International Union for
Biological Sciences

Taxonomic Databases Working Group
http://www.tdwg.org

Technical Roadmap 2006

1. Introduction

The TDWG Infrastructure Project was given the remit to devise an umbrella architecture for TDWG standards. The purpose of this architecture is to:

- Provide a unified vision of the existing and proposed TDWG standards. It is important for the credibility of TDWG that its proposals are seen as part of an integrated whole.
- Suggest how TDWG standards should evolve so that they are interoperable with each other and external standards in the future.
- Maximise the effect of the limited resources of TDWG.

A meeting (TAG-1[1]) of representative from groups currently active within TDWG was held in April 2006 in Edinburgh. This meeting produced a series of recommendations for the TDWG architecture in a report[2] that was widely circulated and adopted by the TDWG executive in Madrid in June 2006.

Although the TAG-1 report contains the major recommendations for the new architecture, it is not in a form that is easily accessible to those previously unaware of the issues involved. The report does not stipulate how or when these recommendations should be adopted.

This document builds on the TAG-1 report and combines it with conclusions coming out of the GUID meetings to give a broader context and roadmap for the adoption of appropriate technologies by TDWG.

2. Core Definitions

There are two core roles that participants in the exchange of biodiversity data may adopt:

- **Data Provider:** An entity (organisation or individual) that makes data available to others by publishing it to the Internet using Data Provider Software.
- **Data Consumer:** An entity that consumes data from Data Providers using Client Software.

Other roles are based on these two roles. A portal, for example, will use client software to consume data from data providers and use provider software to publish it to other consumers. A research scientist may use client software to consume data, and provider software to publish their own related data back to the Internet.

---

[1] http://wiki.tdwg.org/twiki/bin/view/TAG/TagMeeting1Report

[2] http://wiki.tdwg.org/twiki/pub/TAG/TagMeeting1Report/TAG-1_Report_Final.pdf

3. Core Principles

TAG-1 established two foundational principles:

- "*The architecture is concerned with shared data.*"
- "*Biodiversity data will be modelled as a graph of identifiable objects.*"

3.1. Principle 1: Shared Data

The TDWG architecture applies to data that is shared between entities. Only when data crosses boundaries does the format[3] matter. The architecture should not dictate Data Providers or Data Consumers internal structures. A successful architecture should enable the interoperability of providers and consumers with radically different internal implementations.

3.2. Principle 2: Identifiable Objects

Exchange limited to literals (strings and numbers) in unlabelled packages is of no value. The number '55.7' has no meaning to a data consumer on its own. If it is combined with other literals in a labelled package then it is useful. For example:

SamplingStation	
id	18439279
longitude	-2.7
latitude	55.7
name	Lauder 2

But what is a SamplingStation? There is no written description here. What is the datum used for the longitude and latitude? This is an instance of an object of type SamplingStation but if we are to wrap literals in objects, we need **a type catalogue or ontology** where information about the semantics of objects can be stored and retrieved – both by humans and increasingly, by machines.

If we want to refer to this particular instance of a SamplingStation we could use the contents of its id field but the scope of the id cannot be known without further information. Is it unique to all sampling stations or just those from one particular data provider or perhaps all objects in the entire network? We need a system of **Globally Unique Identifiers (GUIDs)** if we are to refer to instances of objects across all Data Providers.

How do we find out more about this SamplingStation? If the id was resolvable then we could use it to get a response. Alternatively we could run a query against one or more Data Providers. To do this we need well-defined **data exchange protocols** that our client software can exploit.

4. A Conceptual Model: The Three Legged Stool

Modelling biodiversity data as a graph of identifiable objects implies an architecture that stands on 3 legs:

[3] The word 'format' is used as a convenience term in this report to mean the combination of protocol and conceptual schema. Existing exchange standards often bind conceptual schemas to the transport protocol. DiGIR does this using substitution groups. SPARQL is bound to RDF. Other schemas are used almost entirely with a single protocol. ABCD is used with BioCASe although will soon be available via TAPIR.

- A type catalogue or ontology – here referred to as the Semantic Hub.
- A system of Globally Unique Identifiers.
- Well defined data exchange protocols.

A three legged stool is a useful metaphor because the legs are all equally important: remove one and the architecture fails; there are multiple dependencies between the legs.

This model will be the focus of the remained of this document.

5. The Semantic Hub

The TAG-1 report identifies a need to develop Core and Base Ontologies to act as a typing mechanism for exchanged objects. There are problems with the word *ontology*. For some, the word implies W3C semantic web technologies – specifically an ontology defined in OWL or a similar language. For others *ontology* is a general term – the Gene Ontology is a vocabulary that may or may not be expressed in OWL.

The term *Semantic Hub* will be used because any TDWG typing system will have to work across multiple technologies. It must be stressed that the hub is a place of convergence for the semantics **not** the data. Initially the Semantic Hub must supply:

- RDFS and/or OWL definitions of objects so that they can be returned in the metadata response of LSID authorities as well as mapped into pure semantic web applications based on SPARQL such as WASABI (formerly known as “DiGIR2”).
- GML Application Schema definitions of objects so that they can be supplied as part of an OGC Web Feature Service.
- XML Schema definition of objects for use as TAPIR output models.
- In the future it may need to be presented as an ebXML RIM[4], OGC Feature Type Catalogue or in some other way.

The same basic semantics must persist across these technologies. A TDWG Specimen needs to be a TDWG Specimen whether it is expressed in GML, OWL or different XML Schemas. This translation process (translation of conceptual schemas or semantics **not** data) must be explicit, documented and preferably automatic.

The only alternative to a Semantic Hub is to advocate the adoption of a single technology for all interactions within the TDWG community. If the architecture were to advocate the adoption of a purely W3C Semantic Web based approach then integration with OGC GML applications would not be possible. Similarly, the use of bespoke XML Schemas as conceptual schemas (as at present) does not permit integration with Semantic Web technologies or GML-based technologies, and use of a purely GML-based approach does not permit integration with bespoke XML Schemas or Semantic Web technologies. The semantics of the TDWG community need to be mapped into all these technologies.

5.1. Populating the Semantic Hub should not be a new modelling exercise.

TDWG has already modelled its domain and the semantics are available in the existing schemas. Establishment of the hub should be a process of translation, refactoring and mapping.

[4] http://www.ebxml.org/

5.2. Implementation

There are constructs that can be expressed in OWL that can't be expressed in UML or GML and in UML that can't be expressed in either GML or OWL etc.

If it is important to allow for objects to be described in different semantic languages, it is important only to use the subset of constructs that is supported by all the different technologies. Formal mappings must be defined to explain how to express each construct in each target language. Modelling needs to be restricted to this subset of common constructs. At present it is believed that this subset may consist of a single inheritance hierarchy of classes possessing properties with ranges restricted to instances of other classes or literals. It is outside the scope of this document to explore the choice of constructs further. The precise subset is still a research/discussion issue that should stabilise by October 2006.

A web-based application (Tonto) is being developed to manage a semantic hub. Tonto should act as a collaborative environment for the creation of object descriptions. Tonto should automatically map these object descriptions into different technologies and place appropriate files in the rs.tdwg.org domain. This is an on going effort by the TDWG Infrastructure Team.

6. Globally Unique Identifiers

The Semantic Hub addresses the typing of objects (nodes) within the graph of biodiversity data. The Hub does not address the arcs in the graph, i.e. how objects reference each other. How does a specimen object refer to the collection it is a part of? The most appropriate answer is through the use of GUIDs.

The use of GUIDs within the biodiversity informatics community has been the subject of two international meetings and numerous pilot projects[5]. This process has led to the selection of Life Science Identifiers (LSIDs) as the preferred standard.

The TDWG architecture should remain as technology-neutral in the selection of particular GUID technologies as it is with modelling technologies. The role of the architecture is to span different domains which may use different GUID technologies. For example, the publishing industry is widely adopting Digital Object Identifiers (DOI)[6] and many W3C semantic web applications are likely to use URL style URIs or Permanent URLs (PURL)[7].

LSIDs do however have particular features that support the development of the graph of identifiable objects:

- **Resolvability:** Client software can dereference an LSID to a data or metadata object. This mechanism works at a global scope across the Internet so any client application can navigate the graph of identifiable objects just as a human user may follow links between web pages. The LSID technology comes with its own protocol.
- **Response Type:** The default return type for LSID metadata is RDF. This means it is possible for client applications that are navigating the graph to parse any node identified by an LSID in the same way. Clients can discover the semantics of the object represented by that node using standard Semantic Web methods (such as rdf:type or rdfs:isDefinedBy) from the TDWG Semantic Hub or a third party (if the class is not in the TDWG domain).

---

[5] http://wiki.gbif.org/guidwiki/

[6] http://www.doi.org/

[7] http://purl.org/

6.1. Implementation

- LSID has been adopted as a GUID technology within the biodiversity informatics community.
- Pilot studies have had positive results using existing and bespoke LSID Authority software.
- Developers of both WASABI and PyWrapper have indicated that they will provide support for LSIDs in their packages.
- Several data providers (initially focussing on nomenclature) have agreed to establish LSID-supported production systems in the near future.
- The primary barrier to the widespread adoption of LSIDs in biodiversity informatics is the lack of RDF or OWL vocabularies to which to map the metadata. These vocabularies will be provided by the Semantic Hub.

7. Data Exchange Protocols

Modelling biodiversity data as a graph of identifiable objects is of little use without protocols to access those objects. It is however unlikely that there will ever be a single data exchange protocol used throughout biodiversity informatics community, for at least two reasons:

- No single protocol will be optimal for all use cases i.e. there will be applications where a specialised protocol is required. It would be inefficient to use a query protocol like SPARQL for simple GUID resolution when the LSID HTTP GET bindings would be more appropriate.
- Adoption of new protocols is time consuming. When a new protocol or a new version of a protocol is introduced it will take some time to be deployed across the network. During this time two versions will be in use.

Data transformation will therefore have to take place if data access is to be ubiquitous rather than restricted to protocol-specific silos. There are three possible locations where transformation can occur:

- **At the data provider:** Providers could map their internal data structures to multiple external formats. The PyWrapper2[8] and WASABI (formally DiGIR2) packages are both pursuing this strategy. BioCASe[9] providers already publish data mapped to multiple versions of the ABCD conceptual schema.
- **At an intermediate transformation service:** Portals or other service providers could map between the different formats. The GBIF data portal is pursuing this strategy.
- **At the client:** Client software packages could ‘understand’ multiple formats or make use of transformation services.

Because transformations have to occur to assure ubiquitous communication, it is important to make them as simple as possible. The Semantic Hub will facilitate this. It will not provide the actual transformation services but will provide normative mappings between representations of the same concepts in different technologies; GML, XML Schema, OWL, RDF etc. This mapping will enable data providers to map their data to a single conceptual view and automatically expose it using different conceptual schemas bound to different

[8] http://www.pywrapper.org/

[9] http://www.biocase.org/

exchange protocols. Client applications will be able to take a similar approach. Such transformations will not be possible without the Semantic Hub.

The architecture does not attempt to solve the problem of multiple exchange protocols and their associated conceptual schemas. As the Semantic Hub matures, concepts used by Data Providers will become more broadly shared and will be mapped to multiple representations. Providers and consumers will come to use the most appropriate formats.

7.1. Implementation

A summary of the current technologies in use is given in Appendix B of the TAG-1 meeting report. This report concludes that there will be in excess of 400 data providers using a combination of the DiGIR, BioCASe, TAPIR, SPARQL and LSID protocols by 2007.

The following steps must occur to unify these protocols:

- The Semantic Hub must provide a single set of semantics for biodiversity data.
- The Semantic Hub must rendered the semantics in ways that are appropriate for each of the exchange protocols e.g. 'Flat' XML Schemas with appropriate substitution groups for DiGIR; OWL ontologies for SPARQL etc.
- Data Providers must map their data to these views.

These steps will take time to accomplish but the infrastructure should be in place and the process begun within a year – See below.

8. Time Line

The semantic leg of the 'architectural stool' is currently far shorter than the others and efforts are underway to try and rectify this.

8.1. October 2006

- TDWG Collaborative Architecture and new Standards Process will be launched.
- TAPIR finalized and documented.
- PyWrapper2 will be deployed as an initial implementation of TAPIR.
- WASABI will be deployed as a SPARQL data provider.
- Tonto will be launched as a management tool for the Semantic Hub.
- Initial Core and Base ontologies will be presented within Tonto.
- RDFS vocabularies used for LSID metadata will be presented within Tonto.

8.2. October 2007

- Tonto will be at version 2 - to take account of requirements from data providers and data consumers revealed in the past year.
- New data provider deployments will use the Semantic Hub as their source for concepts and output models.
- TAPIR will be implemented on WASABI.
- Demonstration projects that expose the same data through multiple exchange protocols to radically different clients will be presented e.g. GIS tools, Semantic Web browsers, data portals.

8.3. 2008 and beyond

- Existing data providers will migrate to Semantic Hub based schemas within their normal cycle of software replacement.
- The semantically integrated approach taken by TDWG will encourage the development of more sophisticated client software packages and services which will help catalyse the release of more data of higher quality and so facilitated research and decision making. These benefits will continue to accrue from initial adoption in 2007 into the future.