Identifiers for the Life Sciences

A Primer for Biologists

What is a "GUID"?

A Globally Unique IDentifier (GUID; rhymes with "squid") is a unique identification code that is applied to "something" that can be unambiguously referenced. GUIDs have their greatest value in computer databases. The "something" that a GUID references can be a database record, a physical object such as a specimen in a Museum, an abstract construct such as a collecting event, or taxonomic concept that is represented by a database record.

What is an "LSID"?

Life Science IDentifiers (LSIDs) are an example of a GUID. They were developed by IBM specifically for the life-sciences community, and therefore have features that support biological datasets. For example, unlike DOIs (a GUID used to identify publications) there is no fee to issue LSIDs. LSIDs are also relatively easy to apply to existing systems. LSIDs follow the template:
urn:lsid:<authority>:<namespace>:<ObjectID>:[version]
An example of an LSID is: "**urn:lsid:ncbi.nlm.nig.gov:GenBank:T48601:2**".

Why are LSIDs valuable?

LSIDs are valuable because they are unambiguous. In biology, we have information about 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 catalog number (e.g., "BPBM 37615"), or a publication by a citation (e.g., "Baldwin & Smith, 1998"). But these identifiers are ambiguous: there are two specimens with the catalog number "BPBM 37615" (one is a fish, the other a mollusk), 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's difficult for a computer to correctly understand the context.

Why are LSIDs any less ambiguous than the identifiers we're used to?

As with all GUIDs, the two qualities of LSIDs that make them useful are **uniqueness**, and **persistence**. The same LSID cannot refer to more than one physical, abstract or electronic object, because if it did, it would not be *globally* unique and it would not be a GUID. LSIDs are also *persistent*, in that any given LSID will *always* refer to the same thing – now, and in perpetuity. Once assigned and established, LSIDs never change. A particular object may get more than one LSID (either intentionally or by accident), but both LSIDs remain persistent, and cross-reference each other once the equivalency is discovered. The identifiers we're used to seeing – people names, taxonomic names, place names, catalog numbers – are neither globally unique, nor (in many cases) persistent.

Can't we just enhance our existing identifiers to make them unique and persistent?

We could extend the identifiers we're used to seeing to try to make them unique. For example, the catalog number "BPBM 37615" could be qualified as "BPBM-I 37615" for the ichthyollogical specimen and "BPBM-M 37615" for the malacological specimen. As long as no other institution used the abbreviation “BPBM” for their data, the numbers would be globally unique. This is the approach that GBIF initially took to uniquely identify specimens – that is to use the three-part identifier system of the DarwinCore standard “InstitutionCode+CollectionCode+CatalogNumber”. With nearly 100 million records accessible through the GBIF data portal, however, non-uniqueness and lack of persistence were significant enough that GBIF is looking to implement LSIDs. While it is theoretically possible to implement uniqueness with museum specimen catalog numbers, it is not so easy in other realms of biologically-relevant data such as taxon names, peoples names, publication citations, localities, etc. Persistence remains a problem as specimens often change institutions, and receive new catalogue numbers.

What good are LSIDs if they are “unfriendly” to read?

LSIDs are not like catalogue numbers or publication citations. LSIDs are intended for efficient computer to computer communication. LSIDs should be transparent to most users; they are designed to be generated and interpreted by computers; not people – just as your computer’s MAC Address (another kind of GUID) is used whenever you connect to a network. Most users never see a MAC address, but they are fundamental to the operation of computer networks. Each part of an LSID serves a specific purpose, both for ensuring uniqueness, and for allowing *resolution* of the LSID.

What is “Resolution”, and why is it important?

An identifier is useful if it is resolvable; that is – how the full details of the object referenced by an identifier can be accessed. This is similar to a web address URL. The text “http://www.gbif.org” by itself isn’t very useful in itself. But if you enter that text into the address bar of your web browser, you gain access to far more information. LSIDs are self-resolving identifiers as they act as both an Internet address *and* as an object identifier.

Why should I support the implementation of LSIDs for biological databases?

Although the internet has provided unprecedented access to biological data, including images, specimen data, DNS sequences, publications, taxonomy, and more…broad implementation of LSIDs would dramatically improve the ease of access to this information. LSIDs 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.. Indeed, LSIDs represent a critical enabling component of an online “Encyclopedia of Life”.

Further Reading:

An informative paper on LSIDs and RDF: http://jbi.nhm.ku.edu/index.php/jbi/article/view/25
LSID Home Page: http://lsid.sourceforge.net/
GBIF GUID Wiki: http://wiki.gbif.org/guidwiki/wikka.php?wakka=HomePage