GUID-2 Workshop Report

Second International Workshop on Globally Unique Identifiers for Biodiversity Informatics (GUID-2)

10th – 12th June 2006
e-Science Institute, Edinburgh, UK

1. Background

The Taxonomic Databases Working Group (TDWG, http://www.tdwg.org/) and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org) held the second of two workshops to address the need for a system of globally unique identifiers (GUIDs) for use in biodiversity informatics. The first workshop was held at NESCent in North Carolina in February 2006 and recommended the adoption of Life Science Identifiers (LSIDs) for this purpose (see: GUID1Report). Subsequent activity has focussed on the development of prototypes, integrating LSIDs with a key data sets and on-line discussion of key infrastructure and policy implications of using LSIDs.

2. Summary

The second GUID workshop was held at the e-Science Institute (e-SI) in Edinburgh, Scotland, 10-12 June 2006. The meeting re-iterated the conclusions from the first workshop and developed the following recommendations and priorities for the use of LSIDs, including:

- Work first with the nomenclators, particularly IPNI, Index Fungorum and ZooBank, to associate LSIDs with these reference lists of scientific names, recognising their foundational place in biodiversity informatics.
- Update software tools for LSID resolution and develop packaged installers.
- Support integration of LSID functionality into TDWG data provider packages.
- Develop documentation and publicity materials for managers, biologists and IT specialists.
- Approach external software projects to encourage the support of LSIDs.

3. Goals

1. To review the outcomes of the first GUID meeting and subsequent developments.
2. To confirm the choice of LSIDs as persistent identifiers (GUIDs) for TDWG functional data types as will be defined in the TDWG Core Ontology (currently under development, http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology).
3. To discuss the governance of GUID allocation.
4. To identify issues and discuss the options for ensuring long term availability and archival strategy for GUID objects (or their data sets).
5. To identify gaps and weaknesses in current GUID server and resolving software and to list the actions required to address them.
6. To find ways to promote the benefits of using GUIDs to the taxonomic, biological research and biodiversity information communities.
7. To identify the technical and 'sociological' support required to assist potential users to start serving GUIDs.
8. To review and make recommendations on the incorporation of GUIDs and RDF into TDWG standards.

4. Discussion

- The meeting identified the importance of promoting the use of LSIDs to the broader community. This goal evolved into the formation of an outreach group [GuidOutreach] to identify priorities for accelerating the update of LSIDs within the broader community associated with TDWG.

- A discussion on the relative merits of using HTTP URI (e.g. PURL or URL) rather than LSIDs was motivated by the current lack of support for LSID resolution in client software and browsers versus the ability to resolve ubiquitous URLs. The use of DNS as the current resolution mechanism for LSID makes them appear similar to regular URLs but with this client side impediment. Having considered these issues, the meeting concluded that the selection of LSIDs made at the GUID-1 meeting remained sound because:
  - o URLs are bound to a single transport protocol (http). It would not, therefore, be possible to separate the resolution mechanism from the identification mechanism in future.
  - o LSIDs are names and abstract the resolution mechanism, and that mechanism may change in the future.
  - o LSIDs provide a mechanism for binding data and metadata that would have to be invented for the use with URLs.
  - o The LSID specification recommends RDF as a metadata format. This would have to be mandated with the use of URLs.
  - o Socially, people do not trust URLs to be permanent because of the 'link rot' they have encountered on the hypertext web. LSIDs are more likely to gain acceptance for use in publications etc (as is the case with DOI).

- Software development – Discussion of the availability and functionality of both provider and client software resulted in a 'break-out' group that produced prioritised recommendations for filling the gaps, improving usability and

extending functionality including how to get client software to resolve URNs (see GUID2WgSoftwareDevelopment)

- What gets an LSID? - The recommendation is that the architecture requires that an object identified by an LSID should be 'typed'. Therefore anything that can be explicitly typed, preferably with the TDWG ontology, could be named with an LSID. The depth of granularity is subject to what the provider wishes to manage. We recognised that we needed an explicit policy statement about LSIDs being attributed to abstract vs. descriptive objects.

- Metadata annotation and link out. Foreign Authority Notification (FAN) can be useful for notification but cannot be effectively used to create back-links to external annotations, because of management issues. A possible solution would be to create a central service where agents could register any annotations that would be available to anyone in the community. If providers are willing to expose their systems to web crawlers, normal discovery methods (e.g. through Google) would also locate distributed references to GUIDs.

- There was insufficient evidence from case studies to provide definitive rules about versioning. There is a need for a TDWG policy statement backed with examples. See: VersioningOfLSIDs

5. Outcomes

Recommendations

A number of specific, mainly technical recommendations were made by the end of the conference:

1. LSIDs will be used as unique identifiers.
2. HTTP GET will be the default binding for getMetadata calls.
3. RDF will be the default metadata response format.
4. The TDWG ontology (and other well-known vocabularies) should be used wherever possible.
5. The use of non-RDF as an accepted_types for metadata is deferred.
6. There should be a central DNS SRV hosting service (see CentralSrvService).

Actions

In addition to technical recommendations a number of actions were identified and allocated, with a commitment to reporting back on progress at the TDWG annual meeting in October 2006.

- Lee to write a 1-page document on the case for LSIDs aimed at senior administrators/managers by the end of June 2006
- Rich to develop a document/presentation on the case for LSIDs to biologists for presentation at TDWG 2006 (August 2006)
- Ricardo, with assistance from Rod and Ben to evaluate software resources collected during GUID-2 and identify gaps in requirements and in the associated documentation (August 2006)

- Roger to evaluate the status of the documentation necessary to promote the uptake of LSIDs within our community (August 2006)
- A narrative explanation for the use of LSIDs and RDF in the biodiversity infrastructure to be written by Roger (July 2006)
- Stable informative material is to be loaded from the Wiki to the new TDWG website
- The need for applicability statement and guidelines for biodiversity organisations serving LSIDs
- Ricardo and Roger to attempt to produce a packaged installer and improved documentation for setting up LSID server software (October 2006).
- Ben to approach IBM to improve access to and use of their LSID resolver plug-ins (with support from Ricardo)
- Encourage development of 'good news' demonstrations through ant and fish communities
- Review tasks involved and costs of setting up the host for a central DNS server service for LSIDs

6. Conclusions

- There has been considerable discussion about the utility of a URN vs. URL strategy. The meeting concluded that our community should implement LSIDs. Experience of the prototypers suggest that implementation of LSID Authorities has been straight-forward.
- Priority actions identified by the meeting should be completed within the next three months to achieve significant progress that must be reported to TDWG 2006.
- The server and client software and associated documentation must be improved to encourage uptake of LSIDs. Key improvements are: i) simplified installation of LSID server software ii) native or 'plug-in' support for LSIDs in semantic web tools, ontology software and browsers.
- Promotional documentation targeted separately at managers, biologists and IT specialists is urgently required.
- Implementation of LSID Authorities at a few key nomenclators and two operational examples with a few associated collections is a priority.
- There is a sound case for establishing a central DNS server service for LSIDs, which will be further assessed and costed.

Annexes

Presentations

The meeting included presentations, break-out group discussions and plenary discussions. The presentations given were:

1. Introduction and overview - *Donald Hobern*
2. Summary of architectural recommendations from TAG - *Roger Hyam*
3. LSID Tester - *Roderic Page*
4. LSID resolver for specimens in DiGIR2 - *Steve Perry*
5. Image LSID resolver prototypes - *Bob Morris*
6. Index Fungorum LSID authority setup - *Kevin Richards*
7. Port of LSID framework to .NET - *Kevin Richards*
8. LSID resolution in SEEK - *Jessie Kennedy*
9. Existing software components and gap analysis - *Ricardo Pereira*
10. Core Ontology - *Jessie Kennedy*
11. Additional components of an LSID network - *Donald Hobern*
12. Publication Bank - *Robert Huber*

See also the list of presentations at the eSI website.

Attendees

Damian Barnier (University of Queensland)
Lee Belbin (TDWG)
Stan Blum (California Academy of Sciences)
Charles Copp (EIM, Technical Writer
Simon Coppard (NHM, London)
Rob Gales (University of Kansas)
Sally Hinchcliffe (RBG, Kew)
Donald Hobern (GBIF)
Robert Huber (Marum - Zentrum fur marine Umweltwissenschaften)
Roger Hyam (TDWG)
Jessie Kennedy (Napier University, Edinburgh)
Chuck Miller (Missouri Botanical Garden)
Bob Morris (University of Massachusetts, Boston)
Tom Orrell (ITIS, Smithsonian Institution)
Roderic Page (University of Glagow)
Ricardo Pereira (TDWG)
Steve Perry (University of Kansas)
Andrew Polaszek (NHM, London)
Rich Pyle (Bishop Museum, Hawaii)
Kevin Richards (Landcare Research NZ Ltd)
Tim Robertson (GBIF)
Vince Smith (NHM)
Ben Szekeley (IBM)
Dave Vieglais (University of Kansas)