Towards an LSID Policy for the Australasian biodiversity federation

Recommendations

Canberra LSID Workshop
ANBG, 2 – 4 April 2007

Biodiversity
Information
Standards
TDWG

International Union for Biological Sciences
Biodiversity Information Standards (TDWG)
www.tdwg.org

25th May 2007

Compiled by Alex Chapman, Greg Whitbread, Ben Richardson & Lee Belbin
from the Minutes of the LSID Policy Workshop
hosted by the Australian National Botanic Gardens
Canberra, between the 2nd – 4th April 2007

Table of Contents

Meeting participants

- Bill Barker (State Herbarium of South Australia; Department of Environment and Heritage; Adelaide)
- Lee Belbin (TDWG Infrastructure Project)
- Alex Chapman (Western Australian Herbarium; Department of Environment and Conservation; Perth)
- Paul Coddington (South Australian Partnership for Advanced Computing; The University of Adelaide; Adelaide)
- Jim Croft (Australian National Herbarium; Centre for Plant Biodiversity Research; Canberra)
- Paul Flemons (Australian Museum; Department of Communications, Information Technology and the Arts; Sydney)
- Piers Higgs (Gaia Resources - representing Western Australian Museum; Department of Culture & the Arts; Perth)
- Robyn Lawrence (Australian Biological Resources Study; Canberra)
- Robert Morris (South Australian Museum; Adelaide)
- Peter Neish (National Herbarium of Victoria; Royal Botanic Gardens Melbourne; Melbourne)
- Laurence Paine (Manager Information Technology; Department of Tourism, Arts and the Environment; Hobart)
- Ricardo Pereira (TDWG Infrastructure Project; Campinas; Brazil)
- Robert Raven (Queensland Museum; Department of Education, Training and the Arts)
- Kevin Richards (Allan Herbarium; Manaaki Whenua - Landcare Research)
- Ben Richardson (Western Australian Herbarium; Department of Environment and Conservation; Perth)
- Dan Rosauer (Department of the Environment and Water Resources)
- Steve Shattuck (Australian Biological Resources Study; Department of the Environment and Water Resources; Canberra)
- Cameron Slatyer (Australian Biological Resources Study; Department of the Environment and Water Resources; Canberra)
- Kevin Thiele (Western Australian Herbarium; Department of Environment and Conservation; Perth)
- Jeff Tranter (Environmental Resources Information Network; Department of the Environment and Water Resources; Canberra)
- Greg Whitbread (Australian National Herbarium; Centre for Plant Biodiversity Research; Canberra)
- Aaron Wilton (Allan Herbarium; Manaaki Whenua - Landcare Research; Christchurch)

Executive summary

A three-day workshop developed recommendations for a policy to apply of Life Science Identifiers (LSIDs) within the Australasian biodiversity federation. The workshop also developed a roadmap for LSID implementation by and for local data providers and biodiversity informatics networks.

Primary reference was made to the pioneering work of Biodiversity Information Standards (TDWG), an affiliate of the IUBS and collaborator with the Global Biodiversity Information Facility (GBIF), in developing data standards for the global bioinformatics community.

The workshop generated recommendations for the delegation of responsibility for allocation, persistence and resolution of LSIDs and drafted a roadmap and workplan for the implementation of this technology within the Australasian biodiversity federation.

The Roadmap aims to actively position early adopters to deliver content to the 'Atlas of Living Australia' and concept projects in time for the next TDWG meeting in September 2007.

Aims and context of the workshop

A policy for LSID implementation, and a road-map for the integration of this technology, within the Australasian biodiversity federation

Uniquely identified data objects are fundamental to the successful implementation of our biodiversity information networks.

Globally Unique Identifiers (GUIDs) are the key to delivering on many of the basic requirements expected of distributed biodiversty information systems. Custodianship; discovery of duplication; effective validation procedures; data update, indexing and caching services; verification of derived product; tracking of annotations etc. Much depends on our ability to return to individual records.

GUIDs provide these benefits by tagging data objects with standardised information. GUIDs can then help locate data objects and provide access to standardised metadata about them.

TDWG is recommending Life Science Identifiers (LSIDs) as the standard GUID technology for the biodiversity informatics community. To this end the TDWG GUID subgroup has produced an LSID Applicability Statement, a Roadmap for LSID implementation and a range of documentation to support LSID deployment.

This workshop was supported by the TDWG Infrastructure Project (TIP). The Project considered that a policy for the application of LSIDs within the Australasian biodiversity federation would provide valuable feedback to the TDWG GUID Group and also inform international LSID implementation projects.

Building on the TDWG LSID Applicability Statement and LSID Roadmap the aims of this workshop are to:

- clearly describe the technology;
- establish key principles for LSID integration;
- prioritize LSID assignment;
- make recommendations as to the delegation of responsibility for allocation, persistence and resolution of GUIDs and;
- document a workplan and timeline for the implementation of GUID technology.

A general introduction

The importance of Life Science Identifiers (LSIDs) for the biodiversity community

Preamble

The World Wide Web revolutionized the way in which we broadcast and access digital information. The next revolution will come from new technologies that allow us to synthesize, manage and integrate the web's vast quantities of information - the so-called semantic web. These technologies will evolve the Web from an electronic notice board into a truly connected, dynamic and flexible knowledge collaboration.

Globally Unique Identifiers (GUIDs) are a critical building block in this new revolution. GUIDs are small, standardised tags attached to digital objects. Database records, documents, images, names, or any other object that will be electronically shared may be uniquely identified and described using GUIDs. Such tagged objects may then be integrated and combined with other information to bring new insights and allow new knowledge discovery. GUIDs may also function as calling-home cards - an GUID on a digital object can be used to find, attribute and identify its original owner, no matter where the object travels on the web.

One type of GUID – Life Science Identifiers (LSIDs) – have been chosen as an agreed standard in the global biodiversity community, supported by the Global Biodiversity Information Facility and other global and Australasian biodiversity peak bodies. LSIDs are decentralized, collaborative and free. Individual institutions - the custodians of data - manage the deployment of LSIDs for their own shareable data assets rather than relying on a centralized issuing authority. This provides LSIDs with much-needed flexibility in the fast-evolving web.

The risks of supporting LSID technologies are inherently low. The only costs in deploying an LSID service are the time necessary for a data manager to establish the service (typically days to weeks). LSIDs are simple, lightweight, and promote rather than impede normal data management workflow.

By contrast, the risks of not supporting LSID technologies are high. Institutions that fail to deploy LSID services will become increasingly disconnected from the emerging web of knowledge, and will be unable to share their data effectively with the world or share the world's data for better decision support.

A joint meeting of information professionals from the combined Australasian herbarium and museum communities has recommended the adoption and deployment of LSIDs by all major Australian biological collections and their host organizations and institutions. The recommendation is endorsed by the Global Biodiversity Information Facility and Biodiversity Information Standards (TDWG) organization. Your institution is urged to support their deployment using the attached implementation plan and strategy.

Discussion

Life Science Identifiers (LSIDs) are small, lightweight, globally unique digital tags that can be attached to any digital object. Objects that carry an LSID can be uniquely identified and attributed, even when the object is shared, merged into other objects, or moved from its local context. Three properties of LSIDs contribute to their flexibility and utility.

1. Think global, then everything's local

Databases use unique identifiers to manage records. For example, specimen records in a specimen database are often identified using accession numbers, and names databases generally assign NameIDs for each name. The uniqueness of the identifier allows a each record in the database to be unambiguously identified – clearly important in managing, using and maintaining the data. Identifiers are however almost always local to the particular database in which they are assigned. If data from two or more databases are combined in some way, uniqueness of the combined identifiers cannot be guaranteed. The outcome of this is that it may then no longer be possible to unambiguously refer to any given record, and all records will need to be cumbersomely renumbered after which many broken processes and links will need to be fixed.

Imagine if every database record in every database in the world had an identifier that could be guaranteed to be globally unique. Then when two databases are merged or share data it would be immediately possible to use the newly accessible records with no possibility of identity clashes or ambiguity.

LSIDs provide just such a way of tagging records in databases with globally unique identifiers. An LSID is a string of text of the form

urn:lsid:authority:namespace:identifier. An example would be

urn:lsid:herbarium.PERTH.lsid.org.au:specimen:02344759

If the authority (herbarium.PERTH.lsid.org.au) is a unique address, and the authority can guarantee that the record 02344759 in its specimen table is unique, then the LSID is globally unique and can identify that specific record. LSIDs may be applied to any type of digital object that may at some time be shared, not just records in a database. LSIDs in exactly the same form can be applied to specimen records, names, descriptions, characters and character states, documents, images, spreadsheets, phylogenies – to any digital object of any kind.

Applying LSIDs to objects is cost-free, and LSIDs are assigned by custodians of data with no requirement for a centralized issuing authority. For these reasons, LSIDs have been adopted by the international biodiversity community as the principal system of globally unique identifiers for use in the life sciences domain. LSIDs are seen as an enabling technology for the next generation of web applications, processes and operations.

2. Have calling card, will travel

LSIDs are more than just unique identifiers for records and other digital objects. They also act as calling-home cards for the objects they are attached to. This means that objects with LSIDs can never get lost on the Internet, and can always be ascribed to their custodian or owner using standard protocols.

Consider a database (the client) which aggregates records from several source databases. The owner of the client database may need to query the source databases at intervals for updates to their records. To do this the client would need to maintain systems for identifying each record in its source database, and for querying each source database for the updates. The updating would almost certainly be a cumbersome and expensive operation.

If, however, the records carry LSIDs and the source databases establish simple resolving services, a straightforward mechanism for updates can be established. Part of an LSID is the address (eg. herbarium.PERTH.lsid.org.au) of the authority which maintains the resolving service of the source database. Free web tools are available which will accept an LSID and find from this address the source's LSID resolver. The tool sends the LSID as a query to the source, which recovers from it the pointer to the original record in its database (e.g. specimen:02344759). The resolving service then returns information about the original record in a standard format. The returned information will normally include essential (meta-)data about the record, and this can be used by the client to update its copy of the record.

If all records carry LSIDs, one process attached to the client's database can be used to update records from all sources. In addition, one process at the client databases can be used to supply updates for all clients. Substantial time and cost savings are available at source and client ends using LSID technology.

3. Carry meaning, not just data

Over time, the ability of LSIDs to recover information about digital objects from their custodians will establish the true power of LSIDs, and play a part in the evolution of the World Wide Web into the Semantic Web – a flexible and intelligent web of knowledge. The key to the power of LSIDs is that the information returned when the LSID of a digital object is queried can be made meaningful to machines as well as humans.

Consider, as an example, Google Images. When this was new a few years ago it was considered pretty cool. But it is simply an early and somewhat primitive example of a data aggregator that is suffering from the lack of LSIDs.

Google Images is powered by web robots which trawl the web for image objects embedded in web pages. When an image is found, the robot returns to Google a thumbnail of the image and an extract of the html page text that surrounds it. From this text, Google Images makes a guess at the meaning of the image – is it an image of Copacabana Beach or of a funnel-web spider? The thumbnail image, a link to the original image and the inferred meaning is then databased ready for querying. The weak link is the inference part – these days any query using Google Images will return some images that correctly match the query but many are wrong – a funnel-web spider image returned from a query about Copacabana Beach is ample evidence of failed inference.

If images are progressively tagged with LSIDs, it will become possible to build vastly more accurate inference engines. If the funnel-web spider image is tagged with an LSID it will be possible to directly query the original custodian of the image to ask for information about it, the metadata. Such a query will return tagged, machine-readable, information using standardized and well-structured formats. One tag may say “This image is of an organism” while another may say “The name of the organism is *Atrax robustus*”. Immediately, an inference engine like Google Images will be able to accurately identify the image, because it has real information from the custodian of the image rather than simply the context of the image on its page.

A system of LSIDs becomes more powerful when LSIDs point to other LSIDs. For example, if the name of the funnel-web spider changes, inference becomes more difficult; it will be hard for a machine to determine that the name has changed and what it has changed to. In the above example, however, if the “name” tag of the funnel-web spider image said “the name of this organism can be found at urn:lsid:museum.NSW.lsid.org.au:name:117858 “ then the current name can be retrieved by the same process, through a query to another authority. In this way, whenever the name changes, the image will be automatically referred to its correct name rather than to an out-of-date name.

Tools have already been built to collate species information across multiple databases based on unique identifiers. For example, inference about the names of a taxa has been built to discover unknown Genbank sequences for that taxa.

Simple examples like these show the power that LSIDs are bringing to the World Wide Web. The global biodiversity community has a real and immediate need for LSID technology; indeed, the success of initiatives such as the Atlas of Living Australia and the ePedia of Life depend on LSIDs being used extensively by our community’s information systems. It’s probable that the next generation will use LSIDs in reasoning and inference engines to create information structures that can hardly be imagined today.

All Australian herbaria and museums have been asked to implement LSIDs as quickly as possible with available resources. Early benefits expected to flow include more effective management of specimen records between institutions, better handling of taxon names and concepts, and more (and more flexible) electronic floras, faunas and identification keys. The average client may not see the LSIDs in the background, but their presence will ensure that available knowledge is accessible.

Subsequently published as:

Thiele, K.R. (2007). Life Science Identifiers (LSIDs) in the global and Australasian bio-diversity communities. *Australian Systematic Botany Society Newsletter* 130: 8-10.

Recommendations

A series of thematically grouped recommendations, were generated during the workshop. These recommendations provide a firm basis on which to establish high-level policy for the strategic coordination of biodiversity data within the Australasian biodiversity federation.

The recommendations are primarily aimed at the peak Australasian bodies that together are the custodians of major authoritative biodiversity datasets, such as CHAH – the Council of Heads of Australasian Herbaria and CHAFC – the Council of Heads of Australian Fauna Collections, representing the Herbarium and Museum institutions respectively.

We expect these bodies to use this document to formulate, integrate and communicate their plans for implementing the architecture for building collaborative information systems in the international arena. These recommendations therefore have relevance to the global biodiversity organisations such as the Global Biodiversity Information Facility (GBIF) and Biodiversity Information Standards (TWDG).

Policy Recommendations

Participants in this workshop recommend that CHAH and CHAFC:

1. ratify and adopt the recommendations in this document
2. ensure each participating institution allocates resources towards implementing LSID technologies according to the proposed Roadmap (see page 13)
3. encourage content purchasers and aggregators to also adopt compatible technologies so that the opportunities for sharing data and building innovative biodiversity products is maximised
4. participate actively in policy decisions and standards development that support the adoption of enabling technologies such as LSIDs.

Management Recommendations

1. LSIDs will be applied to a core set of data objects

   Now:

   - Specimens
   - Names
   - Taxa
   - Descriptions

     Five institutions are exposing taxon descriptions; these could be given LSIDs as a test case. A Description is defined here as a block of text describing a lifeform, as distinct from a Taxon Profile that contains a Description plus any associated photos, maps and other content.
   - Images

   As funding becomes available:

   - Core Ontology
2. Data objects will "wear" their LSIDs
3. LSIDs will be resolvable
4. Data custodians will be responsible for LSID assignment
5. LSIDs will be treated as opaque
6. Follow best practice in the creation of LSIDs
7. Relationships within or between data objects will be exposed using LSIDs

Implementation Recommendations

1. Establish indicative generic processes
   - use TDWG guidelines
   - share implementation plans
   - document costs
   - track issues
2. Obtain commitment and funding
   - at the institutional level
   - at the community level
   - from external projects (eg. ALA)
3. Apply LSID's to data stores
4. Update data providers
5. Install and configure Resolvers
   - makes LSID's visible for exploration, retrieval, harvesting
6. Educate consumers and potential implementers via:
   - TDWG technical web site
   - TWDG general language explanation of LSID's
   - domain or institution-specific 'dummies guides'
   - an outreach program to identify and 'recruit' data sources to using LSID's
7. Extend support networks
   - TDWG 'Help Desk'
   - web-based community forums
   - onsite 'SWAT' teams of local experts
8. Data Use Agreements
   - to mandate persistence of LSIDs within downloaded data records
9. Keep up to date
   - maintain your feed
   - participate in standards bodies
   - provide feedback for refining/extending vocabularies and ontologies
10. Register lsid.org.au

Technical Recommendations

As data custodians in federated information systems, we will:

1. Accept TDWG's recommendation to use LSIDs when sharing data
   - We need to move to true GUIDS and will use LSIDs
   - Our LSIDs follow TDWG standards and will guarantee uniqueness using the format urn:lsid:Authority:Namespace:Identifier
2. Reference a range of best practice documents to support LSID implementation, including
   - TDWG LSID Applicability Statement
   - LSID Best Practices
   - Proposed "LSID for Dummies" document will also refer to best practice (being written by Ricardo)
   - The Authority component should be a domain name under control of the data provider
3. Actively participate in the development of the TDWG Ontology, Protocols and GUIDs (the 'three legs of the stool' representing the TDWG Technical Architecture)
4. Use the available tools to implement TDWG standards, such as pywrapper and TAPIRlink
5. Treat LSIDs as opaque
   - to avoid problems associated with semantic content
6. Ensure that data objects will always include their LSID (data objects need to have the LSIDs stored with them, and it should remain with them at all times)
7. Apply LSIDs to the "core set" of data objects using LSID vocabularies:
   - Specimens - some organisations and projects are currently sharing these now, but moving to a GUID standard - LSIDs - is important for globally sharing the data objects (TDWG vocabulary exists - SpecimenLsidVoc)
   - Names - some organisations are currently sharing these now, but LSIDs will enable linkages and meet the current demands of name resolution nationally and globally (TDWG vocabulary exists - TaxonNameLsidVoc)

- Concepts – only a few institutions can do this now (others will need to undertake both philosophical and technical upgrades of their own thinking/systems to enable this) - TDWG vocabulary exists - TaxonConceptLsidVoc
- Descriptions - projects such as ALA and ABIF would like us to deliver descriptions, and LSIDs will enable us to do this (globally) - no vocabulary developed yet by TDWG
- Images - again, projects will also require images, and again LSIDs will enable us to do this (globally) - no vocabulary developed yet by TDWG

8. Work towards applying LSIDs to the remaining objects (the "optimum set") identified during the meeting and will assist in the development of LSID Vocabularies:
    - lots
    - observations
    - places
    - people
    - publications / literature references
    - keys
    - characters
    - states
    - phylogenies
    - clades
    - vegetation types
    - institutions (Herbaria, Museums)
    - collections (assemblages)
    - IP/copyright
9. Be responsible, as Data custodians, for LSID assignment
10. Ensure that all LSIDs will be resolvable
11. Recommend that data custodians/providers should also be LSID resolvers, but LSID resolvers may also be provided by other organisations/projects (eg. AVH, TDWG)
12. Expose relationships within / between data objects using LSIDs
13. Recommend that TDWG generates some tools for managing and reporting upon the development of particular vocabularies and tools to the wider community.

Proposed Roadmap

Task	Description	Date
	Policy recommendations completed and circulated to the management committees	May 2007
	Recommendation to CHAH/CHAFC to formalize high level collaboration	May 2007
	Setup TDWG Oceania Interest Group - OIG	May 2007
	Ratification by CHAH, CHAFC at next quarterly meetings	June 2007
	Report and covering letter from these two groups to ALA Steering Committee	July 2007
	Formalize collaboration between HISCOM and CHAFC equivalent - consider ad hoc issue specific meetings	July 2007
	OIG group meet (workshop or e-conference) to handle regional issues outside the remit of the current project-focussed groups	July 2007
	TAPIRLink Resolver release (some data services reliant upon this)	July 2007
	Project management tools adopted -- TDWG wiki, Issue tracking	August 2007
	Implementation Planning -- Approved recommendations implemented -- in time for TDWG meeting	August 2007
	NZ Landcare (names, concepts, references) data	available now
	CANB data available - will include names, taxa, specimens, images	June 2007
	ABRS data available	September 2007
	Progress reports to CHAH, CHAFC, TDWG	September 2007
	TDWG 2007 - demonstration of LSID's to include Australian data	September 2007
	ALA prototype project from partnership with 'early adopters'	October 2007

Appendix – Workshop Minutes

The full workshop minutes, including presentation content and further linkages are available at:
http://wiki.tdwg.org/twiki/bin/view/GUID/AustralasianBiodiversityFederationLsidPolicy