



Building a Global Biology Digital Library
Progress toward taxonomic data standards:
The 2001 Taxonomic Data Working Group
Meeting

P. Bryan Heidorn
University of Illinois

Presented at the Illinois Natural History Survey
December 17, 2001

International Working Group on Taxonomic Databases (TDWG)

The Taxonomic Databases Working Group
(TDWG) was founded in 1985:

"to establish international collaboration
among biological database projects so as to
promote the wider and more effective
dissemination of information about the
World's heritage of biological
organisms"—[TDWG Constitution, Art. 1.]

2001 Meeting of TDWG

Sydney, Australia

9th November to 11th November

hosted by the Royal Botanic Gardens,
Sydney.

A Call to Revise the TDWG Standards Development Process

. Position Paper by Stan Blum

California Academy of Sciences

Chairperson TDWG

http://www.tdwg.org/process/tdwg99_blum.html

Three Standards Types

- Semantic
- Exchange
- Vocabulary or Authority

Semantic

- meanings of words or symbols
- data dictionary or an information model

Exchange

- Semantic and Syntactic components
- HISPID3

Vocabulary or Authority

- vocabulary standard specifies a set of values that are mutually understood among a group of people specifies the valid values of an attribute
- Authority List of Scientific Names
<http://www.tdwg.org/plants.html>
- Authority of Author Names
- Abbreviations
- Character Lists and Values (Very difficult!)

JOURNAL/PERIODICAL ABBREVIATIONS:

- Lawrence et al. 1968. *Botanico-periodicum-huntianum*. Hunt Institute for Botanical Documentation, Pittsburgh. (Current electronic version) TDWG Standard.
- Bridson, G.D.R. & E.R. Smith 1991. *Botanico-periodicum-huntianum/supplementum*. Hunt Institute for Botanical Documentation, Pittsburgh. TDWG Standard.

Some Standards in Use

- **AUTHORS:** Brummitt, R.K. & C.E. Powell 1992. Authors of plant names. Royal Botanic Gardens Kew. (Current electronic version) TDWG Standard.
- **DATA TRANSFER FORMAT:** Botanic Gardens Conservation Secretariat 1987. International transfer format for botanic garden plant records. Plant Taxonomic Database Standards No. 1, Hunt Institute for Botanical Documentation, Pittsburgh. TDWG Standard.

DELTA

- Dallwitz, M.J. & T.A. Paine 1986. Users guide to the DELTA system. CSIRO Division of Entomology Report No. 13, pp. 3-6. TDWG Standard.

STANDARDS IN ELECTRONIC FORMAT:

- Conn, B., Editor 1996. HISPID3. Herbarium Information Standards and Protocols for Interchange of Data. Version 3. Royal Botanic Gardens. Sydney. TDWG Standard.
<http://www.rbgsyd.gov.au/HISCOM/HISPID/HISPID3/hispidright.html>

FLORISTIC REGIONS & GEOGRAPHICAL CODES:

- Takhtajan, A. 1986. Floristic regions of the world, pp. vii-xiii. University of California Press. TDWG Standard.
- Hollis, S. & R.K. Brummitt 1992. World geographical scheme for recording plant distributions. Plant Taxonomic Database Standards No. 2, Hunt Institute for Botanical Documentation, Pittsburgh. (Current electronic version) TDWG Standard. Second Edition 2001

New Initiatives

- CODATA Working Group on Biological Collection Data Access: A joint CODATA and TDWG initiative
- Structure of Descriptive Data group

Biological Collection Data Access

- Subgroup on Content Definition
- Subgroup on Protocol Development
- design documents:

<http://digir.sourceforge.net>

Biological Collection Data

Access: Content Description

- Charles Copp presented an initial schema
- Based on the BioCISE information model (<http://www.bgbm.org/biodivinf/docs/CollectionModel/>) and the British NBN/Recorder model
- Berendsohn, W. G., et al.. (1999): A comprehensive reference model for biological collections and surveys. *Taxon* 48: 511-562.
<http://www.bgbm.org/biodivinf/docs/CollectionModel/ReprintTNR.pdf>

Content Description

<http://digir.sourceforge.net/prot/darwin3.xsd>

```
<?xml version="1.0"?>
```

```
<xsd:schema
```

```
  targetNamespace="http://www.namespaceTBD.org/darwin3" xmlns:digir="http://www.namespaceTBD.org/digir4"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns="http://www.namespaceTBD.org/darwin3"
  elementFormDefault="qualified">
```

```
<xsd:import
```

```
  namespace="http://www.namespaceTBD.org/digir4"
  schemaLocation="http://digir.sourceforge.net/prot/digir4.xsd"/>
```

Content Description Elements

```
<xsd:element name="ScientificName" type="xsd:string"  
  substitutionGroup="digir:alphaSearchCondition"/>
```

```
<xsd:element name="Kingdom".../>
```

```
<xsd:element name="Phylum".../>
```

```
<xsd:element name="Class".../>
```

```
<xsd:element name="Order".../>
```

```
<xsd:element name="Family".../>
```

```
<xsd:element name="Genus".../>
```

```
<xsd:element name="Species".../>
```

```
<xsd:element name="Subspecies".../>
```

Content Elements

InstitutionCode, CollectionCode, CatalogNumber,

```
<xsd:element name="Collector" type="xsd:string"  
  substitutionGroup="digir:alphaSearchCondition">
```

```
<xsd:annotation>
```

```
<xsd:documentation>this value could be a list of  
  collectors, specifically in result  
  instances</xsd:documentation>
```

```
</xsd:annotation>
```

```
</xsd:element>
```

Longitude

```
<xsd:element name="Longitude"  
  substitutionGroup="digir:numericSearchCondition">  
<xsd:annotation>  
<xsd:documentation>a decimal in the range of -180.0 to  
  180.0</xsd:documentation> </xsd:annotation>  
<xsd:simpleType> <xsd:restriction base="xsd:decimal">  
<xsd:minInclusive value="-180.0"/>  
<xsd:maxInclusive value="180.0"/>  
</xsd:restriction></xsd:simpleType>  
</xsd:element>
```

Future work on the schema

- will evolve through the work of experts on specific "subschemas" such as botanical names or geography.
- Strong typing will be used.
- Short element names will be used.
- Syntactically identical entities will be represented with a single element definition
- Namespaces of other communities will be used as far as possible.

Protocol Development

- Messages between "Portal" and "Provider"
- The primary purpose of the provider is to identify subsets of information identified by a client provided filter and return some portion of that content in a format requested by the client.

Request Structure

<!ENTITY request (header, operation)>

<!ATTLIST response version #REQUIRED>

<!ENTITY header (sendTime,sourceID,
destinationID)>

<!ENTITY sendTime #PCDATA>

<!ENTITY sourceID #PCDATA>

<!ENTITY destinationID #PCDATA>

<!ENTITY operation ANY>

Request Structure Definitions

- ElementDescription
- request -The top most element of any Request message
- header Contains information generic to all types of Request and Response messages.
- version The version of the service that the message is being directed to. The first valid version identifier will be "1.0.0".

Request

- **sendTime** The time (GMT) formatted according to ISO-8601 that the request message was generated. The actual value of sendTime should indicate as close as practicable when the message is sent.
- **sourceID** An identifier for the source of the message. Where the message is machine generated, this should be the IP address of the machine that generated the message.

Request

- destinationID: The IP address of the target of the message.
- Operation: The operation element identifies the operation that the provider is being requested to perform and the parameters required to perform that operation.

The type attribute of the operation element indicates the requested operation. There are currently two types of operations defined: "search" and "status".

Request Example

```
<request><header>  
<sendTime>1988-04-07T18:39:09Z</sendTime>  
<sourceID>129.237.201.230</sourceID>  
<destinationID>129.237.201.228</destinationID>  
</header>  
<operation type="operationName">  
... operation definition ...  
  </operation>  
</request>
```

Response Structure

<!ENTITY response (header, content, diagnostics)>
<!ATTLIST response version #REQUIRED>
<!ENTITY header (sendTime, sourceID, destinationID)>
<!ENTITY sendTime #PCDATA>
<!ENTITY sourceID #PCDATA>
<!ENTITY destinationID #PCDATA>
<!ENTITY content ANY>
<!ATTLIST content type #REQUIRED>
<!ENTITY diagnostics (diagnostic*)>
<!ENTITY diagnostic #PCDATA>
<!ATTLIST diagnostic code #REQUIRED>
<!ATTLIST diagnostic severity #REQUIRED>

Operations

Search Structure

<!ENTITY operation (dbName,filter, recordStruct)>
<!ATTLIST operation type (search | scan) #REQUIRED>
<!ENTITY dbName #PCDATA>
<!ENTITY filter (LOP | COP)>
<!ENTITY LOP (COP, COP)>
<!ATTLIST LOP type (and | or | andNot | orNot) #REQUIRED>
<!ENTITY COP (term, concept)>
<!ATTLIST COP type #REQUIRED>
<!ENTITY term #PCDATA>
<!ATTLIST term type #IMPLIED>
<!ENTITY concept>
<!ENTITY records #PCDATA>
<!ATTLIST records start #IMPLIED>
<!ATTLIST records count #IMPLIED>

Operations

Search Structure

<operation type="search">

<dbName>birdCollection</dbName>

<filter>

<comparison type="3">

<concept>14</concept>

<term type="2">1970</term>

</comparison>

</filter>

<records>

<format>

<http://speciesanalyst.net/DarwinCore/1.0/recordStruct/full.xsd>

Search Response Structure

- <!ELEMENT content (structure, record*, moreRecords)>
- <!ATTLIST content type (search | status) #REQUIRED>
- <!ELEMENT structure ANY>
- <!ELEMENT record ANY>
- <!ELEMENT moreRecords #PCDATA>

Structure of Descriptive Data Challenge Cases

http://www.tdwg.org/tdwg2001/SDD_TDWG_2001.htm

1. Represent basic descriptions of two taxa from one (natural language) source
2. Basic description of one taxon, two sources, basic markup
3. One treatment, several contributing authors
4. Define basic character list
5. Remote vs local character list

SDD Challenge Cases

6. Remote and local taxon lists
7. Several descriptions at different levels (e.g. family (genus (species (specimens))))
8. Universal treatment, several outputs (natural language, NEXUS, interactive key)
9. Representation of legacy structured descriptive data

SDD Challenge Case 1

1. Represent basic descriptions of two taxa from one (natural language) source

Discaria pubescens Rigid, spreading shrub to c. 1m high and wide; stems glabrous. Leaves soon deciduous (particularly on older plants), +/- oblong, (4-)6-10(-15) mm long, 2-3 mm wide, apex obtuse or minutely mucronate within an apical notch, glabrous or a few hairs present near tip; stipules dark reddish-brown, c. 1 mm long, often shallowly joined around the node; spines stout, 1.5-4 cm long

SDD Challenge Case 1

- *Discaria nitida* Slender shrub, to 5 m high; stems glabrous. Leaves persistent (rarely deciduous), elliptic to obovate, (8-)10-20(-30) mm long, 3-7 mm wide, glabrous, shining; spines not developed at each node, to c. 1 cm long.

Issues

- missing data (e.g. stipule characters for *D. nitida*)
- modifiers (e.g. leaves of *D. nitida* rarely deciduous)
- freeform comments (e.g. *D. pubescens* "particularly on older plants")
- structured data vs marked-up text (e.g. let's say that stipule arrangement will not be a character - how to retain the data "stipules often shallowly joined around the node")

Representation option (Kevin Thiele)

<Document>

<Description Name="Discaria pubescens">

<Feature Name="Habit">Rigid, spreading

<Feature Name="Life form"> <Value>shrub</Value>
</Feature>

<Feature Name="Height"> to c. <Value MinValue="0"/>
<MaxValue>1</MaxValue> <Units>m</Units> high
and wide </Feature> </Feature>;

<Feature Name="Stem indumentum">stems
<Value>glabrous</Value></Feature>.

Representation option (continued)

```
<Feature><Name>Leaves</Name><Feature  
Name="Longevity">soon  
  <Value>deciduous</Value>(particularly on older  
plants)</Feature>, <Feature Name="Shape">+/-  
<Value>oblong</Value></Feature>, <Feature  
Name="Length">( <OutMinValue>4</OutMinValue  
>-) <MinValue>6</MinValue>-<MaxValu  
e>10</MaxValue>(-  
  <OutMaxValue>15</OutMaxValue>)<Units>  
mm</Units>
```

SDD Challenge Case 2

2. Basic description of one taxon, two sources, basic markup
 - referencing of source (i.e. this piece of data comes from this article)
 - ascription of data (i.e. this piece of data was supplied/interpreted by/observed by this contributor)