

TDWG Technical Roadmap 2007

Technical Architecture Group | Convenor: Roger Hyam | 27th August 2007

Introduction

The Technical Architecture Group of the Biodiversity Information Standards (TDWG) is charged with maintaining an overview of TDWG standards, their relationships to each other and to standards produced by other organisations. The charter of the TAG states

that a yearly report will be produced called the 'Roadmap'. This is the 2007 Roadmap document. It supersedes the 2006 Roadmap that was presented at the St Louis meeting.

Standards Architecture

In 2005 the TDWG Infrastructure Project (TIP) was given the remit of devising an umbrella architecture for TDWG standards. A meeting (TAG1) was called in April 2006 and this led to the establishment of the basic principles for underlying the standards architecture, as well as the Technical Architecture Group itself. The TIP have been promoting adoption of this common architecture over the last 18 months.

Why have a standards architecture?

From the point of view of exchanging data – such as in the federation of a number of natural history collections – there is no need for a standards architecture. The federation is a closed system where a single exchange format can be agreed on. The federation can grow by adding new members whose needs are met by the format. This model has worked well in the past but it does not meet the primary use case that is emerging. Biodiversity research is typically carried out by combining data of different kinds from multiple sources. The providers of data do not know who will use their data or how it will be combined with data from other sources. The consumer needs some level of commonality across all the data received so that it can be combined for analysis without the need to write computer software for every new combination. This commonality needs to seamlessly extend to new types of data as they are made available. An architecture is required to provide this commonality.

What form should the architecture take?

A degree of commonality could be achieved simply by specifying how data should be serialised. If all suppliers passed data as well-formed XML, for example, it would

provide a degree of interoperability but clients would still not know how the elements within one XML document relate to those in another or how the items described in those documents were related. At the other extreme, the architecture could provide a detailed data type library which described the way in which each kind of data should be serialised at a fine level of granularity – which XML elements must be present and what they should contain – but it is highly unlikely that a single set of serialisations would meet all needs any more than a single federation schema would. It remains a requirement of some thematic networks that they have well defined data types to ensure that the data passed is valid and fit for purpose.

The architecture therefore has to meet two needs. It has to allow generic interoperability but also restricted validation of data for some networks. It does this by taking a three pronged approach.

- 1) An ontology is used to express the shared semantics of the data but not to define the validity of that data. Concepts within the ontology are represented as URIs (Universal Resource Identifiers).
- 2) Exchange protocols use formats defined in XML Schema (or other technologies) that exploit the URIs from the ontology concepts.
- 3) Objects about which data is exchanged are identified using Globally Unique Identifiers.

This means that (although exchanges between data producers and clients may make use of different XML formats) the items the data is about and the meaning of the data elements is common across all formats.

Implementing the architecture

The Ontology

Prior to the TDWG standards architecture, data exchange has been based solely on passing XML documents. This is good for federation networks but it is not as suitable for sharing different types of data across a generic data exchange network – which is emerging as the primary use case. Combining documents is difficult because the meaning of the elements within the documents depends on their context. If we initially model the shared data as an ontology of linked classes of object rather than documents, it becomes possible to construct documents from the perspective of different base classes that map

directly to the ontology. Clients can then combine documents from different perspectives (and of different formats) because they understand the ontology the documents are based on.

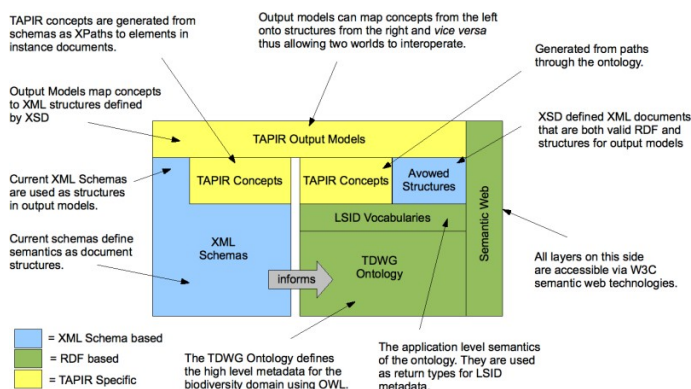
Applying the principle of separation of concerns, it is possible for the definition of the validity of the documents exchanged to be defined outside the ontology. The ontology can be used to specify the meaning of the namespaces whilst XML Schema (or some other technology) can be used to specify valid document structures for any particular exchange application. An ontology is therefore central to unifying disparate application schemas.

Last year a team lead by Jessie Kennedy and including representatives from across TDWG interest groups developed an initial high level ontology of the biodiversity informatics domain. This ontology is available through the TAG Wiki. Creating the ontology was a valuable exercise but everyone involved recognised it needed more work before it could be put to production use. At the same time, a programme was actively rolling out LSID (Life Science Identifiers) authorities. The metadata returned by LSIDs is in RDF format and, to be useful, requires an RDF vocabulary or ontology that at least defines the object types. The TDWG ontology was not going to be developed to a sufficient level of detail in time. The decision was therefore taken to develop a series of smaller ontologies that could serve as an application layer within the larger TDWG ontology and to only loosely link them into the higher classes in the ontology. These two ontologies are referred to as the "LSID Vocabularies" and the "Current TDWG Ontology".

The LSID Vocabularies are now entering production use and, in due course, there will be a requirement for them to be linked to a higher level ontology so as to permit inference. This process requires investigation. It may be possible not to hard link the two parts of the ontology but apply a separation of concerns again. There are multiple ways in which the basic classes of exchanged data could be related. No one set of these relationships is suitable for all applications. It is therefore important not to impose a top-down interpretation of the data but to allow for the possibility of multiple higher level classifications of which the Current TDWG Ontology may be one. It may also be important to link the LSID Vocabularies directly to other upper ontology efforts such as UMBEL (<http://umbel.org/>), SUO (<http://suo.ieee.org/>) and SUMO (<http://www.ontologyportal.org/>).

Exchange Protocols

The architecture is independent of exchange protocols provided the protocols are capable of mapping their concepts to the ontology. Current data suppliers within the community make use of XML Schema based protocols, either DiGIR or BioCASE. These are being replaced by the TAPIR protocol. It is important therefore that the TAPIR protocol integrates with the architecture. The diagram below shows how this is achieved.



This solution relies on the melding together of two technologies that are often thought to be antagonistic: the Resource Definition Framework (RDF) and XML Schema. RDF is a modelling language that describes

everything in terms of subject-predicate-object statements (known as triples). RDF can be serialised in many ways. One of those ways is as XML. XML Schema is a language for defining XML document structures. It is possible to define an XML document structure using XML Schema so that the resulting documents are valid serialisations of RDF. TAPIR is data exchange protocol designed to pass XML messages. The output from a TAPIR data provider is described using XML Schema. TAPIR knows nothing about RDF but by using XML Schemas that define RDF instance documents, it is possible for a TAPIR data provider to behave as an RDF data source.

One of the strengths of the TAPIR protocol is that it allows the definition of custom response types (output models). This can act as a mapping point between conceptual schemas. It should therefore be possible to map other TAPIR concepts into RDF that use the TDWG ontology. This has been demonstrated using data sources mapped to DarwinCore. It should also be possible to map any TAPIR data source to generic RDF.

TAPIR is not the only exchange protocol being used:

The GBIF data portal has its own HTTP based web service that uses XML based on the LSID Vocabularies.

A demonstration OAI-PMH data provider has been built that uses a metadata format based on the LSID Vocabularies. OAI-PMH has also been integrated into the .NET TAPIR provider software.

A number of LSID based projects are making use of the LSID Vocabularies in their metadata responses. LSID support has been added to the TapirLink provider software.

The W3C SPARQL protocol and query language should support the architecture. There are numerous implementations of the protocol and a demonstration using the TDWG ontology is planned.

It is therefore becoming possible to combine data that was retrieved from a number of sources using different protocols.

GUIDS

The biodiversity information community has made primary data available for environmental analyses and decision making. Information on a million scientific names is now available through data providers such as the Integrated Taxonomic Information Service (ITIS), Species2000 and the Catalogue of Life (CoL) and almost one hundred million specimen records are provided by Herbaria and Natural History Museums around the world.

To use these data more effectively, however, clients need mechanisms to:

- Refer to authoritative information resources.
- Facilitate data integration.
- Detect duplicates of the same resource.

To achieve these goals, a system of globally unique identifiers (GUID) is needed.

In 2005 The TDWG Infrastructure Project established a TDWG Globally Unique Identifiers Task Group (TDWG-GUID) to provide recommendations for use of GUIDs in our domain. The GUID members concluded that the Life Sciences Identifiers (LSIDs) were the most appropriate

technology to address current problems.

Life Science Identifiers are unique, persistent, location-independent, resource identifiers for uniquely naming biologically significant resources, such as species names, concepts, occurrences, genes or proteins. Life Science Identifiers are a way to identify and locate pieces of biological information on the web that overcomes some of the limitations of naming schemes in use today. A standard (a TDWG Applicability Statement) is in preparation that specifies how the LSIDs should be applied within the biodiversity informatics domain. This standard will specify the use

TAPIR Provider Implementations

TapirLink

A PHP based TAPIR provider designed to run in similar environments to the current PHP DiGIR providers. Experimental LSID authority software is included.

All providers for MaNIS/ORNIS/HerpNet are actively undergoing transition to TAPIR from DiGIR (~20 in place as of 22 Aug 2007). Both will remain in use until a TAPIR-based portal application is constructed within the coming year, at which point the DiGIR providers will be deprecated. Associated with these activities is the development of a TAPIR-based Darwin Core Feed Builder (<http://code.google.com/p/dwc-feed-builder/>) to detect changes in data from Tapir providers.

Tapir.NET

A Microsoft .NET based implementation that includes

Barriers to Adoption of Architecture

There are two potential barriers to the adoption of the new architecture.

Education

The existing technologies have been successful. The new architecture offers great potential for future developments but is not perceived as adding any particular advantages to current projects. This is largely due to a lack of understanding of:

- 1) The new technologies themselves – what is required to comply.
- 2) The emerging use cases – what could be done with compliant standards.
- 3) The significance of semantic web technologies in general

Targeted resources (mainly documents) are required to

TDWG Credibility Gap

There is a credibility gap between what TDWG proposes and what it presents. TDWG presumes to be the standards organisation for electronic biodiversity data exchange yet, when the list of its standards are compared with what actually happens in the biodiversity community, two things are apparent.

1) Most of the standards are more than 10 years old and paper based. Some are available electronically but the status of the paper and electronic

of classes from the TDWG ontology in the RDF metadata returned for any LSID.

The past year has seen LSID adoption by a number of projects with the biodiversity community to include:

- Species 2000 / Catalogue of Life
- CATE
- ZooBank
- IPNI
- Index Fungorum
- Herb IMI

experimental OAI-PMH support.

This implementation has just been completed and includes functionality handling the majority of the TAPIR specification. Providers have been set up for ZooBank taxon names and HerbIMI specimen data.

PyWrapper

A Python based implementation. LSID authority module nearly completed

PyWrapper?v3.1a includes latest modification to TAPIR protocol during 2007. An installation and configuration manual is being prepared by ETI. The BioCASE network will be switched to PyWrapper v3 once someone is found to maintain PyWrapper (the only TAPIR provider with full ABCD support). Hopefully this will occur during 2008.

mitigate this risk. It is important that any such documentation is specifically targeted. Biologists should not have to delve deeply into computer science issues. Working examples would also be highly useful.

Data Standards

The majority of TDWG members value data standards (such as geographic areas and lists of abbreviations) rather than technical data exchange standards. Data standards should be the core standards of TDWG but they have been neglected for the past decade in favour of the exchange standards and this has led to a degree of alienation within the membership. TDWG could be viable without any exchange standards of its own if it had well maintained data standards and associated applicability statements on the use of appropriate third party technologies. Recommendations on data standards are made below.

versions is not clear. For example, can paper versions of standards be scanned and freely distributed?

2) Some of the most widely used data exchange standards that are associated with TDWG are not ratified. Darwin Core is the classic example of this but the exchange protocols BioCASE and DiGIR should also be considered.

Recommendations

It is the role of the TAG to ensure that a unified picture of TDWG standards is presented to the public. For this to happen two things are required:

1) The status of older and deprecated standards must be clear. For example, the standard "Plant Names in Botanical Databases" from 1995 has the same status as the "Taxon Concepts Schema" from 2005. It appears TDWG recommends HISPID3 over HISPID4 and in parallel to ABCD! How does HISPID3 integrate with other TDWG standards? Should it be integrated? The two courses of action for an older standard are either to dedicate resources to integrating it into the standards architecture or to give it a status indicating it is not recommended for adoption. Resources in TDWG are extremely scarce and so the latter option has to be available.

The Executive Committee must put a mechanism in place to make retirement of standards possible – if only for technical reasons.

TDWG should issue applicability statements for

technologies that were never ratified specifying how people should migrate to their modern equivalents.

2) TDWG must address the issue of 'Data Standards'. Some of the more successful TDWG standards are not data exchange standards but controlled lists of abbreviations or resources such as Index Herbariorum, Authors of Plant Names, TL2 and BPH. These standards are all ratified as paper-based publications. They have all been superseded by on-line versions that have no status within TDWG. None of the on-line versions have interfaces that allow integration into other standards.

TDWG needs a mechanism whereby a dynamic list that is only available through a web service can be a standardised. This mechanism needs to specify technical data access and legal issues.

The Executive Committee should seek out resources to establish this mechanism within the next year as a matter of urgency.

Specific recommendations regarding individual technologies and standards are given in the appendix.

Appendix: Review of Technologies and Standards

TDWG Standards Documentation Specification

Notes: Specifies how standards should be documented under the new TDWG process introduced in 2006. This standard must be ratified first because all other standards are dependent on it. The specification will shortly go for public review.

Recommendation: Speedy executive review and public consultation. This standard is blocking the path for the approval of any other standards.

Responsible Group: Process IG

Access to Biological Collection Data - version 2.06 (ABCD)

Notes: An XML Schema based standard ratified in 2005 that facilitates the exchange of data between natural history collections. ABCD is the most widely used exchange format in Europe. ABCD is widely deployed in the BioCASE network using the BioCASE protocol. ABCD is a more complex equivalent to DarwinCore. ABCD is actively managed and updated.

Recommendation: Reconciliation with DarwinCore and the needs of observations community possibly through use of TDWG ontology. This is the core business of the OSR IG.

Responsible Group: Observation and Specimen Records IG.

BioCASE

Notes: BioCASE is the protocol used to exchange ABCD documents in the BioCASE network. BioCASE has never been ratified as a TDWG standard and is likely to be replaced by TAPIR.

Recommendation: TAPIR should be used in preference to BioCASE in new networks. An applicability statement or other document should be produced to aid migration to TAPIR.

Responsible Group: Technical Architecture Group

Darwin Core

Notes: There are several versions of the XML Schema based exchange standard. Dwc is the most widely used format but it has never been ratified as a TDWG standard. A unified version, with extensions is ready to be submitted. The vast majority of the fields in the schema are represented in LSID Vocabularies as 1:1 mappings.

Recommendation: A decision is needed as to whether to incorporate entirely within the ontology or propose as a separate standard with links to ontology. This must be reconciled with ABCD and the needs of observations community.

Responsible Group: Observation and Specimen Records IG.

DiGIR

Notes: The exchange protocol used to serve DarwinCore and other federation schemas that have been bound to it. It has never been ratified as a TDWG standard. It is now being replaced by TAPIR.

Recommendation: TAPIR should be used in preference to DiGIR in new networks. An applicability statement or other document should be produced to aid migration to TAPIR.

Responsible Group: Technical Architecture Group

Structured Descriptive Data

Notes: SDD is an XML Schema based standard ratified in 2005 that facilitates the encoding of taxonomic descriptive data and diagnostic keys. SDD is a replacement for the DELTA language. SDD is actively supported and there is discussion on producing RDF compatible SDD-Lite in the future.

Recommendation: The relationship of SDD with the new architecture needs to be explored but has not been resourced up to this point. Possible production of 'lite' version for use with RDF if supported by user requirements.

Responsible Group: Descriptive Data IG and Technical Architecture Group

Taxonomic Concept Transfer Schema

Notes: An XML Schema based standard ratified in 2005 that facilitates the transfer of nomenclatural and taxon concept data. TCS promotes the separation of nomenclature and taxonomy in data exchange and is represented almost entirely in the LSID Vocabularies of the TDWG Ontology. Most deployments of TCS use the ontology representation rather than one based on the XML Schema.

Recommendation: Full documentation of the relationship with the version based on the TDWG Ontology is required. Possible version 2 of standard that specifies use of ontology.

Responsible Group: Taxon Names and Concepts IG

Herbarium Information Standards and Protocols for Interchange of Data (HISPID3)

Notes: HISPID3 is non-XML, flat file based exchange standard ratified in 1996 that has been replaced by HISPID4 (which is not ratified by TDWG).

Recommendation: A clear statement on which version of HISPID TDWG recommends (if any) should be on the standards cover page.

Responsible Group: Observation and Specimen Records IG.

Economic Botany Data Collection Standard

Notes: A book ratified as a standard in 1995.

Recommendation: Establish an Economic Botany Interest Group or a statement by Executive Committee on the future of Economic Botany standards in its absence.

Responsible Group: Executive Committee

Plant Occurrence and Status Scheme

Notes: POSS is a controlled vocabulary of terms ratified as a standard in 1995 that is used as a lookup tables in curation databases.

Recommendation: These controlled vocabulary terms need to be made accessible as URIs as part of the TDWG Ontology for integration with legacy systems.

Responsible Group: Geospatial IG

Plant Names in Botanical Databases

Notes: Recommendations for storing botanical names in databases ratified as a standard in 1994. This work is largely superseded by Taxon Concept Transfer Schema and its associated documentation. It may form the basis of older curatorial database schemas.

Recommendation: A clear statement is needed on the standards cover page as to whether TDWG

recommends the use of this standard.

Responsible Group: Taxon Names and Concepts IG.

Authors of Plant Names

Notes: A book of author abbreviations ratified as a standard in 1992 and maintained as a database accessible through. <http://www.ipni.org>. The on-line version is not ratified by TDWG. This is a potential TDWG data standard.

Recommendation: A data standards mechanism is required within TDWG so that a dynamic version of this standard can be formalised.

Responsible Group: Technical Architecture Group

World Geographical Scheme for Recording Plant Distributions

Notes: A list of geographic regions widely used in curation of databases and ratified as a standard in 1992. The boundaries of these regions have been defined as ESRI Shape files but these files are not ratified. The regions have been described as part of the TDWG Ontology <http://rs.tdwg.org/ontology/voc/GeographicRegion>.

Recommendation: Two new standards should be proposed to formalise the boundaries and the vocabulary version of the codes if it can be shown that these are needed for the integration of legacy data.

Responsible Group: Geospatial

XDF - A Language for the Definition and Exchange of Biological Data Sets

Notes: An early XML based standard now deprecated.

Recommendation: Clear statement needed on the standards cover page as to whether TDWG recommends the use of this standard.

Responsible Group: Executive Committee

Botanico-periodicum-huntianum and its supplement.

Notes: There is one standard the original publication and one for the supplement. These are books that provide standard abbreviations for 12,000 journals dealing with plants and approximately 12,000 non-standard abbreviations for those same titles found in other works. Abbreviations appear to be freely available through the Harvard University Herbaria http://asaweb.huh.harvard.edu:8080/databases/publication_index.html.

Recommendation: A mechanism is required within TDWG so that it is clear how a dynamic version could be standardised. The legal status of the information in the books may need clarifying. Can it be freely distributed in its entirety?

Responsible Group: Literature & Executive Committee

Index Herbariorum. Part I: The Herbaria of the World

Notes: A book giving standard abbreviations for herbaria and ratified in 1990. This work is on-line by

New York Botanic Gardens at <http://sweetgum.nybg.org/ih/>.

Recommendation: A mechanism is required within TDWG so that it is clear how the dynamic version could be standardised. A clear statement on the standards cover page is required about TDWG's recommendation about the paper and dynamic online version.

Responsible Group: Natural Collections Descriptions IG

International Transfer Format for Botanic Garden Plant Records

Notes: A non-XML based transfer format ratified in 1987 that has been widely used in botanic gardens community. Some of the controlled vocabularies within ITF2 have been adopted as lookup tables in curation databases.

Recommendation: The controlled vocabularies need to be made available as URIs as part of the TDWG Ontology so that they can be efficiently synonymised with any new standards.

Responsible Group: Observation and Specimen Records IG.

Floristic Regions of the World

Notes: A book ratified in 1986 of floristic regions of the world. This work could be useful if it was made available in an electronic format.

Recommendation: An individual needs to step forward to champion the on going maintenance of the standard or the IG needs to make a statement about its future on the standards cover page.

Responsible Group: Geospatial IG.

User's Guide to the DELTA System

Notes: DELTA, ratified in 1986, is a description Language for Taxonomy that has been widely implemented. DELTA has largely been superseded by SDD and proprietary formats. It is very unlikely that this standard corresponds to any live system.

Recommendation: A clear statement on which version of DELTA TDWG recommends (or none at all) on the standards cover page.

Responsible Group: Descriptive Data IG

Taxonomic Literature, ed. 2 and its Supplements

Notes: A series of books containing "A selective guide to botanical publications and collections with dates, commentaries and types". Now available on-line (<http://tl2.idcpublishers.info/>) by subscription or free to IAPTA members.

Recommendation: Clarification is needed concerning access by TDWG members and this should be on the standards cover page. Can TDWG have a standard that is not 'freely' available to its members? Abbreviations may now be freely available through the Harvard University Herbaria (http://asaweb.huh.harvard.edu:8080/databases/publication_index.html). Clarification of whether a data

standard is required here.

Responsible Group: Literature IG & Executive Committee

Natural Collections Descriptions (NCD)

Notes: An emerging standard for the description of biological collections that resulted from a collaboration between the European Union SYNTHESIS and RAVNS. The standard is being developed as a XML Schema that is integrated with the LSID Vocabularies.

Recommendations: Demonstration providers and documentation needs to be finalised and the standard submitted

Responsible Group: Natural Collections Descriptions IG

TDWG Access Protocol for Information Retrieval (TAPIR)

Notes: TAPIR is a unification of the BioCASE and DiGIR protocols that is being rolled out across both the BioCASE and DiGIR networks. Full implementations of the protocol can support custom response formats. TAPIR is capable of serving RDF that uses the TDWG ontology. A specification will be submitted to the standards process in the near future.

Recommendations: TAPIR is recommended as an XML Schema base exchange protocol.

Responsible Group: Technical Architecture Group

Life Science Identifiers (LSID)

Notes: Life Science Identifiers have been proposed as the preferred GUID technology for the key objects within the TDWG domain. A standard is in preparation that specifies how the LSIDs should be applied within the biodiversity informatics domain. As LSIDs are an OMG standard, TDWG's recommendations will take the form of an Applicability Statement.

Recommendations: LSIDs should be implemented (in accordance with the LSID applicability statement) for classes of object that are not already issued with resolvable URIs such as DOI.

Responsible Group: Technical Architecture Group

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Notes: The primary use of DiGIR and BioCASE providers is to expose data sources for harvest by caching indexers such as GBIF. The same outcome can be achieved using a simpler protocol from the wider community such as OAI-PMH. This was accepted as a strategy in the original TAG1 meeting but did not appear in the 2006 Roadmap document. A test service has now been implemented using the TDWG Ontology as the metadata format and as part of the TAPIR.NET software.

Recommendations: Full documentation on recommendations for use in the form of an Applicability Statement should be proposed as a standard.

Responsible Group: Technical Architecture Group