GUID-1 Workshop Report

Version 1.1 (14-Feb-2006)

Introduction

The Taxonomic Databases Working Group (TDWG) and the Global Biodiversity Information Facility (GBIF) completed their first Workshop on Globally Unique Identifiers for Biodiversity Informatics[1] (GUID-1) at the National Evolutionary Synthesis Center (NESCent), Durham, NC, USA on Feb 1-3, 2006.

Motivation

A GUID framework is foundational in facilitating systems interoperability in biodiversity informatics. It meets the need for a universally adopted system for assigning and recognizing identifiers in the domain.

A GUID framework will help to manage and cross-link the many different types of entities that are manipulated analytically in biodiversity informatics and will improve interoperability with other related life sciences domains, such as bioinformatics and ecology.

The Group

The workshop delegates consisted of a representative cross-section of domain experts from around the world (see Appendix A).

Goals

The goals of the workshop were to:

- Discuss the requirements for globally unique identifiers for biodiversity informatics
- Select an optimal GUID technology (LSID, DOI, Handles or other)

---

1 Please refer to the TDWG GUID Wiki at http://wiki.gbif.org/guidwiki for more information on this effort. For a complete set of materials presented during the workshop, visit the GUID-1 workshop minutes page at: http://wiki.gbif.org/guidwiki/wikka.php?wakka=GUID1Minutes.

- Begin to identify key parameters for implementing an effective system
- Investigate the use of a RDF-based metadata architecture for GUIDs
- Form working groups to address key identified issues before the GUID-2 workshop

Outcomes

- Life Science Identifiers (LSID) seem the most appropriate GUID strategy in biodiversity informatics.
- The use of LSIDs does not preclude the use of other technologies where appropriate.
- LSID authorities must use the Domain Name Service (DNS) to support identifier resolution. (The LSID specification allows for other resolution mechanisms, but DNS is currently the only mechanism in use.)
- Although it is not possible to prevent multiple data providers from issuing alternate identifiers resolving to the same data record, the community should develop processes and tools to coordinate issuing of single identifiers for some classes of data (e.g. taxon names).
- Metadata should be provided as RDF serialized as XML and should exploit existing vocabularies such as Dublin Core wherever these are in wide use.
- The LSID getData method should be used only where it is possible and appropriate to return an unchanging series of bytes. In other cases only the LSID getMetadata method should be used. (This reflects the use of the terms “data” and “metadata” in connection with LSIDs.)

Justifications

The main criteria leading to the selection of LSID technology were:

- The cost-model of DOI. That technology is predicated on the idea that a revenue stream can be constructed for the identified objects, typically sufficient to defray the cost. That this is not the case for most, if not all, of the objects that are likely to be identified in our systems.
- The more dynamic nature of LSIDs, which does not require prior registration of every individual identifier before use.
- The open nature of the LSID protocol and software stack, and the ease of implementing LSIDs on different platforms.

Technology Comparison

The group compared the GUID technologies according to the following criteria:

Opacity: Is the identifier free from embedded semantic information?
Opacity was identified as a possibly important criterion in that genuinely opaque identifiers could not be used to make false inferences about the object represented by a GUID. Handles, DOIs and LSIDs all include similar levels of embedded information.

Governance: Is there a body that monitors the assignment of identifiers?
DOI has a more formal governance model for identifiers than the other standards. Assignment of identifiers is a more strongly contractual matter and all identifier assignment and access is mediated through the DOI registration infrastructure. Several use cases for GUIDs in biodiversity informatics require more dynamic assignment and resolution paths.

Guaranteed persistent: Is there any guarantee that identifiers will remain resolvable other than the commitments made by the assigning authority (commitments which must be made regardless of which technology is adopted)?
The central DOI infrastructure holds the registered identifiers and makes some commitments to host orphaned data.

Registration of assigning organisations: Must institutions register before being permitted to issue identifiers?
Issuing authorities for Handles and DOIs are registered centrally. LSID resolvers must be registered in DNS but do not need to be identified to a central LSID authority.

Registration of identifiers: Must institutions register each identifier before use?
DOIs are only resolvable if they are known to the central authority.

Metadata: Do the identifiers have a standard association with metadata?
Both DOIs and LSIDs have mechanisms to provide access to metadata.

Resolvable: Does the identifier include a mechanism to retrieve the associated metadata and data?
Handles, DOIs and LSIDs are all resolvable in this way.

Globally unique: Is there a commitment that the identifier will uniquely identify a single object?
Handles, DOIs and LSIDs all involve commitments to global uniqueness.

Relocatable: Can an organisation's identifiers be transferred for a different organisation to resolve (e.g. upon closure of the issuing institution)?
An assigner of Handles, DOIs or LSIDs can pass responsibility for resolution to another resolver organisation.

Individually relocatable: Can individual identifiers be transferred for a different organisation to resolve?
Individual Handles or DOIs may be assigned to other organisations to resolve. This is not possible with LSIDs.

Open architecture: Does TDWG have the ability to take over ownership of the standard and software if others stop supporting it?
Handle and DOI are both based on proprietary technologies. LSID is based on a more open strategy.

Affordable: Is the technology affordable for TDWG, GBIF and its partners?
TDWG partners together expect to assign many millions of GUIDs and have no model to fund the cost of DOIs. The cost of licensing Handle technology is unclear. LSIDs will involve costs in development of processes and infrastructure, but TDWG has more control over the process.

Summary Technology Comparison

The following table includes catalogue numbers and taxon names for comparison as these are examples of identifiers currently in use for data integration in biodiversity informatics.

Criterion	Catalogue numbers	Taxon names	Handle	DOI	LSID
Opaque	+/-	-	-	-	-
Governance	+/-	+	-	+	-
Guaranteed persistent	N/A	N/A	-	+	-
Registration of assigning organisations	-	-	+	+	-
Registration of identifiers	+/-	+/-	-	+	-
Metadata	-	-	-	+	+
Resolvable	-	-	+	+	+
Globally unique	-	-	+	+	+
Relocatable	-	-	+	+	+
Individually relocatable	-	-	-	+	-
Open architecture	-	-	-	-	+
Affordable	+	+	?	-	+

Table 1 – Summary Comparison of GUID Technologies

Work plan

There are still many issues to address before our community can fully implement an identifier system based on LSIDs. The workshop addressed a number of specific issues and developed working groups to address the following issues:

- Developing white papers to address best practices and key infrastructure questions.
- Prototyping activities.

The Infrastructure Working Group

This **group** was formed to address the key issues regarding the deployment of LSID as the GUID technology for biodiversity informatics. The mandate of this working group is **to identify required or desirable policies and infrastructure components to ensure robust, long-term operation of shared GUIDs**.

The following activities were identified:

1. Specify minimal standards (including tools and services) for GUID issuance.
2. Investigate long-term archival of LSIDs and associated data and metadata.
3. Investigate establishing (optional) central registration authority.
4. Investigate establishing repository for data and (orphan) datasets with GUIDs
5. Investigate the feasibility, existing actors and requirements for a ”Publication Bank” (a resource to act as a central registry of taxonomic literature and its digital representations, including assigning GUIDs to each publication.
6. Clarify the distinction between GUIDs assigned to data objects and to conceptual entities.

7. Investigate 3rd-party annotation and link-out mechanisms.
8. Develop materials to communicate with wider community.
9. Develop best practices for assigning resolver namespaces for LSIDs.
10. Perform review of LSID specification to identify possible enhancements.
11. Perform gap analysis of LSID software.

The outcomes of this group will be a series of white papers addressing the key infrastructure issues. Those will be reviewed during the second GUID meeting later this year.

The Prototyping Working Group

Our community must experiment with LSID technology and Ontology Engineering if we are to implement a production quality LSID system. The working group will develop prototypes of test cases to test aspects of a GUID infrastructure.

The group will develop test LSID resolvers using data objects provided by each domain, such as names, specimens, and concepts. This activity will also help involve (and train) the community in developing appropriate RDF ontologies, leading to concrete recommendations and implementations.

The potential prototypes to be developed and respective Conveners are:

1. LSID resolver for **taxon names** – developed by nomenclators using IPNI database and an RDF version of TCS-Names – Group responsible: Roger Hyam, Sally Hinchcliffe, Paul Kirk
2. LSID resolver for **specimens** using DarwinCore (also ABCD?) - Steve Perry.
3. LSID resolver for **taxon concepts** by SEEK using TCS.
4. LSID resolver for **observations** by SEEK using EML.
5. LSID resolver for **character data** by Damian Barnier, Kevin Thiele
6. LSID resolver for **images**: Greg Riccardi (MorphBank), Bob Morris

The **taxon names resolver has the highest priority**.

Prototypes will address one or more of the following (but may not be full implementations of an LSID resolution service):

- Hardware and software (including LSID stack)
- RDFS/OWL vocabulary for domain
- Data mapping between local data store scheme and shared ontology

Other important tasks identified by this group are:

- Development of ontologies to represent metadata for the various domains. Coordinated by TDWG TAG with help of experienced ontology engineers.
- To set up a real live LSID server to perform scalability testing.
- A project to demonstrate the potential from LSID-based integration of data for a particular group (Ants) – LSIDs, taxonomic lit, specimen, images, names, sequences from Genbank – Rod Page
- To use SEEK Taxon resolution server (alpha) for testing.

This group will have 3 months to work on the specified tasks before preparing for the second GUID workshop.

Next Workshop: GUID-2

The TDWG Infrastructure Project is planning a second GUID workshop in late May or early June, 2006. At the time of writing a venue has not been decided.

The second workshop should cover the following:

- Review of the material produced between the workshops by both working groups (prototypes and white papers).
- Summary of the lessons learned in the process.
- Identify open issues and devise specific work plans to address them.
- Draft concrete recommendations on GUIDs for production systems – in general and for each specific domain (names, specimens, concepts, images, etc).

Information about GUID-2 will be distributed as soon as possible.

Appendix A: List of Participants

Participant	Institution	Country
Andrew Jones	Catalogue of Life	UK
Benjamin Szekely	IBM	USA
Bob Peet	University of North Carolina	USA
Cliff Cunningham	National Evolutionary Synthesis Center (NESCent)	USA
Dag Terje Endresen	International Plant Genetic Resources Institute (IPGRI)	Italy
Damian Barnier	Centre for Biological Information Technology (CBIT)	Australia
Donald Hobern	Global Biodiversity Information Facility (GBIF)	Denmark
George Garrity	Bergey's Manual Trust - Michigan State University	USA
Gerald Guala	United States Department of Agriculture (USDA)	USA
Greg Riccardi	Florida State University	USA
Hideaki Sugawara	DNA Data Bank of Japan	Japan
Jessie Kennedy	Napier University	UK
Joel Kingsolver	National Evolutionary Synthesis Center (NESCent)	USA
Kevin Richards	Landcare Research	New Zealand
Lee Belbin	TDWG Infrastructure Project	Australia
M. I. Zuberi	University of Rajshahi	Bangladesh
Matt Jones	National Center for Ecological Analysis and Synthesis (NCEAS)	USA
Patricia Gensel	University of North Carolina at Chapel Hill	USA
Paul Kirk	CABI Bioscience	UK
Peter Dawyndt	The World Federation for Culture Collections (WFCC)	UK
Ricardo Pereira	TDWG Infrastructure Project	Brazil
Richard L. Pyle	Bishop Museum	USA
Robert Huber	MARUM - Institute for Marine Environmental Sciences	Germany
Roderic Page	University of Glasgow	UK
Roger Hyam	TDWG Infrastructure Project	UK
Sally Hinchcliffe	International Plant Names Index (IPNI)	UK
Scott Federhen	GenBank	USA
Simon Coppard	International Code of Zoological Nomenclature (ICZN)	UK
Stan Blum	California Academy of Sciences	USA
Steve Perry	University of Kansas	USA