen once a Persistent Identifier scheme has been implemented on your information resource. These events (Table 1) are presented in

order of impact. Those relating to the authority (organisation) are high impact, while those relating to the object are lower impact (but likely to happen more frequently).

Table 1. Potential events impacting an existing Persistent Identifier scheme.

Level	Event	Desired effect	What is needed to achieve desired effect			
			URI	**PURL**	**DOI**	**LSID**
Authority	Organisation renamed	Old organisation name is maintained for the identifiers	DNS domain name must be kept and redirected	PURL resolver redirects to the new organisation domain name	DOI resolver redirects to the new organisation domain name	LSID SVN records need updating and LSID resolver redirects to new organisation domain name
	Organisation restructured	New organisation department maintains old identifiers	New department takes control of the web site that serves the URI documents	New department takes control of the web site that serves the PURL documents	New department takes control of the web site that serves the DOI documents	New department takes control of the LSID resolver
	Domain name lost	Old identifiers are maintained	Not possible	PURL resolver is redirected to the new domain name	DOI resolver is redirected to the new domain name	Not currently possible. LSIDs require control of the authority domain name for DNS lookup.
Context	Dataset transferred to a different authority	New authority maintains the old identifiers	Only possible if the old authority redirects to the new authority for resolution	PURL resolver is redirected to the new authority	DOI resolver is redirected to the new authority	LSID resolver is redirected to the new authority LSID resolver
	Datasets merged / split / renamed	Both dataset identifiers remain resolvable	Mapping (URI redirect) required from old dataset name to new dataset	PURL resolver redirects to new dataset resolver	DOI resolver redirects to new dataset resolver	LSID resolver redirects to new dataset resolver
Object	Expression of metadata changes (different vocabulary is used)	The meaning and understanding of the resource the identifier	Document that the URI resolves to needs updating	Document that the PURL resolves to needs updating	Document that the DOI resolves to needs updating	Metadata response for that LSID needs updating

		refers to remains the same				
	Content of metadata changes	Metadata includes an indication of when the record was last updated	Document that the URI resolves to needs updating	Document that the PURL resolves to needs updating	Document that the DOI resolves to needs updating	Metadata response for that LSID needs updating
	Resource is destroyed	Persistent Identifier continues to resolve. Metadata explicitly states that the resource has been destroyed. Object identifier is not reused.	Document that the URI resolves to needs updating to specify this record is deprecated	Document that the PURL resolves to needs updating	Document that the DOI resolves to needs updating	Metadata response for that LSID needs updating. Version of the LSID may be updated if there is a replacement record

Example

For our example, we will use a combination of UUID and URI identifiers. As UUID identifiers are opaque and unchanging, we will use these as the base identifier of our records. This will ensure that any data manipulation, merging and reorganisation we do internally to our dataset will not upset the identification of that resource (internally or externally). We can then use URI identifiers in conjunction with the UUID to enable web resolution of our identifiers. This is shown in Figure 5 below.

SpecimenId	SpecimenURI	AccessionNo	DateCollected	Collector	…
…					
10FC9784-B30F-48ED-8DB5-FF65A2A9934E	http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E	EC 12921	2001-08-01	Smith, B.	
…					

Figure 5. Specimen record with UUID identifier and resolvable URI identifier.

3. Managing Persistent Identifiers

a. Classes of data for identifier assignment

One thing that must be decided when issuing Persistent Identifiers for your data is what type of data you will be providing. Several issues must be considered when looking at this, including the data type and the scope of the details provided for a single identifier.

- **Type of data**. What kind of data are you providing? Is it data about your specimens, about an image library you have, about scientific names, or perhaps observations in the field? It is important to define the types of data you have in your digital repository, and which of those you want to be able to provide over the Internet. This will be influenced by the predicted or requested uses of your data. For example, users may want to use your identifier service for referencing the resources that you curate.
- **Scope of the data, or granularity**. This was discussed in section 1b, but is an important consideration to review when defining the data to be identified. At this point you need to decide what core data types it makes sense to apply identifiers to (i.e., what are the core data resource types that you work with). If you are very focused on a specific type of resource, say an observation dataset, and have little need to refer specifically to peripheral resource types (such as the collector of the observation), then it may be that you would generate identifiers only for the observation resources. All data related to an observation would be included in the record about the observation. If, however, you are often referring to common peripheral resources, such as the taxon to which the observations are identified, it would be good to provide your own identifiers for these too.

In our example, we have **Specimens**, which are collected during a **Collection Event** and **Identified** to a **Taxon**, as shown in Figure 6.

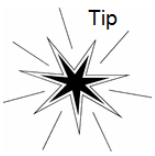
Specimen
UUID: 10FC9784-B30F-48ED-8DB5-FF65A2A9934E
URI: http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E
Accession Number: EC 12921
Collection: Example Collection (EC)
CollectionEvent ID: 942F4061-885E-497F-B5EE-0E689CDA8E66
Identified To ID: 6EF499AE-9F5B-4106-82C3-E8330A71F591
...

Collection Event
UUID: 942F4061-885E-497F-B5EE-0E689CDA8E66
URI: http://www.example.org/collection-event/942F4061-885E-497F-B5EE-0E689CDA8E66
Collector: Smith, B.
Date Collected: 2001-08-01
Locality: Lincoln, Canterbury, New Zealand.
...

Identification
UUID: 6EF499AE-9F5B-4106-82C3-E8330A71F591
URI: http://www.example.org/identification/6EF499AE-9F5B-4106-82C3-E8330A71F591
Identified To: Amanita alba Lam.
Identified To Taxon ID: 7217D220-836A-11DF-8395-0800200C9A66
Determiner: Smith, B.
Identified Date: 2001-08-03
...

Taxon
UUID: 7217D220-836A-11DF-8395-0800200C9A66
URI: http://www.example.org/taxon/ 7217D220-836A-11DF-8395-0800200C9A66
Scientific Name: Amanita alba Lam.
Publication: *Encycl. Méth. Bot .***1**(1): 107 (1783)
Taxon According To: Example Collection (EC)

Figure 6. Our data broken down into the resources we want to specifically identify.



To decide if a resource type needs an identifier, ask yourself:
- "Do I refer to a specific resource of this type multiple times?"
- "Will other people want to specifically refer to a resource of this type?"
- "Is this a regulated field or value, e.g., Species Name?"

Common classes of data for the biodiversity domain include:
- Collection
- Specimen
- Observation
- Taxon Name
- Taxon Concept
- Publication and Citations
- Identification
- People (Collector, Determiner, Observer, etc.)
- Taxon description
- Event
- Locality, Geo Region

b. When to change the identifier if the data changes

Once you have decided that you have some digital resources to apply Persistent Identifiers to, and know the type and scope of those resources, there is an immediate assumption, or even a requirement, to keep the data for those resources consistent indefinitely. This may sound difficult or even impossible – how is it possible to never change a database record? It is therefore important to decide on what constitutes the primary properties of the identified resource, and what degree of change will mean the resource has fundamentally changed and therefore will require a new identifier.

For example, Figure 7 shows the taxon name "Amanita alba Lam.", published in "*Encycl. Méth. Bot .***1**(1): 107 (1983)".

Taxon
UUID: 7217D220-836A-11DF-8395-0800200C9A66
URI: http://www.example.org/taxon/ 7217D220-836A-11DF-8395-0800200C9A66
Scientific Name: Amanita alba Lam.
Publication: *Encycl. Méth. Bot .***1**(1): 107 (1983)
Taxon According To: Example Collection (EC)

Figure 7. Example taxon data.

Perhaps we discover there was a typo in our data, such that the year the name was published was supposed to be **1783**, not **1983**. This is a fundamental piece of information about our taxon name. Does this mean it is a different name? Or is it just a correction? The

answer to this will often depend on the context in which the data have been created and in which they will be used. Generally, a good place for these decisions to be made is within working groups of the appropriate community. Some data operations will always modify the data to a degree that requires the creation of a new digital resource, for example changing the "according to" literature citation for a taxonomic name [Güntsch]. One possible approach is to require a new identifier if the data for a resource change, but not require it if only the metadata change (see section 1c for discussion of data and metadata).

A few rules to help make this decision are:

- ✓ Define the types of resources that are being identified.
- ✓ Define what constitutes "data" and what constitutes "metadata" with respect to your resource types.
- ✓ Define the fundamental properties of each resource type that define that resource. These properties should be values that, when changed, will change the meaning and understanding of that resource.
- ✓ Define relationships between resources. Check that if any properties of a resource change, that this will not fundamentally change the meaning of the relationships with other resources.
- ✓ Define the degree to which each property can change before it results in a different resource. Sometimes this will be nil, where a value cannot change at all, and sometimes there will be a degree of tolerance of change.
- ✓ Examine the number of changes applied to your dataset over a period of time. Be aware that the scale of change may not be consistent over time – a dataset may undergo a period of intense data cleaning prior to internet publication, after which the number of changes may be fewer.
- ✓ Consider making public the scale of changes to your data so that those referencing your data using Persistent Identifiers are aware of this dimension.

c. Versioning of identifiers

When the data for a resource do change to a degree that results in a fundamentally different resource, you will need to decide how to handle this. In some cases you may decide that it does result in a completely new resource, and sometimes you may decide it results in a different version of the same resource. Either way it is important to maintain links between the two editions of the resource.

For example, if we look at the publication of our taxon name "*Encycl. Méth. Bot .***1**(1): 107 (1783)", and follow a few edits, we can examine different possibilities for versioning the data for the publication. The taxon name publication is based on a PDF of the publication article (Figure 8).

Original Taxon Record:

UUID:	7217D220-836A-11DF-8395-0800200C9A66
URI:	http://www.example.org/taxon/ 7217D220-836A-11DF-8395-0800200C9A66
Scientific Name:	Amanita alba Lam.
Taxon Rank:	species
Publication:	*Encycl. Méth. Bot .***1**(1): 107 (1784)
PublicationID:	996793C1-DA90-47C8-87C3-AF24031AA21A
Taxon According To:	Example Collection

Publication Record:

UUID:	996793C1-DA90-47C8-87C3-AF24031AA21A
URI:	http://www.example.org/publication/ 996793C1-DA90-47C8-87C3-AF24031AA21A

File:	emb01010107.pdf
Citation:	*Encycl. Méth. Bot* .**1**(1): 107 (1784)
Journal:	*Encycl. Méth. Bot.*
Volume:	*1*
Page:	*107*
Year:	*1983*

Figure 8. Example of a publication record referencing an original taxon record.

If we consider a correction to this record (i.e. the year has been wrongly recorded as 1983, then corrected to 1783), then this could be seen as a minor metadata correction as this is not a change to the data representing the resource (i.e. the PDF document has not been modified). This will not require a new version.

If we then consider a more significant change, say the PDF has been altered to include more pages, then this would be seen as a change to the actual resource and therefore require a version change. This could be achieved by assigning a new identifier to the "new" resource and referring to the original in the metadata. This could be reflected in the records as follows.

UUID:	C25BD906-04BC-4311-A86F-7F81D72031A3
URI:	http://www.example.org/publication/ C25BD906-04BC-4311-A86F-7F81D72031A3
File:	emb01010107-1.pdf
DerivedFrom:	996793C1-DA90-47C8-87C3-AF24031AA21A
Citation:	*Encycl. Méth. Bot* .**1**(1): 107 (1784)
Journal:	*Encycl. Méth. Bot.*
Volume:	*1*
Page:	*107*
Year:	*1783*

Figure 9. Example of a version change requiring a reference to the original.

4. Publishing Persistent Identifiers and their data

a. Working Groups and Recommendations

During recent years several working groups have undertaken the task of reviewing the status, technologies and best practices of working with Persistent Identifiers.

One of these, the TDWG [TDWG] GUID subgroup spent several years working on and discussing the best practices for working with identifiers. The outputs of these activities can be seen on the GUID wiki of the TDWG community, http://wiki.tdwg.org/GUID/.

The outcomes of the TDWG GUID working group included:

- Recommending the adoption of the LSID Persistent Identifier technology for biodiversity informatics (but not to the exclusion of any other suitable identifier technology).
- The best practices when implementing LSIDs for biodiversity data, including the minimal resolution services that should be set up and the best way to define a naming scheme for identifiers.
- Recommending the use of standard vocabularies and schemas for formatting data, including TDWG standards, Dublin Core [DC] and FOAF [FOAF].
- Recommending the reuse of any existing, common identifiers and thus encouraging standardisation of biodiversity data.

Another recent, major activity was the GBIF [GBIF] LGTG task group (LSID and GUID Task Group). This group was primarily focused on undertaking a gap analysis of the technical infrastructure and capabilities of the biodiversity community. The result was a report to GBIF recommending actions needed to improve these issues. The report is available on the GBIF website (http://links.gbif.org/gbif_lgtg_report.pdf).

The recommendations for action by GBIF included the following:

- Taking a leadership role in driving the application of identifiers in the biodiversity community.
- Providing education and support for community members who are in the process of adopting Persistent Identifiers (e.g., this guide).
- Encouraging interconnected data, and reuse of existing Persistent Identifiers.
- Providing services for supporting the use of Persistent Identifiers in the biodiversity community.

b. Vocabularies

Vocabularies and ontologies are types of Knowledge Organisation System (KOS). A full discussion of the various systems which range from simple glossaries and dictionaries (i.e., flat vocabularies) through classification schemes, taxonomies, thesauri and ontologies is beyond the scope of this publication. The reader is referred to Hodge [Hodge2000] for an introduction. The terms vocabularies and ontologies, as here, are often used interchangeably although the term ontology has a specific meaning in the context of informatics:

> *"In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)." [Gruber2009].*

It is critical to define the vocabularies used within any community that is interested in sharing their data over the web. The TDWG community has carried out much work in this

area and now supports a range of biodiversity standards and schemas, including some for Taxon Names and Concepts, Specimens and Observations, Images and Multimedia, Natural Collections and Literature. A schema can be considered as a data exchange or message-passing format for the data described using particular vocabularies.

It is also very important to reuse, where appropriate, the vocabularies and schemas that other communities have developed, to aid interoperability and save reinventing the wheel. This is most likely to work where there are areas of common interest among multiple, varied communities and less likely in specialist areas of your community. There are a few generic vocabularies such as Dublin Core that could be used in nearly every case.

Because biodiversity informatics is a fairly specialised area of expertise, it is likely that a large proportion of the vocabularies and ontologies required for this domain will need to be developed within this community.

The following is a summary of the main current schemas, vocabularies and ontologies, both generic and biodiversity-specific, in use within the biodiversity informatics domain.

Generic Vocabularies

Dublin Core
A generic vocabulary used for basic metadata for any type of digital resource. This is a very commonly used, well accepted vocabulary and is a good one to use as a first step. More details can be found at http://dublincore.org.

Friend Of A Friend (FOAF)
A vocabulary used to describe people and their relationships. See http://foaf-project.org for more information.

Simple Knowledge Organization System (SKOS)
Provides an RDF schema for describing Knowledge Organisation Systems such as vocabularies and ontologies. For more information see http://www.w3.org/2004/02/skos/.

Biodiversity Vocabularies and Schemas

Taxonomic Concept Transfer Schema (TCS)
A schema used to describe and transfer taxonomic name and concept data. This includes information about nomenclatural name data, taxonomic concept data and relationships between taxon concepts. More information can be found at http://www.tdwg.org/standards/117/.

Access to Biological Collection Data (ABCD)
A schema to describe biological specimens and observation data. For more information see http://www.tdwg.org/standards/115/.

Structured Descriptive Data (SDD)
A schema to describe descriptive biodiversity data such as formal descriptions of taxa and dichotomous keys. For more information see http://www.tdwg.org/standards/116/.

Natural Collections Descriptions (NCD)
A schema for describing collections of natural history material. For more detail see http://www.tdwg.org/standards/312/.

Darwin Core (DwC)
A set of terms to facilitate the sharing of information about biological diversity. For more information see http://www.tdwg.org/standards/450/.

Multimedia Resources Metadata Schema (MRTG)
An emerging schema that allows description of multimedia resources. The work done by this group also includes useful discussions of schemas and RDF in general. For more information see http://www.keytonature.eu/wiki/MRTG.

Geospatial Vocabularies and Ontologies
OGC [OGC], Open Geospatial Consortium, is the major player in the geospatial arena. There are a range of standards and service specifications available through the OGC, the primary relevant standard being the Geography Markup Language (GML) [GMLES]. For more information see http://www.opengeospatial.org/standards/gml.

Consider the following points when deciding which schemas, and vocabularies to use:

- If you need to describe general data such as **creation date, owner, title**, then the use of the Dublin Core schema is a good idea (in widespread use).
- Do you need to describe core biodiversity data such as **Specimens, Taxon Names** and **Observations**? Use the TDWG schemas as a starting point.
- For other domains, research the area of interest to locate any existing working groups and standards organisations. It is likely that someone has already had a similar need and collaboration with this group and reuse of existing standards will be beneficial.
- If no existing standards can be located, custom schemas can be created for the task at hand.
- Decide on the scope of the task to determine whether a generic schema or a more specific schema is appropriate (depends on the amount of detail available for each digital resource).

Using the Dublin Core and Darwin Core vocabularies with our example, a possible XML representation is shown in Figure 8.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:dwc=http://rs.tdwg.org/dwc/terms/
      xmlns:dc="http://purl.org/dc/elements/1.1/">
   <rdf:Description rdf:about=" http://www.example.org/taxon/ 7217D220-836A-11DF-
8395-0800200C9A66">
      <rdfs:type rdf:resource="http://rs.tdwg.org/dwc/terms/Taxon"/>
      <dwc:basisOfRecord rdf:resource="http://rs.tdwg.org/dwc/dwctype/Taxon"/>
      <dc:modified>2007-05-04T18:13:51.0Z</dc:modified>
      <dc:language>en</dc:language>
      <dwc:scientificNameID rdf:resource="http://www.example.org/taxon/
      7217D220-836A-11DF-8395-0800200C9A66" />
      <dwc:acceptedNameUsageID rdf:resource="http://www.example.org/taxon/
      7217D220-836A-11DF-8395-0800200C9A66"/>
      <dwc:originalNameUsageID rdf:resource="http://www.example.org/taxon/
      7217D220-836A-11DF-8395-0800200C9A66"/>
      <dwc:scientificName>Amanita alba Lam.</dwc:scientificName>
      <dwc:acceptedNameUsage>Amanita alba Lam.</dwc:acceptedNameUsage>
      <dwc:parentNameUsage>Amanita alba Lam.</dwc:parentNameUsage>
      <dwc:originalNameUsage>Amanita alba Lam.</dwc:originalNameUsage>
      <dwc:nameAccordingTo>Smith, T.  Fungi of Earth</dwc:nameAccordingTo>
      <dwc:namePublishedIn>Encycl. Méth. Bot .1(1): 107
      (1783)</dwc:namePublishedIn>
      <dwc:higherClassification>Fungi, Basidiomycota, Agaricomycetes,
```

```
      Agaricomycetidae, Agaricales, Amanitaceae</dwc:higherClassification>
      <dwc:kingdom>Fungi</dwc:kingdom>
      <dwc:phylum>Basidiomycota</dwc:phylum>
      <dwc:class>Agaricomycetes</dwc:class>
      <dwc:order>Agaricales</dwc:order>
      <dwc:family>Amanitaceae</dwc:family>
      <dwc:genus>Amanita</dwc:genus>
      <dwc:specificEpithet>alba</dwc:specificEpithet>
      <dwc:scientificNameAuthorship>Lam.</dwc:scientificNameAuthorship>
      <dwc:taxonRank>species</dwc:taxonRank>
      <dwc:nomenclaturalCode>ICBN</dwc:nomenclaturalCode>
      <dwc:taxonomicStatus>accepted</dwc:taxonomicStatus>
   </rdf:Description>
</rdf:RDF>
```

Figure 10. Example marked up using some available vocabularies (Darwin Core and Dublin Core)

c. Reuse of Persistent Identifiers

Multiple applicable identifiers for the same resource

It is often the case that two people have the same data for a very similar resource (which may well be exactly the same resource). In this case it would be very useful and beneficial for those people to use the same identifier for that resource. Wouldn't it make the lives of the people using that data so much easier if they knew two different instances of data were referring to the same physical resource? It would then be possible to link a variety of data, e.g., Observation records, Taxonomic records and Specimen records, using that same identifier. For this reason it is quite important to reuse Persistent Identifiers wherever possible.

However, it is often difficult to determine if two different instances of data are referring to the same "real life" thing. By "real life" thing we mean the material resource, abstract concept, or physical resource that exists in the real world that we are trying to describe. There are some basic principles to follow that will help this issue.

- ✓ Ensure that there are no existing Persistent Identifiers that refer to the same "real life" resource before generating a new one.
- ✓ If an existing resource exists that is an exact match to the resource you are describing then use the identifier of that existing resource.
- ✓ If a similar resource exists, link to that resource, via data or metadata links, to allow people to see that they are related. This can be done using various ontological constructs such as **owl:sameAs**, **rdfs:seeAlso** or **skos:related**.

Use of the owl:sameAs construct is shown below in Figure 9.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
                 xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
                 xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:owl="http://www.w3.org/2002/07/owl#">
<rdf:Description rdf:about=" http://www.example.org/taxon/ 7217D220-836A-11DF-
8395-0800200C9A66">
            <dc:modified>2007-05-04T18:13:51.0Z</dc:modified>
            <dc:language>en</dc:language>
            <dwc:basisOfRecord>Taxon</dwc:basisOfRecord>
            <dwc:scientificNameID rdf:resource="http://www.example.org/taxon/
7217D220-836A-11DF-8395-0800200C9A66"/>
```

```
      <owl:sameAs rdf:resource="urn:lsid:indexfungorum.org:names:494891"/>
…
      </rdf:Description>
</rdf:RDF>
```

Figure 11. An example showing how to refer to a strongly related resource.

Citing the correct identifiers

A consideration when providing data over the web is how to reference any original data or related data that your data are based on. This can be done as described in the previous section using constructs such as owl:sameAs and rdfs:seeAlso.
Another useful approach here is to be more explicit and specify the original data as the source of your own data. One way to do this is using the Dublin Core term **dc:source**.
Our example is shown in Figure 10.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
                 xmlns:dwc=http://rs.tdwg.org/dwc/terms/
                 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.example.org/taxon/ 7217D220-836A-11DF-
8395-0800200C9A66">
            <dc:modified>2007-05-04T18:13:51.0Z</dc:modified>
            <dc:language>en</dc:language>
            <dwc:basisOfRecord>Taxon</dwc:basisOfRecord>
            <dwc:scientificNameID rdf:resource="http://www.example.org/taxon/
7217D220-836A-11DF-8395-0800200C9A66"/>

      <dc:source rdf:resource="urn:lsid:indexfungorum.org:names:494891"/>
      …
      </rdf:Description>
```

Figure 12. An example showing how to refer to a source resource.

When considering which data provider to use for identifying a resource, it is useful to consider if any authorities are used higher up in the "chain" of ontology classes. For example, if a user decides to use taxon concepts from provider X, they have to use provider X's choice of taxon name provider - they cannot enforce the use of a different one. Making explicit the chain of dependencies is obviously useful, so it may often be necessary for concept providers to make clear who their dependent authorities are.

d. Scope of your data

If you have decided that you, or your institute, are the appropriate provider for the data you are exposing, you need to consider how to divide up your data and which digital resources you will apply Persistent Identifiers to. In particular, you should:

- ✓ Decide what forms a coherent dataset. It may be just a subset of all of the data that you want to apply identifiers to. If so, you need to assess which subsets of data have their own distinct identity, i.e., the subset that could potentially move to a separate system in the future.
- ✓ Decide which resources (data classes) are to be identified within that "dataset". Here you also need to assess the delimitations of your own authority, i.e., you may choose to use identifiers for remotely managed data resources as well as supplying identifiers for resources which you yourself manage.

e. Transferring datasets and their identifiers

When considering the infrastructure for providing data via the Internet, some key considerations are:

- Host organisation (who is going to provide services for supplying the data).
- Context name (what project name or organisation domain name will define how to access the information).
- Redundancy (how do we ensure the data are always available, 24/7 and globally).

These considerations will obviously change over time due to factors such as organisation name changes, domain name changes, and funding influences. It is important, therefore, to consider how these issues could be avoided for your data in the future.

Some simple rules to avoid these issues:

- ✓ Use an institute-independent domain name. For example, for an international fungi names index that is driven by an organisation called Example Org, it would be better to use a domain name such as http://fungi.org rather than the domain name http://example.org/fungi
- ✓ Develop a backup plan. You need a host organisation or funded body to be ready to take over hosting of your data if your current institute is no longer able to do so.
- ✓ Keep your data and data structures independent of other areas of your organisation, i.e., if the dataset in question is independent of other organisational aspects then the dataset can be transferred to another institute with reasonably low dependencies and complications.

f. Linked Data

Linked Data is a relatively new approach to data on the web. The idea is based on the fact that every digital resource is de-referenceable using standard web protocols (HTTP), and every resource links to other resources on the web as much as possible.

The traditional Internet is based on documents that are retrievable via global interconnected networks. These documents are located at URLs (Uniform Resource Locators). The problem with this is that documents tend to be very unstructured HTML, and are only really useful for a person using a web browser interface. Linked Data is an approach to put data on the web in a similar way, but in a structured form suitable for machine processing. In a way, this is like making the Internet one big interconnected database.

As mentioned in section 1b, we need to address the issue of whether an identifier refers to a "real life" resource or to a digital representation of that resource. The normal mode of operation on the web is to retrieve a digital representation of a resource, typically a web page. Another approach to solving this issue is to use the Linked Data approach of redirection. The redirection mechanism is a standard HTTP method, using HTTP response type 303. HTTP 303 indicates a redirection to the data for the resource in question. In this way it is possible to use an identifier that points to a non-existent location on the web (URI), where that location then redirects to a location that returns the data for the resource in question. For example, if we are discussing the person "John X. Smith", using identifier http://example.org/person/JohnXSmith, then it is not really possible to send the real person, John, over the Internet when someone tries to resolve that identifier. Therefore, the identifier http://example.org/person/JohnXSmith simply *refers* to the "real" person, and when that identifier is resolved, a redirect to the data for that person is returned to the caller. The caller then retrieves the data from the redirected location.

So for our specimen example, we could do something like that shown in Figure 11.

Physical Specimen Identifier:
http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E
(when resolved, redirects to a digital representation)

Digital Specimen Identifier:
http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E.rdf
(resolves to an RDF representation of the specimen)

Figure 13. Example of various URI identifiers denoting different representations of a resource.

g. Web services

Web services are a key component of data provision over the Web. To provide data for Persistent Identifiers, it is necessary to host web services that return that data to the caller.

Different services are required for different types of Persistent Identifier. The following list summarises the web service requirements for the different types of identifier.

- URI
  - HTTP web service endpoint
- LSID
  - LSID authority service
  - LSID data service
  - LSID metadata service
- PURL
  - HTTP web service endpoint
- DOI
  - DOI web services are provided by DOI.org [DOI]

The degree of required web service technology ranges from a simple HTTP web service, to a more complicated SOAP [SOAP] oriented service. To set up an LSID service you are not required to develop a SOAP service, but WSDL [WSDL] documents must be provided so an understanding of SOAP is useful.

Simple HTTP web service

Any web developer should be able to set up a web service endpoint that enables you to provide documents representing your digital resource over the web.

Several possible ways to achieve this include:

- Set up text documents for each digital resource. Use the Persistent Identifier of that resource for the name of the document. For example, for the previous digital specimen identifier, a URL like the following would work:

http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E.rdf

- Set up some simple web code (e.g. ASP or PHP scripting languages) to access the database containing the digital resources, and return that data as the response to the resolution of that identifier. In this case it may be fine to just have the identifier as the final part of the URL, e.g.

http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E

- If you are following the Linked Data approach then the preferred way to handle this is to handle a call to the URL with no file extension and redirect to the URL with the file extension (following the idea that the non-extension URL points to the

"conceptual" resource and the URL with file extension is the document representation of that resource). E.g.,

http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E redirects to http://www.example.org/specimen/10FC9784-B30F-48ED-8DB5-FF65A2A9934E.rdf

SOAP services for LSID resolution

LSID resolution requires three steps for complete resolution. 1) the client uses DNS to locate the LSID authority endpoint; 2) the client makes a web call to the server to determine the LSID functions that authority supports; 3) the client makes a call to one of those specific functions to get the metadata for that Persistent Identifier (LSID).

It is possible, although not recommended, to support all three stages of this resolution using basic web functionality. This can be achieved by, at a minimum, doing the following:

1. Set up a basic HTTP web service at a web location of your choice. For our example we will use a service located at http://www.example.org/authority/
2. Register an SRV record for your LSID internet domain name. For our example www.example.org, we would need to register the SRV record for www.example.org and point it to our domain at http://www.example.org/
3. Set up the authority web service to handle the minimum LSID calls. For our example this includes calls to:
    a. http://www.example.org/authority/ - this call attempts to get the WSDL [WSDL] for the services the authority supports. At this point, static WSDL files can be used to specify that only the basic HTTP GET service is supported at the location http://www.example.org/authority/metadata/
    b. http://www.example.org/authority/metadata/?lsid=xxx – this call is requesting the actual metadata for the resource, so at this point we need to return the actual document representing the identifier.

LSID Hosting Services

Sometimes, even the simple version described above will seem too much work and require too much expertise. As recommended in the LGTG report (mentioned in section 3) hosting services are intended to be set up at some point to assist people with the LSID resolution process. GBIF [GBIF] are one of the institutes intending to support these services.

In this situation, one of the processes likely to be used is as follows:

1. GBIF provides a domain name for LSID identifiers, e.g., http://lsid.gbif.org. In this way GBIF can handle the DNS resolution step for the LSIDs.
2. GBIF would also provide LSID assigning services so that clients can request new identifiers as required. Namespacing of the identifiers will be important here and one possible way to do this would be to have the "namespace" component of the LSID assigned for a specific client. For our example, we could request GBIF to give us the LSID namespace "Example", and then our LSIDs would take the format urn:lsid:lsid.gbif.org:example:XXX. A GBIF service would then have the functionality to produce a new identifier based on this format and allocate this identifier to us.
3. The third step of LSID resolution is hosting of the actual data and metadata content. The content could either be hosted by the client or by GBIF. If hosted by the client, all they would need to do is provide a web service endpoint that could be used as the base for redirection at that stage of LSID resolution. If hosted by GBIF, then the client would be required to provide their data/metadata to GBIF, and updates of the data as they change.

5. Checklist for implementing Persistent Identifiers

- ✓ Pick a context name for scoping your identifiers (section 1e).
- ✓ Assess the robustness and opacity of the format of your identifier (section 1e).
- ✓ Assess any current identifiers for your digital resources and decide if they will remain unchanged into perpetuity and hence be usable as part of the external Persistent Identifiers (section 1d).
- ✓ Pick a Persistent Identifier format, e.g. LSID, URI (section 2).
- ✓ Define the management processes for connecting your digital resources / database records to the Persistent Identifiers (section 3).
- ✓ Define which type of resources you will be providing Persistent Identifiers for (section 3a).
- ✓ Decide how to handle data changes (section 3b).
- ✓ Define management processes for handling changes to your data and versioning of your resources (section 3c).
- ✓ Decide on vocabularies and schema(s) to represent your data (section 4b).
- ✓ Ensure there are no existing Persistent Identifiers that you could reuse, before creating new ones (section 4c).
- ✓ Decide on what subset (or perhaps all) of your data you want to make available. Keep in mind this subset could be relocated at some point in the future (Section 4d).
- ✓ Decide if you would like to follow Linked Data practices (section 4f).
- ✓ Decide on the web services that will be required for your Persistent Identifiers. At a minimum this should include basic HTTP resolution of the identifiers, or use 3rd party identifier hosting services if this is preferred (section 4g).

6. References

[ABCD] Access to Biological Collection Data. http://www.tdwg.org/standards/115/.

[Baskauf2010] Baskauf, S. J., 2010. Organization of occurrence-related biodiversity resources based on the process of their creation and the role of individual organisms as resource relationship. Biodiversity Informatics 7:17-44.

[DC] DC. Dublin Core Metadata Initiative. http://dublincore.org.

[DwC] Darwin Core. http://www.tdwg.org/standards/450/.

[DOI] DOI Foundation. http://www.doi.org/.

[FOAF] Friend Of A Friend ontology. http://www.foaf-project.org/.

[GBIF] Global Biodiversity Information Facility. http://www.gbif.org.

[GMLES] Geography Markup Language encoding standard. http://www.opengeospatial.org/standards/gml.

[Gruber2009] Ontology by Tom Gruber in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009. http://tomgruber.org/writing/ontology-definition-2007.htm.

[Güntsch] Güntsch, A., Berendsohn, W., Geoffroy, M. Versioning and the use of GUIDs for PESI. http://www.eu-nomen.eu/pesi/index.php?option=com_remository&Itemid=56&func=fileinfo&id=767.

[Hodge2000] Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. http://www.clir.org/pubs/reports/pub91/contents.html.

[IEFTURI] IETF URI specification. http://tools.ietf.org/html/rfc1630.

[IEFTUUID] IETF UUID specification. http://www.ietf.org/rfc/rfc4122.txt.

[ISBN] ISBN. International Standard Book Number. http://www.isbn-international.org/.

[JacobsWalsh] Jacobs, I. and N. Walsh (Eds.), 2004. Architecture of the World Wide Web, Volume One (W3C Recommendation 15 December 2004. http://www.w3.org/TR/webarch/.

[MRTG] Multimedia Resources Task Group. http://www.keytonature.eu/wiki/MRTG.

[NCD] Natural Collections Descriptions. http://www.tdwg.org/standards/312/.

[OCLC] Online Computer Library Center. http://www.oclc.org/.

[OGC] Open Geospatial Consortium. http://www.opengeospatial.org/.

[SDD] Structured Descriptive Data. http://www.tdwg.org/standards/116/.

[SKOS] Simple Knowledge Organization System. http://www.w3.org/2004/02/skos/.

[SOAP] Simple Object Access Protocol. http://www.w3.org/standards/techs/soap.

[TCS] Taxonomic Concept Transfer Schema. http://www.tdwg.org/standards/117/.

[TDWG] Taxonomic Databases Working Group. http://www.tdwg.org.

[URI] Uniform Resource Identifier. http://www.w3.org/Addressing/URL/uri-spec.html.

[UUID] Universally Unique Identifier. http://en.wikipedia.org/wiki/Universally_Unique_Identifier.

[W3C] W3C. http://www.w3.org/.

[W3CTech] W3C Technical Architecture Group (TAG) 2005. httpRange-14: What is the range of the HTTP dereference function? http://www.w3.org/2001/tag/group/track/issues/14.

[WSDL] Web Services Description Language. http://www.w3.org/TR/wsdl.

7. Acronyms

ABCD	Access to Biological Collection Data
ASP	Active Server Pages
CRUD	Create Read Update Delete
DNS	Domain Name System
DOI	Digital Object Identifier
DwC	Darwin Core
FOAF	Friend of a Friend
GBIF	Global Biodiversity Information Facility
GML	Geography Markup Language
GUID	Globally Unique Identifier
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ISBN	International Standard Book Number
KOS	Knowledge Organisation System
LSID	Life Science Identifier
MRTG	Multimedia Resources Task Group
OCLC	Online Computer Library Center
OGC	Open Geospatial Consortium
PHP	PHP: Hypertext Preprocessor
PURL	Persistent Uniform Resource Identifier
RDF	Resource Description Framework
SDD	Structured Descriptive Data
SKOS	Simple Knowledge Organisation System
SOAP	Simple Object Access Protocol
SRV	Service Record
SVN	Switched Virtual Network
TCS	Taxonomic Concept Schema
TDWG	Taxonomic Databases Working Group
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
UUID	Universally Unique Identifier
WSDL	Web Services Description Language
XML	Extensible Markup Language