Towards a Linked Biology – An integrated perspective of phenotypes and phylogenetic trees

Eduardo Miranda[1], Anaïs Grand[2], Régine Vignes Lebbe[3], and André Santanchè[1]

[1] Institute of Computing – State University of Campinas
Av. Albert Einstein, 1251 – Cidade Universitária, Campinas, Brazil
[2] Muséum national d'histoire naturelle, CR2P - UMR 7207 CNRS/MNHN/Univ Paris 06, 57 rue Cuvier, CP48 - F-75005, Paris, France
[3] UPMC Univ Paris 06, CR2P - UMR 7207 CNRS/MNHN/Univ Paris 06, 4 Place Jussieu, Tour 46-56, 5ème étage, F-75005, Paris, France
eduardo.dpm@gmail.com,grandanais@gmail.com
regine.vignes_lebbe@upmc.fr,santanche@ic.unicamp.br

Abstract. A large number of studies in biology, including those involving phylogenetic tree reconstruction, result in the production of a huge amount of data – e.g., phenotype descriptions, morphological data matrices, etc. Biologists increasingly face a challenge and opportunity of effectively discovering useful knowledge by crossing and comparing several pieces of information, not always linked and integrated. Our motivation stems from the idea of transforming these data into a network of relationships, looking for links among related elements and enhancing the ability to solve more complex problems supported by machines. This work addresses this problem through a graph database model, linking and coupling phylogenetic trees and phenotype descriptions. In this paper we give an overview of an experiment exploiting the synergy of linked data sources to support biologists in data analysis, comparison and inferences.

1 Introduction

In spite of several initiatives to publish open data and to combine phylogenetic trees, there is still a high amount of latent knowledge hidden in potentially linkable data, which are fragmented in several heterogeneous datasources. This heterogeneous multitude of resources can be seen as a dataspace [1], where pieces of data maintain unexploited potential links. This work addresses this problem in a specific scenario. We gather together in a graph database data coming from distinct sources, containing phenotype descriptions and phylogenetic trees. This graph subsidizes links discovery, aimed at supporting biologists in the analysis and comparison of phylogenetic information (such as homology hypotheses, characters and trees) of hypothetical phylogenetic trees.

This paper is organized as follows: Section 2 presents our three layer method and the architecture of our system; Section 3 presents our graph-based model, illustrates our approach to discover links based on similarity and the results of the visualization tool; Section 4 presents concluding remarks.

2 Three Layer Method and System Architecture

In this research we propose a Three Layer Method, in which fragmentary data sources are mapped to a graph database, where the data will be pass through integration steps targeting an ontology. Our approach remodels sources from the dataspace to a graph representation, in which the data can be unified and linked, subsidizing the discovery of latent knowledge, which raises from the relations. The graph model was designed to be published on the Web in a Linked Data approach. Graph transformations will be applied for the transition from representations in the dataspace to a more formalized representation through ontologies. This work focuses in the graph representation and its application to support an analytical tool to compare data across studies.

Figure 1 summarizes the general architecture of our system. From a set of heterogeneous data sources (1), we ingest and transform data in a graph (2) stored in a graph database (3). In this stage of the research, we are interested in phenotype descriptions and phylogenetic trees, even though the architecture was designed to afford smooth future extensions to other kinds of biological data. In step (3) each data source will result in a distinct graph. We applied LSIDs to unify Operational Taxonomic Units (OTUs) in the graph referring to the same real world object (4). In step (4), we are developing algorithms to discover relations and find similarities among nodes in the graph, which are made explicit by adding new edges in the graph. The resulting graph can be locally analyzed by a researcher; can be published on the Web in a Linked Data approach to be remotely exploited (6); and will subsidize the expansion and enrichment of ontologies in the future (7).

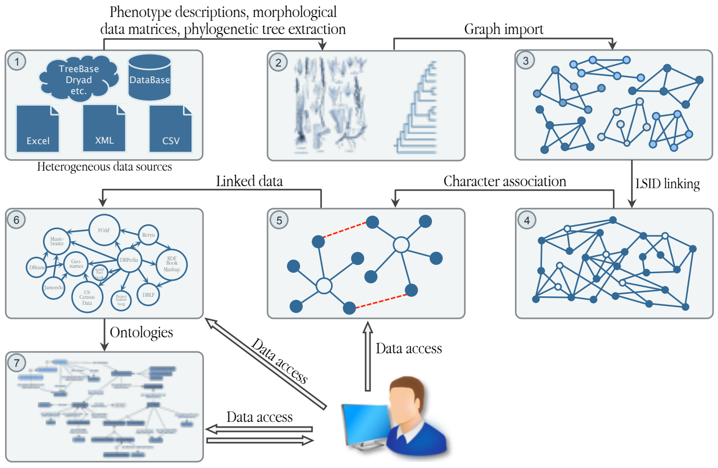

Fig. 1. General System Architecture.

3 Unified Graph Data Model and Link Discovery

In this section we will present an overview of our proposed graph model. From the numerous graph data models proposed, the *property graph* model was adopted in the present work. In *property graph*, nodes and relationships can maintain extra metadata as a set of key/value pairs. Moreover, relationships are typed, enabling to create multi-relational networks with heterogeneous sets of edges.

Figure 2 shows our graph data model. The tables below the nodes/edges represent their types and metadata. Our model maps parts of the Structured Descriptive Data [4] (SDD) format to a graph data model. The SDD format is a XML-based standard for recording and exchanging descriptions of biological and biodiversity data of any type [2]. We mapped as follows: OTUs are entities made nodes. SDD States were mapped to nodes, since equivalent *States* related to several OTUs can be unified and related. Finally, the SDD *StateDefinition* makes a semantic bridge – relationship – between OTUs and characters. This part of the model, is a common denominator of phenotype descriptions, conceived in our previous work [3].

Our model comprises into a single place phenotype description and phylogenetic tree. For this reason a new node called HTU (Hypothetical Taxonomic Unit) is present in this model. HTUs are internal nodes in phylogenetic trees that represent an inferred ancestral organism.

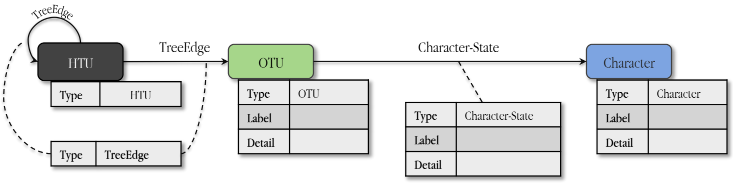

Fig. 2. Property Graph Model

In order to illustrate the possibilities raised by the unification and linking of data of phenotype descriptions with phylogenetic trees, we developed a practical experiment executed in our system, which involves the linking discovery among characters. In this respect, we are proposing a heuristic similarity measure that computes the similarity degree between two morphological character. The similarity degree is based on a weighted combination of factors concerning two characters: their labels; whether they describe the same taxa; the common or similar character-states.

Figure 3 shows the result of a visual tool that creates an edge between each two related characters based on the similarity measure. This is a simple but powerful visualization tool that could play a pivotal role in supporting biologists to understand and detect correlation between characters. That tool was able to show a graph clique among highly related characters – see figure 3. Knowing that characters are similar to some extent can encourage biologists to unify identical relationships.

[4] The Structured Descriptive Data format (*http://wiki.tdwg.org/SDD)*

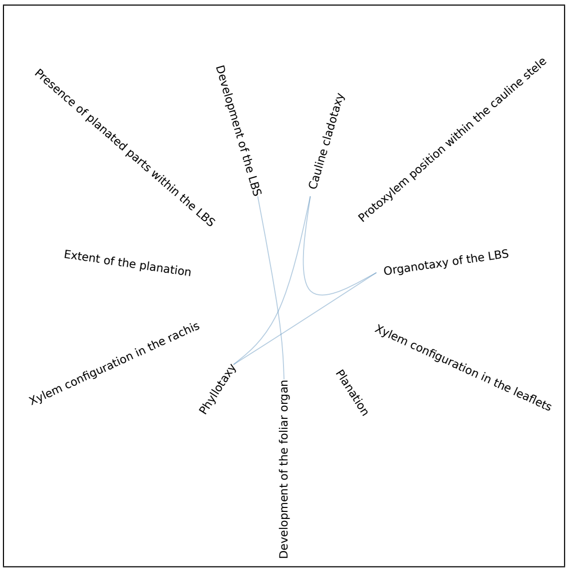
Presence of planated parts within the LBS
Development of the LBS
Cauline cladotaxy
Protoxylem position within the cauline stele
Extent of the planation
Organotaxy of the LBS
Xylem configuration in the rachis
Xylem configuration in the leaflets
Phyllotaxy
Planation
Development of the foliar organ

Fig. 3. Practical Experiment

4 Conclusion

Our unified model enabled us to discover and make explicit the potential semantics raised by linking previously unconnected information. The viability and the potential of our approach were tested by experiments in which 2 distinct authors descriptions of fossils were inserted into a graph database and analyzed by the similarity measure method mentioned in this paper. We developed a tool to visualize how close related are two given characters and some preliminary results are presented. This tool has the potential to indicate the recurring use of the same character in different studies and might support biologists to understand and detect correlation between characters.

Acknowledgments. Work partially financed by (CNPq 138197/2011-3), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project and PRONEX-FAPESP), INCT in Web Science(CNPq 557.128/2009-9) and CAPES-COFECUB, as well as individual grants from CNPq.

References

1. Franklin, M., Halevy, A., Maier, D.: From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. **34**(4) (December 2005) 27–33
2. Hagedorn, G.: Structuring Descriptive Data of Organisms – Requirement Analysis and Information Models. PhD thesis, Universität Bayreuth,Fakultät für Biologie, Chemie und Geowissenschaften (11 2007)
3. Miranda, E., Santanchè, A.: Unifying phenotypes to support semantic descriptions. VI Brazilian Conference on Ontological Research (Ontobras) (09 2013)