

GBIF Species model Workshop


This are some suggestions for the species model exchange standard being developed by GBIF.

Use Cases

1. 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.
2. 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.
3. 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.

Data Types Required

| Name | Contains | Example | Related standards |
|-----------------|---|--|-------------------|
| Metadata | Metadata about the document | dc:creator A Haigh dc:date 2007-04-01 | Dublin Core |
| Scientific Name | The scientific name of the species + objective nomenclatural data | <i>Arum italicum</i> Mill. urn:lsid:ipni.org:names:142144-3 | TCS |
| Taxonomic Data | Taxonomic data. Subjective taxonomic data e.g. subjective synonyms, child, child taxa etc. | has child: <i>italicum ssp. albispathum</i> <i>italicum ssp. canariense</i> <i>italicum ssp. italicum</i> <i>italicum ssp. neglectum</i> has synonym: <i>Arum albispathum Steven ex Ledeb.</i> | TCS |

| | | | |
|-----------------------|--|---|---------------|
| Distribution | The geographical range this species is found over. Probably needs to be qualified for invasive species etc – also distribution could be in natural language or codified form | Turkey (SW Anatolia). or Codified: TUR. or TDWG “Country” Turkey-in-Europe | OGC standards |
| Coded Data | Data codified according to some controlled vocabulary, or data that is expressed as numerical values or ranges or statistical measures even. | tuber: rhizomatous peduncle: much shorter than leaves peduncle length: 4.6 – 16 cm spathe length: 11 – 27 cm (outside maximum 38 cm) | SDD |
| Natural Language Data | Non-parsed data which a human can understand | Tuber rhizomatous. Spathe-limb internally pale-green to almost white. Spadix ¼ to ½ as long as spathe-limb. Staminate flowers yellow before anthesis. (similar to <i>Arum maculatum</i>) | SDD |
| Biotic relationships | Relationships with other taxa – this would require an attribute to contain a term from a controlled vocabulary which denotes the type of relationship | has pollinator: <i>Psychodidae</i> | ? |
| Multimedia | Data about related multimedia resources such as images, audio and the metadata associated with those resources. |  | |
| References | Data about related documents which might be paper-based or might be electronic. | Boyce, P.C. (1995), The genus <i>Arum</i> (Araceae) in Greece and Cyprus. <i>Ann. Musei Goulandris</i> 9 27-38. or doi: 10.1093/aob/mcj022 or http://www.aroid.org/genera/arum/italicum/ | TDWG-Lit? |
| Unit Data | Data about specific living or dead organisms which | K000400287 : type Iraq, Baquba (Agnew & Hadac) 3581 1961 | ABCD / Darwin |

Rational

- Exchange of data depends upon equivalence of format (structure) of the data to be exchanged and equivalence of the semantics (meaning).
- It is important that the exchange standard includes **Natural Language Data** (unparsed) lots of people will want to exchange this and it is probably the type of data which can be most easily encapsulated within a simple structure. I suspect that everyone will want to exchange natural language data on different subjects.
- Given that users are likely to want to specify the subject of their data in a flexible/extensible way and to discover data which is of interest , it may be better to specify the subject of an element by annotating the element using terms from a controlled vocabulary (ontology) which can be added to or extended rather than creating an element for every subject – for example:

```
<NaturalLanguageData>
  <Metadata>
    <dc:subject>controlledTerm1</dc:subject>
  </Metadata>
  <content>
    Lorem ipsum dolor sit amet, consectetur adipiscing elit.
    Mauris aliquet, magna sit amet tincidunt rhoncus.
  </content>
</NaturalLanguageData>
```

rather than

```
<Subject>
  Lorem ipsum dolor sit amet, consectetur adipiscing elit.
  Mauris aliquet, magna sit amet tincidunt rhoncus.
</Subject>
```

- Beyond natural language data I would suggest that **Coded Data** is important too, but this is more complicated, and depends even more strongly on being able to specify a common controlled vocabulary for characters, states, etc. The SDD standard seems to have covered this (in addition to natural language descriptions).
- **Taxonomic Data** is, in my humble opinion, special enough that it deserves an element of its own. Ideally it would specify an Isid for a taxon concept. Failing that, it may be easier for people to specify at least the taxonomic name (as an Isid, or as a name string). Really this data should follow the TCS format. There's no point developing another standard for this data if TCS is sufficient.
- **Distributional Data** is really very important, especially to non-taxonomists. Distribution can be expressed in several ways, and they all have advantages, so I

think that it might be worth specifying that both natural language data and coded data can be about the distribution of a species. Bottom line is: maybe distribution should be a term in the ontology, and not a specific element, allowing users to extend this to specify distributional data like “distribution as an invasive” “distribution naturally occurring”.

- I separated out **Biotic Relationships** because this might be a special type of coded data which could be catered for with a general model, linking the species “page” to another species page or taxon concept, and using a controlled vocabulary to specify the “type” of biotic relationship (in much the same way as TCS specified a TaxonomicRelationship with a type). Examples from the controlled vocabulary could include:
 - host
 - parasite
 - parasitoid
 - symbiont
 - prey
 - disease of
- It seems to me that **Multimedia Objects, References, and Unit Data** are all dealt with in other exchange standards, but it would be important to be able to express relationships between species and these other resources.

Subjects

Given that other projects have tried to address similar problems (for example: SEEK), it seems excessive to develop a new ontology if an existing one will meet the requirements of this project, but I’m in no way an expert on this sort of thing. Useful general terms (I’ve indented some to indicate a includes/ is included in relationship) could be

- Biology
 - Genetics
 - Molecular Biology / Biochemistry
 - Physiology
 - Morphology
 - Ecology
 - Habitat
 - Behaviour
 - Population Ecology (Demography)
 - Community Ecology (Prey/Predators)
 - Biogeography
 - Distribution
 - Evolution
- Conservation Biology
 - Invasiveness
 - Abundance

- Management
- Economic Uses

I've also tried to classify the descriptions in CATE according to an ontology of terms as an example

CATE-Sphingidae

These are terms in the natural language description of the hawkmoths

Morphology

Male

Female

Body

Forewing

Hindwing

Upperside (of body part)

Underside (of body part)

Length (characteristic of body part)

CATE-Araceae

These are terms in the natural language description of the Aroids

Petiol

Leaf-blade

Peduncle

Spathe

Spadix

Altitude

Phenology

Habitat

Etymology

Uses

Conservation Status