"An Unambiguous Identification of Life?"

Lee Belbin
lee@tdwg.org
(Written as a Press Release after GUID-1 meeting at NESCent, February 2007)

Put ten taxonomists in a room to describe a new species and you will probably get more than ten descriptions and names. Biology is complex. Imagine then what it might be like to try to trace changes in the naming and misnaming of a species through time. Identifying the status of a species name can be a frustrating exercise even for experts.

Consider then the millions of observations of hundreds of thousands of species stored in thousands of databases around the world. If we are to conserve the planet's biodiversity, we depend on such data to make quality management decisions. How can informed decisions be based on data where the same species is given five different names across different databases, or one name refers to five different species?

A meeting of 30 international experts in bioinformatics and computing was hosted by the USA' National Evolutionary Synthesis Centre (NESCent: www.nescent.org) to attempt to address this problem. The workshop was organized by the Taxonomic Database Working Group (www.tdwg.org) of the International Union of Biological Sciences and the Global Biodiversity Information Facility (www.gbif.org). It is GBIF's role to provide open access to the world's biological data and it is TDWG's role to provide the standards to make that possible.

The workshop was seeking a system of identifiers for data records that relate, for example, to the naming or occurrence of organisms. This identifier needs to be "globally unique" for each data record. Thus such identifiers are called globally unique identifiers, or GUIDs.

By using GUIDs, the current ambiguity in databases could be greatly reduced over time. A well-designed GUID system could also provide the basis for valuable additional services to those seeking biological information.

The meeting adopted a GUID technology known as Life Science Identifiers (LSID). LSIDs were developed by the Object Management Group (www.omg.org), an open-membership, not-for-profit consortium that produces and maintains computer industry specifications that enable data integration.

LSIDs serve as uniform resource tags that uniquely identify a given data object on the Internet. LSIDs provide standard mechanisms for accessing data and metadata (descriptive information) for the objects they identify. Experts say that they will form anchor-points for a range of layered information services relating to these objects, such as access to further data resources in a wide range of formats. LSIDs can in this way serve as a stepping-stone to integrating biodiversity data.

For example, an LSID issuing authority such as a museum could generate an LSID for a type specimen (the specimen that a species is named from) and link services that enable a scientist on the Internet to see images of the specimen, a set of keys to identifying the species within the family it belongs to, and observations in space and time of that species.

LSIDs are comprised of six components, for example-

urn:lsid:ncbi.nlm.nih.gov:GenBank:T48601:2

where “urn” and “lsid” form a mandatory preface for LSID data; “ncbi.nlm.nig.gov” is the identifier of the organization that assigned the LSID to the data; “GenBank” identifies a class of data objects offered by the organization; “T48601” is the name of the data object; and “2” is an optional version number.

IBM has developed tools that help providers serve LSIDs on the Internet and for clients to resolve those LSIDs to access resources (see http://www-128.ibm.com/developerworks/webservices/library/os-lsid2/#resources).