2014 IEEE 10th International Conference on eScience

From Phenotypes to Trees of Life: A Metamodel-driven approach for the integration of Taxonomy models

Anaïs Grand
Muséum national d'histoire naturelle
CR2P - UMR 7207 CNRS, MNHN, Univ Paris 06
57 rue Cuvier, CP48 - F-75005, Paris, France
Email: grand@mnhn.fr

Régine Vignes Lebbe
UPMC Univ Paris 06
CR2P - UMR 7207 CNRS/MNHN/Univ Paris 06
4 Place Jussieu, Tour 46-56, 5ème étage, F-75005, Paris, France
Email: regine.vignes_lebbe@upmc.fr

André Santanchè
IC - Unicamp - University of Campinas
Av. Albert Einstein, 1251, 13083-852, Campinas - SP, Brazil
Email: santanche@ic.unicamp.br

***Abstract*—Several projects aim at gathering together data concerning life around the world, in order to systematize them and produce a big, unified tree of life. Rather than a static single picture of the living world, this kind of tree: (i) is a result of a dynamic interaction among several models produced by biologists for describing life and expressing how life changes and evolves as time goes by; (ii) is not unique, since there are different competing perspectives describing life (morphology, behavior, ecology, genetics etc.) and different methods of reconstructing evolutionary trees. Our work addresses these problems by proposing a "superimposed metamodel" mechanism, which acts as a modeling skeleton, supporting a unified view and articulation of models/ontologies involved in tasks that start at collecting data from the field towards producing descriptions and evolutionary trees. It enables to externalize specific knowledge as ontologies and to trace the entire rationale from one extreme of the process to the other one. This paper shows practical experiments in which we explore such characteristics as: guiding the expression of evolutionary hypotheses from observational data, going backwards on the provenance path, or evaluating changes of the tree in front of new evidences collected in the field.**

***Keywords*-e-biology; phenotype; phylogenetics; metamodel; model integration; ontology**

I. Introduction

Naturalist biologists gather large amounts of information on the biological groups they study. Their work starts at observations on the living world, which are generalized to characterize concepts such as *taxon* – a generalization of groups of organisms – and *character* – an element to describe or characterize taxa. The discipline that studies taxa via their character description is called *taxonomy*, and *systematics* is the discipline that classifies taxa. The classifications built by systematists aim at reconstructing the history of life on earth and the evolution of living beings; such classifications are called *phylogenies* or *phylogenetic trees*. The hypothesis that organisms have a common history (i.e., a common ancestor) and form a cluster in the phylogenetic tree comes from the knowledge yielded by characters. For instance, among plants the concomitant presence of vascular tissues and of a branched sporophyte as the principal generation phase is traditionally seen as inherited by a common ancestor. This combination of features characterizes a particular taxon: the vascular plants.

Building taxon classifications implies making organisms comparable via their characters. Scientific literature can be seen as a bank of characters. Biologists use and reuse published characters so as to describe, compare and classify taxa. In order to know if an author A uses the same character of an author B, the labels of characters are not sufficient. Further than comparing characters labels, it is more important to compare the concepts behind them. As Brazeau [1] emphasized, characters are structured data more than flat textual statements. Flat textual descriptions – as usually adopted by biologists – do not necessarily make explicit all the semantics comprised in a character, since pieces of information remain implicit. As a consequence, the interpretations of scientists are often ambiguous, namely accentuated by a heterogeneous use of the terminology. However, the reproducibility of phylogenetic analyses depends on non-ambiguous interpretations, providing transparency, traceability and enhanced comparison of characters.

Analyses and inferences may require combining and comparing millions of data items. Since data are produced much faster than they can be digested, we pile up a data repository of potential discoveries. Several partial "islands" of data contain complementary evidences, without explicit representation of connections, sometimes being captured only by specialized software that make implicit associations. In order to enable machines to help in the analysis and inference processes, the available data must be integrated in a semantic level. *Semantic* here stands for formal and explicit, more specifically, based on ontologies. This implies making explicit the relations among

978-1-4799-4287-9/14 $31.00 © 2014 IEEE
DOI 10.1109/eScience.2014.48

IEEE computer society

the “islands”. Explicit semantics can be exploited to support keeping track, tracing, managing and comparing different perspectives of researchers, and also making inferences that connect several complementary pieces of data.

Even though related work has been addressing this issue, there are still open problems. One main challenge, which is the focus of this research, is that semantic phenotypic descriptions and phylogenetic trees require the articulation of several preexisting biology ontologies, which were not originally designed to be related. Besides ontologies related to specific aspects of phenotype descriptions – e.g., quality, anatomy and phylogeny – there is a huge volume of domain specific ontologies in biology. This is our main argument here: on one hand, it is not possible to impose the same ontology for everybody, on the other hand, we need a “discipline” to relate existing ontologies. In order to link existing ontologies, providing a unified perspective, we shifted our attention to the metamodel level, to conceive ontology-based modeling primitives specialized in phenotype description, phylogenetic trees and their interrelation. They are designed to lay over existing ontologies and their models – we call this process to superimpose a metamodel – abstracting them in a unified perspective and linking them with explicit relations.

The remaining of the paper is organized as follows: Section II summarizes existing approaches and digital models to represent and manage phenotype descriptions and phylogenetic trees; Section III presents foundations for our approach as well as related work; Section IV presents our proposal of a superimposed metamodel; Section V details a practical application; and Section VI presents our conclusion and future work.

II. From Phenotypes to the Tree of Life

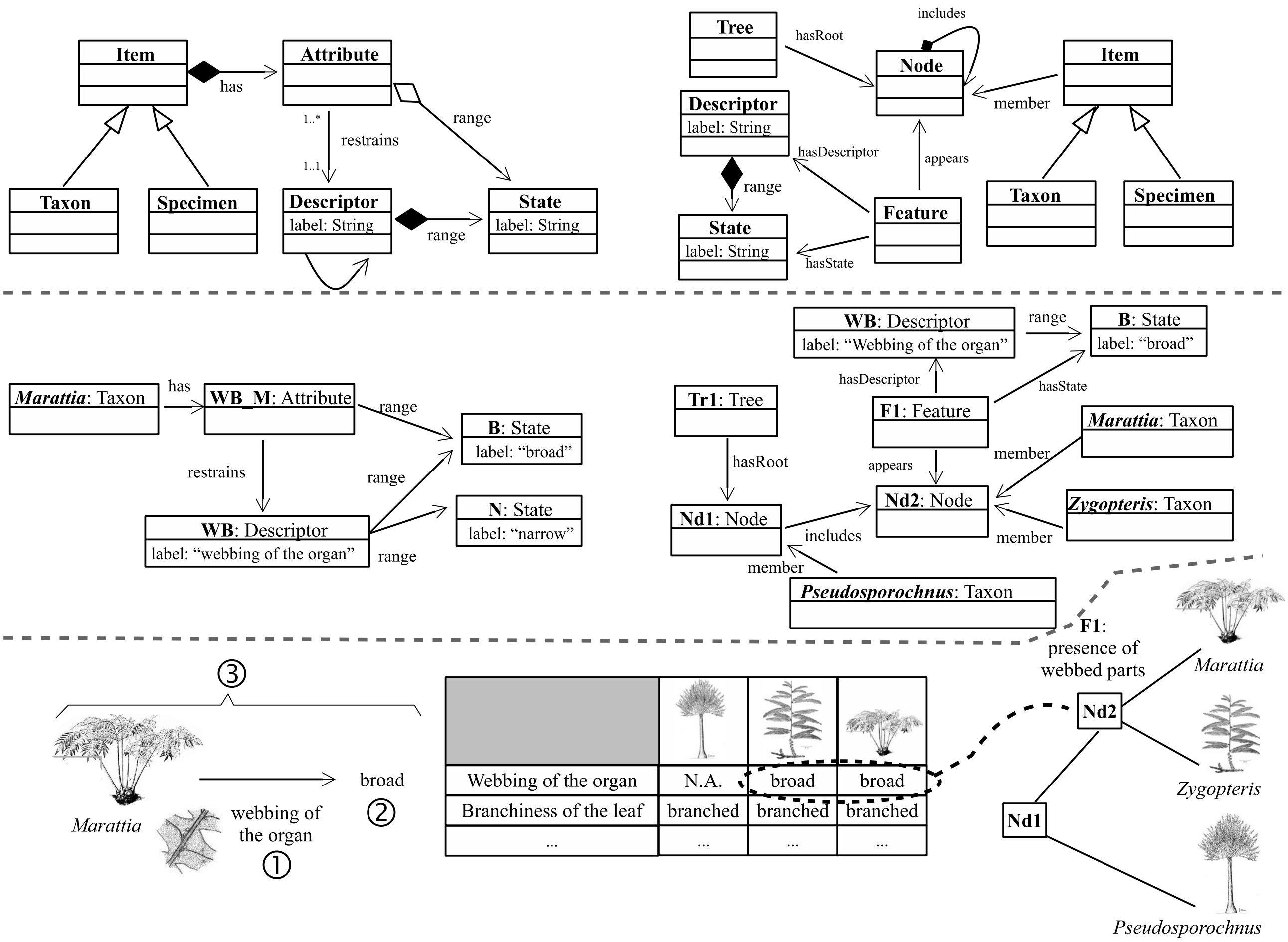
This section summarizes the processes followed by a biologist working on a descriptive and/or phylogenetic model, illustrated in Fig. 1. In order to synthesize these processes in a model of Fig. 1, we combined relevant representation models related to this process, adopted by standards [2], [3] and biology software [4]. The process starts from collecting evidences of living beings from the real world and transforming them in descriptions, going towards generalization of taxa and finally a phylogenetic tree organization. Fig. 1 is organized in three layers. The upper layer presents UML models to describe living beings and to represent phylogenetic trees. The bottom layer shows a practical example of description/classification of plants. The middle layer maps the bottom examples into instances of the upper layer model. The left side of the figure focuses in the phenotype description and the right part in the phylogenetic tree. Even though they are related and information on the left side will be used by a biologist on the right side, we intentionally did not connect the models, thereby emphasizing how they appear in the existing standards and software representations, as unconnected models, in spite of their dependencies.

We now present our practical example – illustrated in Fig. 1 bottom layer (left) – of a botanist describing fern organs [5]. Fern organs can be webbed (i.e., laminated) or not. When leaves are webbed, the flat green part is called a lamina. Traditionally, a phenotype description is composed of sets of ① descriptive statements (e.g., “webbing of the organ”), ② values (e.g., “broad”, “narrow” etc.) and ③ attributions of specific value(s) to an organism (e.g., “the webbing of the organ of the plant *Marattia* is broad”). The phenotype description is represented in the descriptive area by a set of statements “characters” (following [6] terminology) or “descriptors” (following [7] terminology) and their values “character states” or “descriptor states”. At the top of Figure 1 we present a schema of these descriptive primitives, we call `Descriptors`. Besides its label, aimed at human consumption, a `Descriptor` defines a range of possible `States`, which are also characterized by labels. The middle layer presents instances of the schema, representing the “`webbing of the organ`” descriptor and its possible states: “`broad`” and “`narrow`”. These same descriptors/states receive different names in other biology domains. In the evolutionary area, phylogenetic characters indicate the homology, i.e., the sameness relationship between morpho-anatomical entities.

A biologist may systematize relations between descriptive statements, values and fern taxa in a matrix, as illustrated in the center bottom of Fig. 1. Here, fern taxa appear in the columns and descriptors in the rows. A cell value is the state defined for a descriptor attributed to a taxon. “N.A” means “non applicable” and represents the inapplicability condition among descriptive statements.

The model of the top layer (left) in Fig. 1 shows that a given description is designed to be applied to a set of `Items`, which can be individual `Specimens` or `Taxon` entities. The general descriptive primitives are further tailored for each `Item`. An `Attribute` restrains a respective `Descriptor` to accept only a subset of the `States` observed in a related Item. For example, in the middle layer (left), the *Marattia* `Taxon` (an `Item`) constrains the values of the `Descriptor` “`webbing of the organ`” to “`broad`” through an `Attribute`.

This descriptive work carried by biologists can lead to evolutionary studies. The data collected by the biologist in the matrix are used for the construction of a phylogenetic tree. This kind of tree traces the evolution of taxa based on differentiations (i.e., diversifications) expressed within phylogenetic characters, as illustrated in the bottom layer (right). Taxon members of the same node in the tree/character share the same characteristics. The model in the upper layer (right) in Fig. 1 summarizes the main elements of a phylogenetic tree. Phylogenetic trees are made of nested `Nodes`, the most inclusive node being called the Root of the `Tree`. The `Tree` is populated with `Items` connected to the `Nodes`. The `Items` here are conceptually equivalent to the `Items` presented in left side. However, they are intentionally represented apart to emphasize that they are not connected in existing representations, as mentioned in the beginning of this section. A `Feature` (i.e., a `Descriptor State` which is interpreted as a putative characteristic for a group of `Items`) appears in a `Node` of the `Tree`. An `Item` presenting a specific `Feature`



Webbing of the organ	N.A.	broad	broad
Branchiness of the leaf	branched	branched	branched
...	...	...	...

Fig. 1. Models of existing approaches to represent phenotypes and phylogenetic trees.

appears at the corresponding Node. The relationship between Features is seen as being hierarchical by some authors [8]–[10]. As a consequence, Items that are hypothesized to share a Feature, during the Character description, are also represented within a hierarchy. In this particular case, characters share the same hierarchical representation as phylogenetic trees. The relationship between entities/taxa is represented by a hierarchical structure. As illustrated in the bottom layer (right) of Fig. 1: *Marattia* and *Zygopteris* are members of the node characterized by the "presence of webbed parts" whereas *Pseudosporochnus* is connected to the root of the hierarchy. Here, this hierarchy of taxa (*Pseudosporochnus* (*Zygopteris*, *Marattia*)) is a hierarchical phylogenetic character, following [8]–[10].

III. Foundations and Related Work

The model in the upper layer (left) in Fig. 1 is derived from Xper2 [4], a descriptive data management program, which also represents the fundamental descriptive elements of several other description tools, compatible with the Structured Descriptive Data (SDD) standard (http://wiki.tdwg.org/twiki/bin/view/SDD/). The model in the upper layer (right) in Fig. 1 was derived by us from the LisBeth [11] phylogenetic program, which also represents the fundamental tree elements of several phylogenetic applications [12]–[15].

In the Tree, the Feature is linked to the descriptive model elements and assumes an unidirectional interoperability between Xper2 and LisBeth (i.e., Xper2 exports data which can be consumed by LisBeth). With homology hypotheses provided by the biologist, LisBeth automatically reconstructs hierarchical phylogenetic characters from a Xper2 descriptive model. As usual in this context, phenotype description systems are able to produce data (e.g., exporting a file) to be used by phylogenetic systems. The process is unidirectional – from the phenotype descriptions to the phylogenetic tree – the models are not integrated – as mentioned in the beginning of the previous section and depicted in the top layer of Fig. 1 – and each system works in its own subset (i.e., updates in one side will not automatically reflect in the other). The alignment and connection of models is just a first step for integrating phenotypic and phylogenetic data. A rich semantic description is fundamental to support, for example, comparison between hypotheses and consistency checking. We further summarize relevant initiatives, which are tackling this question. In the classical systematics approach, biologists list a set of characters or descriptors for a given living being – usually textual descriptions – and possible states that this character can assume – also textual descriptions. This approach limits the action of computers. Formalizing descriptions through ontologies is an approach which is gaining increasing attention.

Related work aimed to map phenotypical descriptions to ontologies noticed the importance of providing some method in the description process [16]–[19]. In Fig. 2 we synthesize this evolution. In the top, we start by the classical textual based approach, which is previous to any digital system. Inside textual descriptions, as showed by the fragment provided in the figure, biologists refer to description elements (second layer). Most of description digital systems adopted currently have progressed to the *Structured description* layer, devising characters and their states in the descriptions, following the model we presented in the previous section, where we adopted the term descriptor with the same meaning of character here. In this section we use character, as usually referred in the related work.

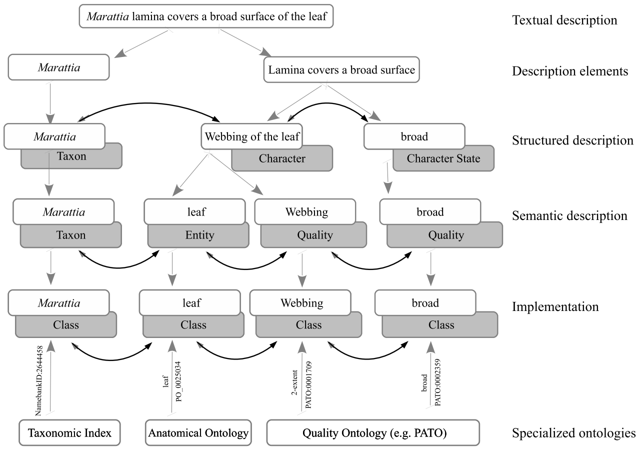


Fig. 2. From textual and structured descriptions to semantic descriptions with specialized ontologies.

The first movement towards the next layer of a *Semantic description* (Fig. 2) – which is now a common perspective – was to migrate from a classical description approach to an "Entity-Quality" (EQ) approach [7], [20]. Continuing the example of the previous section, a character (or descriptor) can be "webbing of the leaf", with character states: "broad" or "narrow". This character mixes two elements: an Entity ("leaf") and a Quality ("webbing"). The Entity is a morphological or anatomical structure that is being observed, a Quality is a property under which the Entity is described. The EQ approach transforms a character into a relation between these two elements.

In order to represent the EQ as ontologies, related work evolved to the *Implementation* layer. Instead of a textual description, they relate each element in the biological description process to an ontology class: an Entity becomes a class of an anatomical ontology of a given organism; a Quality usually becomes a class of the PATO - Phenotypic Quality Ontology (http://purl.bioontology.org/ontology/PATO). Description tools, such as Phenote (http://www.phenote.org) and Phenex [18], follow this perspective allowing the inclusion of more semantics in phenotype descriptions. Despite of the benefits achieved by this stage, related work is aware of its limits. Because classes play distinct roles in this scenario, researchers observed the importance of having some method in the way they relate. Gkoutos and his colleagues [16] proposed a schema, representing it as a diagram and Balhoff and his colleagues [18] presented a descriptive process, which is materialized in their system (Phenex). Elements of the *Implementation* can be related to *Specialized ontologies*, as illustrated in Fig. 1, but there is no formal representation of how these players in the *Specialized ontologies* can be related.

Prosdocimi and his colleagues [15] emphasized the relevance of connecting phenotype descriptions to phylogenetic trees, to supply comparative data analysis. Their formal model focuses on the phylogenetic tree and thus their phenotype description is based in the Character/Character State representation (layer *Structured Description* of Fig. 2). Therefore, there is still an open issue of how representing in a formal way the role played by each element in the ontology based description, their relations and how phenotype descriptions are properly integrated to phylogenetic trees. This work contributes in this sense, by superimposing a metamodel representation, as we detail in the next section. As far as we know, there is no related work able to integrate the complete process in a meta representation, as we propose here.

IV. Metamodeling Phenotypes

Since one of the main tasks in biology concerns systematizing the living world, there are already thousands of taxonomies and ontologies, comprising a wide range of domains – e.g., plants, fishes, mice – and concerns – e.g., anatomy, qualities, phylogeny, description. Therefore, instead of imposing a new unified ontology to integrate everything, we designed a metamodel – as an ontology – to be superimposed on top of existing ontologies, which allow us to connect and integrate them according to their roles. Our metamodel is not meant to be an upper ontology, but rather a (meta)view we project over existing ontologies. It captures the rationale of the process from phenotype description to phylogenetic trees and makes explicit the roles of existing ontology elements, providing a unified abstraction in a metalevel layer and disciplining their relations to integrate them. Beyond related work, it formalizes methodologies to relate ontologies.

We organized the presentation of our approach in Fig. 3 and Fig. 5. As we will detail, Fig. 3 concentrates the metamodel elements and several model elements, which have tight relation with the model. The figures adopted a UML/MOF (MetaObject Facility – http://www.omg.org/mof/) approach to represent metaclasses and classes. In our case, the metamodel and model are represented as part of an OWL ontology. There are differences between the UML object model and the

OWL/RDF model, and the Ontology Definition Metamodel (ODM) [21] is an initiative towards integrating them. Since our representation is in OWL and we are adopting UML based diagrams to visually represent them, in order to simplify the visual presentation, we are adopting the following simplified mapping: UML classes mean OWL classes; UML inheritances are `rdfs:subClassOf` relations between classes; UML instances are `rdf:type` relations between classes and instances; each UML stereotype annotation $\ll mcls \gg$ means that the respective class is instance of a metaclass $mcls$. Each OWL object property is represented by a class with the stereotype $\ll OWLObjectProperty \gg$ and its respective `domain` and `range` as UML associations. In order to represent OWL metaclasses, we adopted four ODM mappings: `OWLClass` `OWLObjectProperty`, `OWLDomain` and `OWLRange`. ,

Entities and Qualities are usually represented as classes in several existing biology ontologies. For example, PATO ontology represents qualities as classes and Plant Ontotology represents the entities as classes. Therefore, our metamodel represents `Entity` and `Quality` as metaclasses – see Fig. 3 up left. In this way, it is possible to define existing ontology classes as instances of these metaclasses – this is the kernel of our superimposition approach, as illustrated in Fig. 4. The metaclass `Taxon` represents any taxonomic classification. We provide classes representing common biology taxonomic classifiers as instances of `Taxon` (Fig. 3 left).

Fig. 4 shows an example of how our metamodel is superimposed on existing ontologies. It is possible to devise two layers: the upper layer has metaclasses of our metamodel; the bottom layer has preexisting or new ontology classes. Preexisting classes are imported from external ontologies, e.g., the `Plant Ontology`'s `Shoot System` class (http://purl.obolibrary.org/obo/PO_0009006) and the `PATO`'s: `2D-extent` class (http://purl.obolibrary.org/obo/PATO_0001709), `Broad` class (http://purl.obolibrary.org/obo/PATO_0002359) and `Narrow` class (http://purl.obolibrary.org/obo/PATO_0000599). In the superimposition process we add a `rdf:type` property from the class to the metaclass, e.g., in order to define `PATO 2D extent` as a $\ll Quality \gg$, we define that it will be an's instance of the Quality metaclass – i.e., a property `rdf:type` from `PATO:2D-extent` to `Quality`. Beyond a semantic characterization of roles, this metaclass association plugs existing ontologies to the overall metamodel. We decided to represent possible states of a given quality class as subclasses of the respective quality class, following the PATO approach to represent quality states. Even though states are specializations of qualities, in the metalevel we define a specific metaclass $\ll QualityState \gg$ for them (see Fig. 4).

In the kernel of the descriptive metamodel there is the `Character` metaproperty, as a specialization of the `OWLObjectProperty` metaproperty – see Fig. 3 center. Instances of the `Character` metaproperty – we refer as $\ll Character \gg$ properties – will be properties playing the role of biology characters. We define, in a metamodel level, how $\ll Character \gg$ properties will be defined in the model level, by specializing the OWL domain and range – see `characterDomain` specializing `OWLDomain` and `characterRange` specializing `OWLRange` in the figure. Therefore, the domain of a given $\ll Character \gg$ property will be constrained to an `Entity` and its range will be constrained to a `Quality` or a `QualityState`. In this way, our metamodel captures and formalizes the rationale of biologists when producing phenotype descriptions. In a stage of the descriptive process, characters are related to `Taxa` and they have their ranges constrained to those values observed in the respective `Taxon` – see our explanation of Attributes in Section II. In this case, a $\ll Character \gg$ property will be related with a `Taxon` in the model through a `characterTaxon` property.

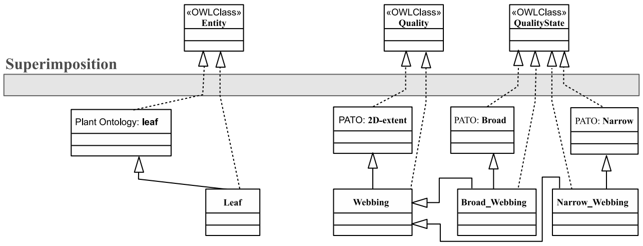


Fig. 4. Example of a superimposed metamodel.

In Fig. 3 right, we present a model to represent a phylogenetic tree, based on the CDAO [15] model. It is here connected to our metamodel by the class `Feature`. Each instance of this class will be associated with $\ll Character \gg$ properties, through the `hasCharacter` property, and specify specific values assumed by the property (states), which are instances of $\ll QualityState \gg$ classes. This connection enables phylogenetic tree nodes to specify the diversification in a semantically richer way, which is connected with the overall phenotype descriptive metamodel/model.

Fig. 5 shows a model which is a practical application of our superimposed metamodel in a case of a biologist describing ferns. The `leaf` class corresponds here to the Plant Ontology `leaf` class. The `2D-extent` class comes from the PATO ontology. Both classes were superimposed by our metamodel as instances of the `Entity` and `Quality` metaclasses respectively. This show how our approach is able to formally incorporate external ontologies without changing their original models.

The `Leaf_Webbing` $\ll Character \gg$ property – instance of the `Character` metaproperty – describes the extent of the lamina within the leaf. Therefore, its domain is the $\ll Entity \gg leaf$ class and its range is a subclass of $\ll Quality \gg 2D - extent$ specialized for leaf, the $\ll Quality \gg Webbing$ class. The `characterDomain` and `characterRange` were used instead of the RDF/OWL domain and range, since they are specializations tailored to properly define a character, as presented in the metamodel. In the left side of Fig. 5, we present the classic descriptive system – detailed in Section II and Fig. 1 – connected with our model which is derived from our superimposed metamodel.

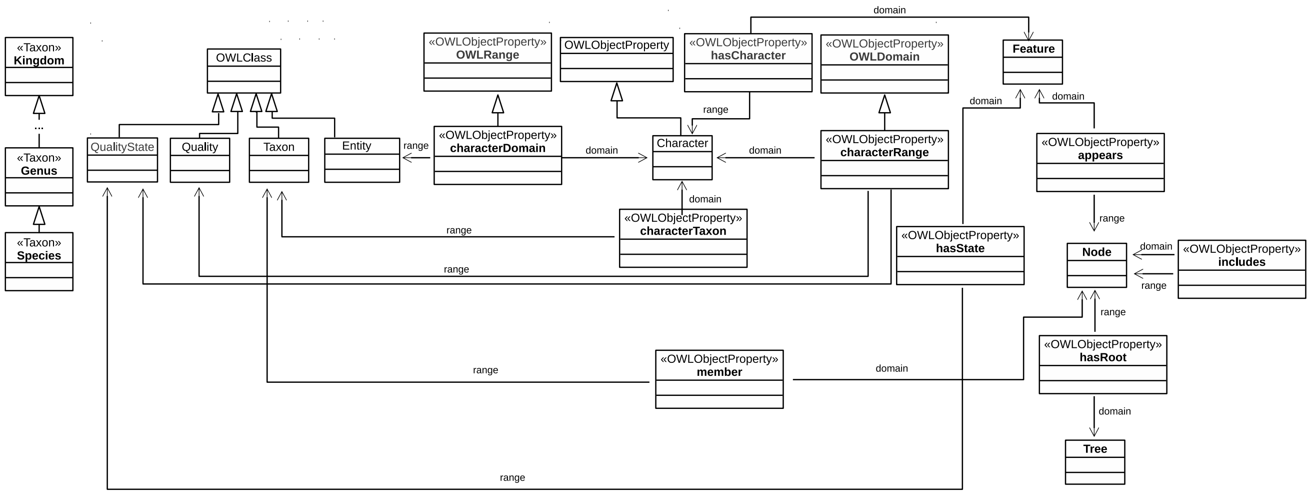

«Taxon» Kingdom
...
«Taxon» Genus
«Taxon» Species
OWLClass
QualityState
Quality
Taxon
Entity
«OWLObjectProperty» OWLRange
OWLObjectProperty
«OWLObjectProperty» hasCharacter
«OWLObjectProperty» OWLDomain
Feature
«OWLObjectProperty» characterDomain
Character
«OWLObjectProperty» characterRange
«OWLObjectProperty» characterTaxon
«OWLObjectProperty» hasState
«OWLObjectProperty» appears
Node
«OWLObjectProperty» includes
«OWLObjectProperty» hasRoot
«OWLObjectProperty» member
Tree
domain
range

Fig. 3. Taxonomic Description Metamodel.

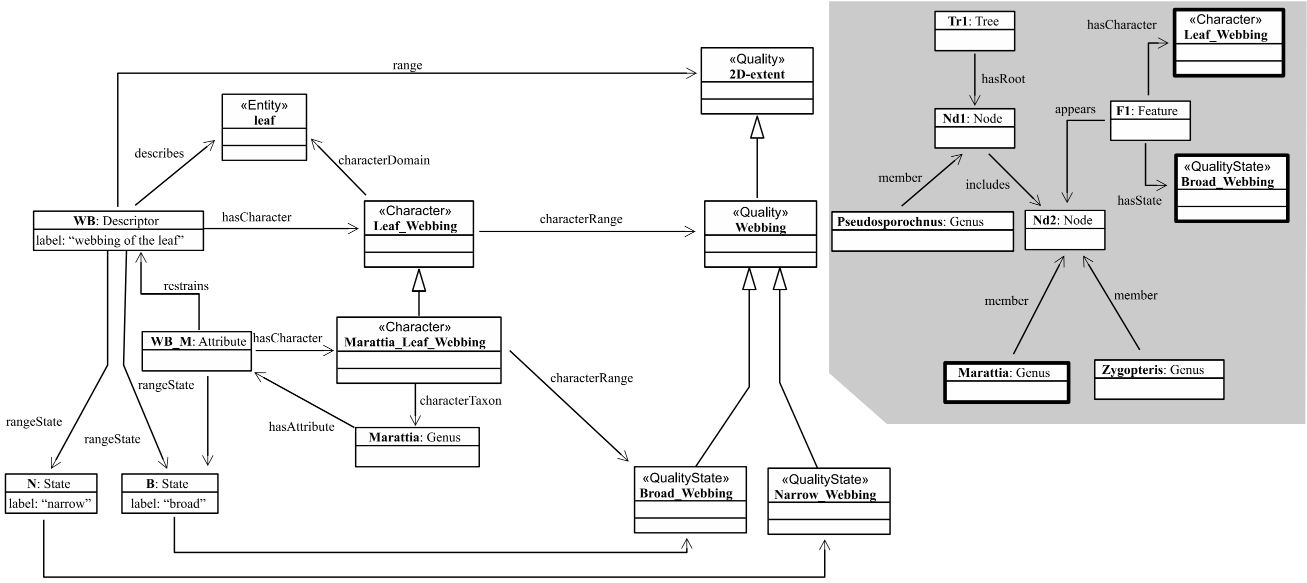


Fig. 5. Applying the Metamodel to a Taxonomic Description Model.

This connection is the basis to bridge existing descriptions to our model, maintaining the traceability. It will be also the basis to connect this new metamodel/model with descriptions produced in the Xper2 system.

As shows Fig. 5, the states of a given descriptor become $\ll\! QualityState \!\gg$ classes. In our example, “broad” and “narrow” become the $\ll\! QualityState \!\gg\! Broad\_Webbing$ and $\ll\! QualityState \!\gg\! Narrow\_Webbing$, which are subclasses of the $\ll\! Quality \!\gg\! Webbing$ class. Since a given taxon restrains the possible values of a given character – as mentioned before – our model defines a new character as a subproperty with a restrained range. In our example, $\ll\! Character \!\gg\! Marattia\_Leaf\_Webbing$ is a subcharacter of Leaf_Webbing and is related to the *Marattia* Genus by a characterTaxon property, as described in the metamodel. Genus represents the $\ll\! Taxon \!\gg\! Genus$ of Fig. 3. This new subcharacter is related to a subclass of $\ll\! Quality \!\gg\! Webbing$ class – the $\ll\! QualityState \!\gg\! Broad\_Webbing$ – which restrains the universe of instances to those observed in the *Marattia* genus, in this case only the “broad”. This model formally represents the observation that in the *Marattia* genus the extent of the lamina covers the whole organ.

In the right side of Fig. 5, we represent the connection of our model with the phylogenetic tree. To simplify the diagram, avoiding excessive crossing lines, we duplicated from the left side the $\ll\! Character \!\gg\! Leaf\_Webbing$, the Genus *Marattia* and the $\ll\! QualityState \!\gg\! Broad\_Webbing$. In the example, three genera – instances of $\ll\! Taxon \!\gg\! Genus$ class – are represented in the phylogenetic tree. The `Feature` related to the node `Nd2` is defined by a $\ll Character \gg Leaf\_Webbing$ in the specific state $\ll QualityState \gg Broad\_Webbing$.

V. Practical Application

In the previous section, we showed how our superimposed metamodel formalizes the rationale of biologists and disciplines the way classes are related – which are two contributions of our approach. In this section we go a step further, emphasizing how to explore the abstraction provided by the metamodel to define rules and queries addressing generalized metamodel elements. Therefore, we produced a generalized and reusable set of rules/queries. Whenever we superimpose the metamodel, we also superimpose the rules and the whole inference rationale. The rules can be applied to an ontology as soon as it is integrated by superimposing our metamodel (see the OWL version of our ontology and examples at: http://purl.org/metabio/).

The distinction between Entity and Quality in a knowledge representation system is essential, considering the biologist needs. Our superimposed metamodel integrates the complete process and makes explicit the role of each component: Quality, Entity, Character etc.

We present here two practical examples of requests:

The first request concerns the comparison of descriptions. A biologist needs to know what entities are comparable, i.e., what entities refer to a same character range or quality. For instance, in the table presented in the bottom layer of Fig. 1 the lateral organ of *Pseudosporochnus* is compared with the leaf of *Marattia* and *Zygopteris* by the means of their "webbing". The generic question "What are the entities which are related with a given quality?" would be relevant to investigate what entities in ferns have similar qualities (e.g., "webbing"), what entities are described considering their webbing, or what entities can present a given webbing (e.g., "broad").

This request can be expressed as a query – e.g., by using SPARQL – or as a rule. In this experiment we opted to express the requests as SWRL rules to emphasize the reusability provided by our approach, as rules can be incorporated in the ontology.

Generic request 1:
"What are the entities which are related with a given quality?" The following rule answers this question by setting a property "related" connecting Entity with the respective Qualities:

```
Entity(?x), Quality(?q), Character(?c),
characterDomain(?c, ?x), characterRange(?c,
?q) -> related(?x, ?q)
```

Refining the request for the $\ll Quality \gg Webbing$:

"What are the entities which are described by the means of their webbing?"

```
Entity(?x), Quality(Webbing), Character(?c),
characterDomain(?c, ?x), characterRange(?c,
Webbing) -> related(?x, Webbing)
```

Applying the rule:

We further show how the system will follow the path to answer that the $\ll Entity \gg Leaf$ is described by means of the $\ll Quality \gg Webbing$:

```
Entity(Leaf), Quality(Webbing),
Character(Leaf_Webbing),
characterDomain(Leaf_Webbing, Leaf),
characterRange(Leaf_Webbing, Webbing) ->
related(Leaf, Webbing)
```

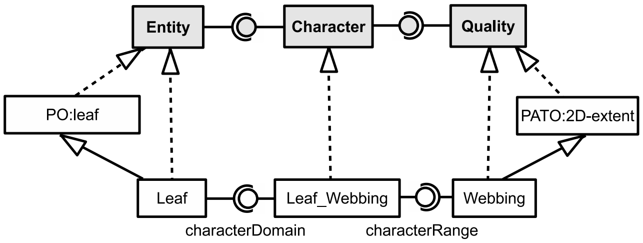


Fig. 6. Practical application of the superimposed metamodel.

Fig. 6 graphically illustrates the previous answer to the query. It shows three basic elements of the metamodel in grey – Entity, Character and Quality – and their expected relations. In white, it shows instances of the metamodel, i.e., the Entities $PO : Leaf$ (from the Plant Ontology) and $Leaf$; the Qualities $PATO : 2D - extent$ (from the PATO ontology) and $Webbing$; the Character $Leaf\_Webbing$. As previously presented in Fig. 4, the $Leaf$ and $Webbing$ are subclasses of external ontologies. It emphasizes an advantage of superimposing our metamodel. It enables to build a query in metamodel terms – i.e., looking for Entity, Quality and Character – which are homogeneous for any member of the knowledge base, even though the elements in the answer can come from different ontologies, which carry their own ontology structure.

The second request illustrates how our superimposed metamodel can connect elements of the complete process. Therefore, we will connect features of the phylogenetic tree with the respective phenotype description. A specific topology of a phylogenetic tree is inferred from the phenotype descriptions and is supported by them. Sometimes, elements of phenotypes must be matched with features related to each node of the tree, to support inferences and discussions. In our example, a biologist wants to investigate chains of differentiations in the tree (i.e., successive differentiations concerning the same Entity in different stages of the tree); for instance, a transformation series.

Generic request 2:
"Is a feature F2 differentiated from a feature F1?"

In order to answer this request the rule will check if there is two successive differentiations in the same Entity.

```
Node(?nd1), Node(?nd2), includes(?nd1, ?nd2),
Feature(?f1), Feature(?f2), appears(?f1,
?nd1), appears(?f2, ?nd2), Character(?c1),
Character(?c2), hasCharacter(?f1, ?c1),
hasCharacter(?f2, ?c2), Entity(?e),
characterDomain(?c1, ?e), characterDomain(?c2,
?e) -> possibleDifferentiation(?f2, ?f1)
```

Evolving the request 2:
"Are unbranched leaves differentiated from branched leaves?"

In this second version we want to refine our request specifying that both features refer to the same Entity (`Leaf`), the same Character (`Leaf_Branchiness`), but in two distinct states: “branched” and “unbranched”.

```
Node(?nd1), Node(?nd2), includes(?nd1,
?nd2), Feature(?f1), Feature(?f2),
appears(?f1, ?nd1), appears(?f2, ?nd2),
Character(Leaf_Branchiness), hasCharacter(?f1,
Leaf_Branchiness), hasCharacter(?f2,
Leaf_Branchiness), hasState(?f1, Branched),
hasState(?f2, Unbranched), Entity(Leaf),
characterDomain(Leaf_Branchiness, Leaf) ->
possibleDifferentiation(?f2, ?f1)
```

As mentioned in Section 4, related work does not have neither a formal approach to distinguish in ontologies roles in a meta-level of abstraction, nor a formal set of relations among them, as our metamodel. Therefore, the rules presented in this section cannot be expressed in this general terms by related work. For example, in our model, we can superimpose any existing biology ontology – e.g., anatomical ontologies, plant ontologies etc. – with the $\ll\mathit{Entity}\gg$ or $\ll\mathit{Quality}\gg$ metaclasses and they will comply with our rules. Related work must write ad hoc rules to any specific involved ontology, as they are not abstracted in a upper level.

VI. Conclusions

In this paper we presented our superimposed metamodel driven approach to integrate and relate biology ontologies. It synthesizes in a unifying metamodel the process from the phenotype description to the phylogenetic tree, supporting inferences crossing the overall representation. Our superimposed metamodel takes advantage of existing ontologies and abstracts, on top of them, the rationale followed by biologists to produce descriptions. It fosters and disciplines the connection among ontologies, which were not originally related. Moreover, our metamodel enables traceability across phylogenetics and descriptions. From the phylogenetic tree yielding homologous features, it is possible to check which characters are involved.

We showed by some practical examples, expressed as rules, that we are able to produce reusable rules addressing our generic metamodel. It materialized in a formal and useful way some techniques, which are discussed in related work, but were not formalized.

Future work include two directions: (i) The integration of our metamodel and related ontologies with our tools – Xper2 and LisBeth – so they can operate at a more semantic level; the ontology will allow both tools to operate in an integrated perspective. (ii) The development of a process to support semi-automatic transformation of existing XML-based descriptions in ontologies.

Acknowledgements. This work was partially supported by grants from Brazilian agencies FAPESP and CNPq, in particular the FAPESP-MSR Virtual Institute (NavScales project), the CNPq MuZOO project and the CNPq Complex Descriptions on the Web project, as well as FAPESP/Cepid in Computational Engineering and Sciences, FAPESP-PRONEX (eScience) and the INCT in Web Science. We thank Paula Mabee for useful comments and advices.

References

[1] M. D. Brazeau, “Problematic character coding methods in morphology and their effects,” *Biological Journal of the Linnean Society*, vol. 104, no. 3, p. 489498, 2011.

[2] M. J. Dallwitz, “A general system for coding taxonomic descriptions,” *Taxon*, p. 4146, 1980.

[3] G. Hagedorn, “Structuring descriptive data of organisms requirement analysis and information models,” 2007.

[4] V. Ung, G. Dubus, R. Zaragüeta-Bagils, and R. Vignes-Lebbe, “Xper2: introducing e-taxonomy,” *Bioinformatics*, vol. 26, no. 5, p. 703704, 2010.

[5] A. Corvez, *L’origine de la megaphylle chez les Monilophytes.*, muséum national d’Histoire naturelle ed., Paris, 2012.

[6] J. A. Hawkins, C. E. Hughes, and R. W. Scotland, “Primary homology assessment, characters and character states,” *Cladistics*, vol. 13, pp. 275–283, 1997.

[7] J. Lebbe, “Représentation des concepts en biologie et médecine. introduction à l’analyse des connaissances et à l’identification assistée par ordinateur,” Ph.D. dissertation, Thèse de doctorat, spécialité Sciences de la vie. Université Pierre et Marie Curie - Paris 6,, Paris, 1991.

[8] N. I. Platnick, “Philosophy and the transformation of cladistics,” *Systematic Zoology*, vol. 28, no. 4, pp. 537–546, 1979.

[9] G. J. Nelson and N. I. Platnick, *Systematics and Biogeography, Cladistics and Vicariance*, columbia university press ed. New York: Guildford, Surrey, 1981.

[10] R. Zaragüeta-Bagils and E. Bourdon, “Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data,” *C. R. Palevol*, 2007.

[11] R. Zaragüeta Bagils, V. Ung, A. Grand, R. Vignes-Lebbe, N. Cao, and J. Ducasse, “LisBeth: new cladistics for phylogenetics and biogeography,” *Comptes Rendus Palevol*, Aug. 2012.

[12] D. L. Swofford, “PAUP: phylogenetic analysis using parsimony,” Champaign, 1993.

[13] J. P. Huelsenbeck and F. Ronquist, “MRBAYES: bayesian inference of phylogenetic trees.” *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[14] P. A. Goloboff, J. S. Farris, and K. C. Nixon, “TNT, a free program for phylogenetic analysis,” *Cladistics*, vol. 24, no. 5, pp. 774–786, Oct. 2008.

[15] F. Prosdocimi, B. Chisham, E. Pontelli, J. D. Thompson, and A. Stoltzfus, “Initial implementation of a comparative data analysis ontology,” *Evolutionary bioinformatics online*, vol. 5, p. 47, 2009.

[16] G. V. Gkoutos, E. C. J. Green, A.-M. Mallon, A. Blake, S. Greenaway, J. M. Hancock, and D. Davidson, “Ontologies for the description of mouse phenotypes,” *Comparative and Functional Genomics*, vol. 5, no. 6-7, pp. 545–551, 2004.

[17] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, “Linking human diseases to animal models using ontology-based phenotype annotation,” *PLoS Biology*, vol. 7, no. 11, p. e1000247, Nov. 2009.

[18] J. P. Balhoff, W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision, “Phenex: Ontological annotation of phenotypic diversity,” *PLoS ONE*, vol. 5, no. 5, p. e10500, May 2010.

[19] C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner, “Method integrating phenotype on tologies across multiple species,” *Genome Biology*, 2010.

[20] P. M. Mabee, G. Arratia, M. Coburn, M. Haendel, E. J. Hilton, J. G. Lundberg, R. L. Mayden, N. Rios, and M. Westerfield, “Connecting evolutionary morphology to genomics using ontologies: a case study from cypriniformes including zebrafish,” *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, vol. 308B, no. 5, pp. 655–668, Sep. 2007.

[21] D. Gasevic, D. Djuric, and V. Devedzic, *Model Driven Engineering and Ontology Development*. Springer Publishing Company, Incorporated, 2009.