Taxonomic Markup Language: Applying XML to Systematic Data

R. Gilmour
Science Library
University at Albany
Albany, NY 12222

Running head: Taxonomic Markup Language

Keywords: XML, taxonomy, systematics

Abstract
Summary: An XML document type definition (DTD) is provided for the description of taxonomic relationships between organisms. Two XSL stylesheets for the graphical presentation of simple taxonomic trees are also provided.
Availability: The DTD and stylesheets along with sample files are available at http://www.albany.edu/~gilmr/pubxml/.
Contact: gilmr@cnsvax.albany.edu

Within the context of the late-20th century information explosion, it has become imperative to format data in ways that facilitate its interchange (see Murray-Rust, 1998). XML (Extensible Markup Language) has provided a means of doing this, using simple text-based markup to describe the structure and semantics of ordered data (Bray et al., 1998). To further facilitate standardization, XML files may be linked to Document Type Definitions (DTDs), which describe the set of markup elements to be used. These DTDs serve as standards within particular disciplines or industries, allowing files to be freely interchanged among users of a DTD and allowing applications to be built which can parse the resulting files in accordance with the DTD. This allows XML data to be passed to databases, spreadsheets, graphics packages, and other applications, making XML an excellent candidate for a nearly universal data format. Freely available validating parsers can be used to check an XML file for conformance to a particular DTD. Within the sciences, XML DTDs have been developed for chemistry publications (Murray-Rust, 1999) and for biological sequence data (Fenyö, 1998 and 1999).

Systematists make use of biopolymer data and codified morphological data to reconstruct the branching phylogenies of groups of organisms. The hierarchical structure of these phylogenies lends itself readily to being description by XML. These phylogenies are created by algorithms that seek to minimize unnecessary assumptions (the rule of parsimony) or maximize the probability of a particular phylogeny occurring (the rule of maximum likelihood). A phylogenetic tree is therefor not a simple statement, but must be qualified by a wide variety of statistics which describe the degree to which the original data is explained by the tree and the statistical strength of particular configurations (Kitching et al., 1998).

In addition to generating “raw” phylogenies (the visual representations of which are variously known as dendrograms or cladograms), systematists also attempt to provide a nomenclatural system which, at least to some degree, reflects the patterns of relationship in the phylogeny. The document type definition (DTD) proposed here seeks to accomplish three things:

1. The description of the structure (topology) of a biological phylogeny.
2. The presentation of statistical metadata about the phylogeny.
3. The option of superimposing a Linnean taxonomy upon the phylogenetic structure.

Note that while the DTD allows taxonomic units to be tagged in a Linnean manner, this is optional, since many practicing systematists do not routinely apply this type of classification or do so only at a very late stage in their research. This is especially true of situations in which species boundaries are not clear or in which the nature of sub-specific taxa is open to interpretation.

The following is an example of how a pectinate phylogeny of organisms A, B, C and D could be described:

```
<sys>
  <terminus name="A"/>
  <taxon rank="1">
    <terminus name="B"/>
    <taxon rank="2">
      <terminus name="C"/>
      <terminus name="D"/>
    </taxon>
  </taxon>
</sys>
```

A dichotomously branching topology for the same taxa would look like this:

```
<sys>
  <taxon rank="1">
    <terminus name="A"/>
    <terminus name="B"/>
  </taxon>
  <taxon rank="1">
    <terminus name="C"/>
    <terminus name="D"/>
  </taxon>
</sys>
```

The elements used in the above example include `sys`, which is the root element of any tree described by this DTD; `terminus`, which is used for the end-points of a tree; and `taxon`, which is used for any internal nodes. The `terminus` element takes a required `name` attribute while the `taxon` element takes an optional `rank` attribute. Rank is defined as the number of internodes between the current node and `sys`, so a `taxon` element with a low rank number is more inclusive than one with a high rank number. (Note that the information provided by `rank` is redundant with information inherent in the document structure. It is included for authors who wish to make the levels of branching more explicit or who wish to use XSL stylesheets (see Clark, 1999) or other applications to perform calculations based on the depth of branching.)

The same topology marked up in a Linnean manner with properly named taxa might be expressed:

```
<sys>
  <genus name="Botrychium">
    <species name="Botrychium lunarioides"/>
    <species name="Botrychium virginianum"/>
  </genus>
  <genus name="Ophioglossum">
    <species name="Ophioglossum engelmannii"/>
    <species name="Ophioglossum pussillum"/>
  </genus>
</sys>
```

Statistical data about the tree may be including by using various attributes. If the statistic refers to the tree as a whole, such as the consistency index, it is expressed as an attribute of `sys`: `<sys consistency="0.75">`. If the statistic refers to a particular node (*e.g.* a bootstrap value), it is expressed as an attribute of that node, while statistics referring to internal branches are treated as attributes of the distal node of that branch. Thus, the branchlength of the internode between `sys` and the `<genus name="Botrychium">` node would be expressed within the latter node as `<genus name="Botrychium" branchlength="14">`. Table 1 summarizes the statistics that may be used with this DTD and how they may be applied.

In addition to statistical data about the tree, some context for the tree should be provided. The `sys` element takes a number of optional attributes to aid in this level of description. These include `datatype`, which has allowed values of `molecular`, `morphological`, and `mixed`; and `gene`, which indicates the gene upon which a molecular phylogeny is based. In order to provide a taxonomic context for the tree, all of the Linnean element names may also be used as attributes of `sys` to describe the types of organisms the tree is depicting. So, for example, the sys tag for a phylogeny of the Euphorbiaceae might look like: `<sys kingdom="Plantae" class="Angiospermopsida" family="Euphorbiaceae">`. It is up to the author of the file to determine what level of detail is appropriate here, but at least including the kingdom seems advisable. It is recommended that other contextual metadata be included in a comment or CDATA field at the beginning of the file (see example cited in abstract).

Table 1: Statistics allowed by DTD.

Statistic	Optional attribute of
consistency_index	<sys>
retention_index	<sys>
rescaled_consistency_index	<sys>
length	<sys>
branchlength	any element except <sys>
bootstrap	any element except <sys> and <terminus>
jacknife	any element except <sys> and <terminus>
decay	any element except <sys> and <terminus>
synapomorphies	any element except <sys> and <terminus>

References

Bray,T.,Paoli,J. and Sperberg-McQueen,C.M. (1998) Extensible Markup Language (XML) 1.0. Available at http://www.w3.org/TR/1998/REC-xml-19980210.html.

Clark,J. (1999) XSL Transformations (XSLT) Version 1.0. W3C Proposed Recommendation 8 October 1999. Available at http://www.w3.org/TR/xslt/.

Fenyö,D. (1998) The Biopolymer Markup Language. Available at http://www.proteometrics.com/BIOML/.

Fenyö, D. (1999) The Biopolymer Markup Language. *Bioinformatics* 15(4): 339-40.

Kitching,I.J., Foley,P.L., Humphries,C.J and Williams,D.M. 1998. *Cladistics: the theory and practice of parsimony analysis*. 2nd ed. Oxford University Press.

Murray-Rust,P. (1998) The Globalization of Crystallographic Knowledge. *Acta Crystalographica* D54: 1065-1070.

Murray-Rust,P. (1999) XML-CML.ORG - The Site for Chemical Markup Language. Available at http://www.xml-cml.org/.