

A Technical Architecture for TDWG Standards

**TDWG Technical Architecture Group
Roger Hyam**

An Architecture for TDWG Standards

- Provide a unified vision of the existing and proposed TDWG standards
- Maximise the effect of the limited resources of TDWG
- Suggest how TDWG standards should evolve to
 - Increase interoperability with each other
 - Increase interoperability with external standards

TAG-1: First Meeting

- **“The architecture is concerned with shared data”**
 - Format matters when data crosses boundaries.
- **“Biodiversity data will be modelled as a graph of identifiable objects”**
 - Documents are nested objects
 - Literals on their own are of no value

Literals in Labelled Packages are Useful

- 55.7 doesn't mean anything on it's own but in

```
<SamplingStation>  
  <id>18439279</id>  
  <longitude>-2.7</longitude>  
  <latitude>55.7</latitude>  
  <name>Lauder</name>  
</SamplingStation>
```

- What is a SamplingStation?
- What is the datum used?
- **Conclusion 1: We need an ontology**

Building on Current Knowledge

- TDWG has already captured much of the semantics for biodiversity informatics
- These semantics are modelled in XML Schemas
- There is overlap between the XML Schemas
- There isn't a single ontology
- ... but the ingredients are there to build one.

How do we link objects?

- Value of the graph increases with number of meaningful connections
- Must refer from one object to another to make the edges of the graph
- Objects need a resolvable identifier
 - `<id>18439279</id>` = Not good enough
 - Is it unique to all sampling stations?
 - Is it unique to all objects?
- **Conclusion 2: We need a GUID System**

How do we find objects?

- What are the neighbouring sampling stations?
- Need to have protocols to ask questions and discover data on the network.
- Crawling the network from object to object is not enough.
- **Conclusion 3: Need well-defined data query and exchange protocols.**

The Three Legged Stool

- Ontology – for typing the nodes
 - Existing XML Schemas: ABCD, DwC, SDD etc
 - Core Ontology developed 2006
 - Ontology Metamodel
- GUID – for joining the nodes
 - LSID: 2 workshops and much work
- Protocols – for retrieving the nodes
 - TAPIR, BioCAsE, DiGIR
 - SPARQL, WFS, LSID, Other.

Bringing It All Together

- Can we use a single protocol?
- No single protocol is optimal for all use cases.
 - Shouldn't bend LSID protocol to query or use SPARQL just for resolution
- Adoption of new protocols is time consuming
 - By the time a new protocol is fully deployed the replacement will be ready
- **Conclusion: Integration must occur at the semantic level not the protocol level**

What should the ontology do?

- Different technologies are in use, we need:
 - RDFS and/or OWL definitions of objects
 - GML Application Schema definitions of objects
 - Custom XML Schema definition of objects
 - In the future it may need to be presented as an ebXML RIM, OGC Feature Type Catalogue or in some other way
- The same semantics must persist across these technologies

How will the ontology do this?

- Need a metamodel that maps to multiple technologies.
 - Metamodel is subject of separate presentation
- Need an application to administer the ontology.
 - 'TDWG **O**ntology **T**ool' has been developed to demonstrate this

Summary of Architecture

- Clients and servers exchange data using the most appropriate protocol.
- Data is modelled as a graph of objects.
- The objects are typed using a central ontology.
- The ontology has one set of semantics internally but different facets for different protocol technologies.
- Objects are identified and linked using LSIDs.

How do I participate?

- Use frame based modelling
 - Classes and Properties
- Read the '*TAG Roadmap 2006*'
 - <http://www.tdwg.org/>
 - Subgroups > TIP > TIP Document Archive