

# GBIF Species Model Workshop

## Introduction

GBIF and TDWG have identified a pressing need to develop a standardised species data model to complement those already available for specimens and observations (Darwin Core and ABCD Schema). Several initiatives have already begun to model species level data and there is a need to bring them together to reach consensus and avoid fragmentation. To start this process, a species model workshop took place at the GBIF Secretariat in Copenhagen from 16 to 18 April 2007 and was attended by representatives from several species modelling initiatives (Plinian Core, GISIN, Nature Serve, FishBase/SeaLifePortal, ETI Informatics), the EDIT and CATE eTaxonomy projects, TDWG, and GBIF.

## Objectives of Workshop

The species model is intended to be a specification of data concepts and structure intended to support the retrieval and integration of data that documents species. The objective of the workshop was to develop a top-level categorisation of species data suitable for the content served by different projects, building on work already done. It is expected that individual specialist groups would then extend this top level categorisation to accommodate their own particular requirements. Several uses cases are envisaged, amongst them:

1. a human agent wishes to aggregate content from several pages dealing with the same species to develop a synthesised view;
2. integrating "portals" like Encyclopedia of Life need a uniform species model to allow them to process in a scalable way the millions of species pages available on the web;
3. automated/semi-automated retrieval and aggregation of content;
4. facilitate more intelligent searching;
5. expression of species model as RDF for Semantic Web applications;
6. facilitate generation of RSS/Atom feeds for alerting and update services.
7. One website publishes taxonomic and biological information about taxa from a particular group (Family, Superfamily, Order, Kingdom). Some of these taxa have homonyms which lie outside of the group under study. This website would like to be able to federate resources from other specialist sites, rather than create content of its own. It does not necessarily know that resources exist for each homonym, but would like to be able to discover such resources if they are available.
8. A community site for horticulturalists wishes to access information about species of a particular group. They are only interested in a subset of the total possible information about a plant, in particular, its natural habitat, climate, range of temperature, soil acidity, moisture, and altitude to enable their members to cultivate such plants.
9. An evolutionary ecologist would like to test whether there is a trade-off between two traits (as a trivial example, reproductive method and population growth rate) within a large group of organisms (for example, higher plants). They need both trait values for various taxa, and their classification.

## Day 1 Afternoon

Each of the five major species modelling projects provided a quick run through of the main elements in their schemas. The presenters were Paco Pando (Plinian Core), Leslie Honey (Nature Serve), Michael Browne (GISIN), Nicolas Bailly (SeaLifePortal), Wouter Addink (ETI Bioinformatics). All take a document centric approach using XML schema (XSD). Following the overviews, Éamonn Ó Tuama presented a consensus list of the top level elements derived from the models, which participants had contributed through email discussions in the weeks prior to the meeting. Some general discussion took place around IPR issues, the need for good image metadata (e.g., to distinguish diagnostic images from general images), how to deal with both general and atomised statements, and targeting audiences. Plinian Core includes a "target audience" element whereas NLBIF dispenses with target audiences. Nature Serve has recently had to face this issue, particularly for children.

We then addressed general modelling issues. Roger Hyam spoke on the object-based vs document centric approach to modelling and, in particular, the TDWG Technical Architecture Group's adoption of objects to better

integrate with the Semantic Web. He described the Universal Biodiversity Data Bus (UBDB) which provides the underpinning architecture and explained how, while XSD documents are good for transfer protocols, they are more difficult to use on the UBDB because mapping between one schema and another is often problematic. The solution is to model using classes and their properties (to keep track of what is going on) and convert to documents when required, an “object oriented documents” approach.

Bob Morris spoke briefly about the GISIN Invasive Alien Species Profile Schema (IASPS) . The schema is extensible, had to allow constraints on use of vocabularies (legal constraints), but also allow data providers to supply their own terms (in defined schemas). TAPIR is used for transport although PyWrapper does not handle recursion well.

Ben Clark gave an overview of the CATE eTaxonomy project and provided some insights gained from modelling such disparate taxa as Araceae and Sphingidae. He stressed the need for inclusion of Natural Language Data in the exchange standard as this would be demanded by many users. In order to allow a user to specify the subject of their data in a flexible and extensible way, he suggested that it would be better to specify the subject of an element by annotation of the element with terms from a controlled vocabulary which can be added to or extended rather than creating an element for every subject. This strategy was adopted for the species model that we subsequently developed.

## Day 2 Full Day

Roger Hyam presented an outline of the proposed model. Described in simple terms, it would provide a root element that referred to a species or taxon concept under which one or more “fact” elements about the species would be listed. Each fact would be of a particular category, drawn from a list of general terms with the possibility of linked controlled vocabularies for each of these terms. There would be provision for including Dublin Core type metadata and for linking out to other standards such as TCS, SDD, Darwin Core / ABCD Schema.

Questions that came up in the initial discussion included:

1. Is our target publications or databases? (aim for publications first)
2. Are we dealing with just terminal taxa or also higher (aggregated taxa)?
3. Do we deal with just intrinsic (coded in genome) species “facts” or also include extrinsic (environment related); most felt that we need to cover both.

The workshop then split into two groups:

1. Domain experts group (Michael Browne, Leslie Honey, Paco Pando, Nicolas Bailly, David Patterson, Andreas Kohlbecker, David Remsen, Éamonn Ó Tuama)
2. Modelling Group (Roger Hyam, Ben Clark, Bob Morris, Wouter Addink, Donald Hobern)

While the modellers worked on defining the species model and its expression (in RDF), the domain experts group took the list of consensus elements and proceeded to create a set of top level categories. For each category, they defined an element name, a human readable name, and a definition (semantics). They also tried to identify controlled vocabularies for some of the categories.

## Day 3 Morning

Roger Hyam presented the species data model now generalised as the Taxon Data Model. Some discussion took place about where to include metadata – under the root element or allow in each element? We then worked through some examples of marking up species data according to the Taxon Data Model, e.g., a FishBase instance and an invasive species example. This proved particularly helpful in gaining a better understanding of the model.

The final part of the workshop was devoted to a discussion on setting up a TDWG Interest Group for the Taxon Data Model. Roger explained the process and procedures involved in doing this. The group agreed to proceed with it with the goal of creating a TDWG standard by next September. We identified particular users of the taxon model: SpeciesBase, Encyclopedia of Life and GISIN. The GISIN network would also provide a suitable example to demonstrate the Taxon Data Model.