Generating Knowledge Networks from Phenotypic Descriptions

Fagner Leal Pantoja, Patrícia Cavoto, Julio Cesar dos Reis, André Santanchè
Institute of Computing
University of Campinas
Campinas, São Paulo, Brazil
{fagner.pantoja, julio.dosreis, santanche}@ic.unicamp.br
patricia.cavoto@gmail.com

***Abstract*—Several computing systems rely on information about living beings, such as *Identification Keys* – artifacts created by biologists to identify specimens following a flow of questions about their observable characters (phenotype). These questions are described in a free-text format, *e.g.*, "big and black eye". Free-texts hamper the automatic information interpretation by machines, limiting their ability to perform search and comparison of terms, as well as integration tasks. This paper proposes a method to extract phenotypic information from natural language texts from biology legacy information systems, transforming them in an *Entity-Quality* formalism – a format to represent each phenotype character (*Entity*) and its state (*Quality*). Our approach aligns automatically recognized *Entities* and *Qualities* with domain concepts described in ontologies. It adopts existing Natural Language Processing techniques, adding an extra original step, which exploits intrinsic characteristics of phenotypic descriptions and of the organizational structure of *Identification Keys*. The approach was validated over the *FishBase* data. We conducted extensive experiments based on a manually annotated Gold Standard set to assess the precision and applicability of the proposed extraction method. The obtained results reveal the feasibility of our technique, its benefits and possibilities of scientific studies using the extracted knowledge network.**

I. Introduction

Within the large set of knowledge bases containing information about living beings, phenotype descriptions play a key role, denoting the visible properties of organisms. These descriptions are written in a textual format and are mainly composed by morphological characters, *e.g.*, "eye", and related qualifiers, *e.g.*, "big".

Many problems arise when descriptions are written in a free-text format, such as the possibility of writing the same description in different ways. For instance, "*median fin skeleton*" can be written as "*unpaired fin skeleton*" and "*axial fin skeleton*". It makes difficult the fully semantic interpretation of data by computers and limits their capacity of supporting accurate analyses over the information. The challenge in this scenario is how to distinguish, as automatically as possible, the characters and their states from free-text descriptions.

In this paper, we propose a method to detect the phenotype descriptions expressed in an *Identification Key* (IK), which is a decision tree to identify a specimen based on observed characters [1]. Our proposal transforms the recognized elements into a semantic-based representation aligned to the *Entity-Quality* (EQ) approach [2]. In an EQ statement, the *Entity* refers to the morphological character (*e.g.*, "*eye*") and the *Quality* stands for a qualifier (*e.g.*, "*big*") that specifies a given state of the *Entity*.

This investigation defines a two-step method. The first step analyses a sentence using a Natural Language Processing (NLP) technique that produces a *Dependency Tree*, establishing dependency relations between the sentence terms. It extracts EQ elements computing matches between ontology concepts and terms of the tree. We assume that the relations among terms in the *Dependency Tree* have latent *Entity-Quality* statements. They reflect the biologists approach to write phenotype descriptions: a term (or a set of terms) representing a given *Entity* has specific kinds of dependency with a term representing its *Quality*. The second step takes advantage of the way that biologists relate and structure the phenotype descriptions. This step explores the correlations between sentences inside the IK. The identified *Entities* and *Qualities* are connected to domain ontologies to make their semantic explicit.

We conducted an experimental evaluation using data from *FishBase* to validate the proposed method. *FishBase*[1] is a fish knowledge base containing information used by researchers, fishery managers and zoologists. We show how recognized EQs can link descriptions of several species to produce a knowledge network and how this network can be explored for data analysis. The results indicate the adequacy and potentialities of our approach.

The remainder of this article is organized as follows: Section II formulates the research scenario and problem. Section III discusses the related work. Section IV describes the proposed method for the extraction and semantic linking of phenotype EQs. Section V reports on our experimental evaluation and shows potential applications of the generated knowledge network. Finally, Section VI draws conclusions and future work.

II. Research Scenario and Problem Definition

Among several types of data managed by FishBase, Identification Keys (IKs) consist in artifacts created by biologists to

[1] http://www.fishbase.org

Couplet	Character	Next	Prev	Link
1 a	One continuous dorsal fin.	2	(1)	
1 b	2 dorsal fins or dorsal fins clearly separated into 3 parts.	4	(1)	
2 a	Spines present in dorsals, sometimes feeble, pelvics present.	3	(1)	
2 b	No spines in dorsal or anal fin; eel-like; pelvics absent.	-	(1)	Apodocreedia, Creediidae
3 a	Mouth extending beyond eye, with elongate maxilla; caudal fin rounded to subtruncate.	-	(2)	Opistognathus, Opistognathidae
3 b	Mouth reaching eye, lower jaw projecting; caudal fin pointed; first 2-3 dorsal rays filamentous, free.	-	(2)	Trichonotidae

Fig. 1: Fragment of *Identification Key* to the *Teleostean families* from East Africa (sub-order *Trachinoidei*). Source: http://www.fishbase.org/keys/description.php?keycode=799.

identify species or any other taxonomic group (called taxon) of an observed specimen [1]. An IK denotes a structured set of phenotype descriptions of organisms. To identify a living being using an IK, users might navigate through a series of multiple choice questions about the specimen characteristics. According to the picked answers, the path leads to the respective taxon. Currently, *FishBase* has 1,668 IKs of fishes containing 25,542 phenotype description sentences.

As an example of IK usage, Figure 1 presents an IK to identify the *Teleostean* families, from East Africa (sub-order *Trachinoidei*). The identification process begins with question 1, which has the pair of options 1a and 1b, with their descriptive texts in the *Character* column. According to the picked answer, the user might navigate to question either 2 or 4, indicated in the *Next* column. Each descriptive text inside the *Character* column is called *Key Question* (KQ). This process is repeated until the biologist reaches a row that does not lead to another question. At this stage, the specimen is identified and its respective taxon appears at the *Link* column.

IKs and other data of *FishBase* are stored in a set of relational tables. Handling all these data manually is a huge challenge for scientists, who face difficulties to analyse some scenarios involving the network of relations (links) among taxa and their characteristics. The overwhelming amount of phenotype descriptions is in free-text format. This format is more flexible and easier to produce, having advantages in the narrative structure and providing better expressiveness. However, this free-text format is inappropriate for some computational tasks, mainly when it involves the interpretation and comparison of the content by machines. It hampers tasks involving information retrieval and integration with other sources, since the description components are "locked" within the text.

Therefore, it is necessary to develop methods that can automate the identification, extraction, and integration of phenotype information hidden in descriptive texts [3]. This research faces the problem through a method that automatically recognizes *Entities* and *Qualities* inside phenotype description sentences and link them to other data managed by *FishBase* (*e.g.*, species, genus, country). It produces a knowledge network, making possible:

- **Reuse of EQs:** If EQs are duly unified in a semantic level, it is possible to identify which IKs refer to the same EQs, making explicit the network among IKs and EQs.
- **No need of previous knowledge:** In *FishBase*, IKs are segmented according to the taxa that they identify. Therefore, users must know beforehand the specimen's taxon to pick a correct IK. This process is laborious and error-prone; In addition, it limits the use of the system only to expert biologists, who could not have previous clues about the specimen to be identified. An explicit and standard semantic representation might enable to correlate EQ elements of several IKs and combine them in a unified identification tree.
- **Relation between taxa and keys:** Unified and semantic-enriched descriptions will enable to perform analyses to understand facts including: (i) which IKs identify similar taxonomic groups; (ii) which EQ elements are determinant to discriminate a taxon of a specimen; (iii) which EQ elements define a specific taxon.

III. Related Work

There is a huge amount of biological data available in free-text format. As the process of producing biological data is expensive and complex, it is necessary to leverage the capability of automatically computing existing data. Thus, there is a challenge of migrating such vast amount of data into machine-interpretable formats, in order to produce semantically explicit knowledge.

These machine-interpretable data can be used by generic identification systems to improve their process and results. These systems implement different identification processes, such as: by descriptive characteristics, by pictures, by morphological measures, *etc*. Besides the generic identification systems, some information systems specialized in specific kinds of organisms may also offer support to build and publish IKs. For example, *FishBase* for fishes and *Bird Id* (http://www.birdid.co.uk) for birds. In the *FishBase* case, the identification process can be conducted in distinct ways, such as by images, by ecosystems, through descriptive characteristics, *etc*.

Several systems allow people to digitally create and publish Identification Keys for organisms [4], for example, *Intkey*, *IdentifyIt*, *Linnaeus II*, *Lucid* [5], *MEKA*, *NaviKey*, *PollyClave*, *XID*, *xPer* [6], *ActKey*, *eFloras*, *SLIKS*, and *KeyToNature* [7]. Technical reviews of some of these tools can be found in Dallwitz *et al.* [8].

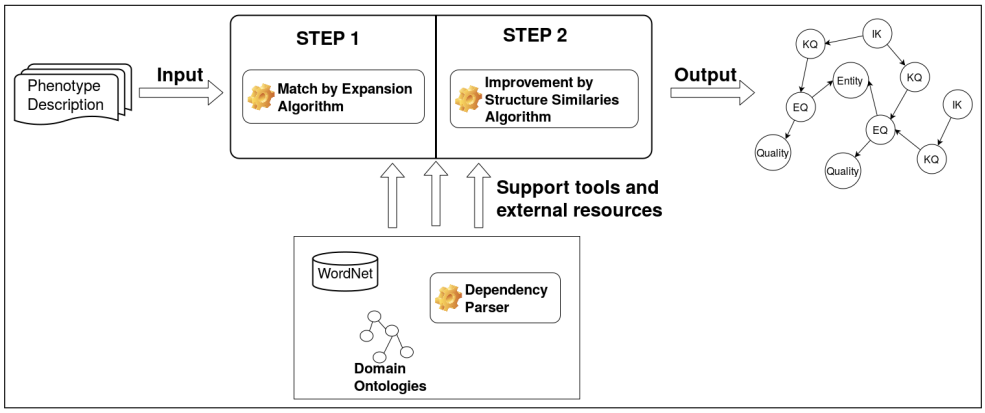

Fig. 2: General view of the proposed approach.

Farnsworth *et al.* [4] give an overview of technical innovations and trends in the area and highlight the importance of ontologies and semantics. They show that there is still space for improvements on the usage of these data, concerning data analysis and the correlation of phenotypes across different taxa and systems.

Existing investigations consider the use of phenotype descriptions in a machine-interpretable format. *Phenoscape*[2] addresses this issue adopting the *Entity-Quality* (EQ) approach to describe phenotypes and developing a scalable infrastructure that enables linking phenotypes across different fields of biology by the semantic similarity of their descriptions.

Concerning how to make explicit the semantics of biological data, Dahdul *et al.* [9] investigated techniques for transforming descriptive biology texts into a format that enables large-scale computation. Based on a previous study, they claim that large-scale computation can benefit from annotating characters with ontology terms. Therefore, they advocate the need of efficient methods to automatically extract and annotate phenotypes from descriptions and consider that NLP tools can be used in the process.

Related work concerning phenotype extraction are mostly concentrated in: interactions among genes, proteins, drugs, and diseases. This can be seen in Ciaramita *et al.* [10], Song *et al.* [11], Pyysalo and Ananiadou [1], Ramakrishnan *et al.* [12], and Fundel *et al.* [13]. Although the domains are similar to our work, we exploit specific characteristics of organisms morphological descriptions to improve the results of our extraction.

Cui [14] presents a method to extract phenotypes that describe leaves, fruits, and nuts of plants. He uses two key techniques: (a) an unsupervised learning algorithm to annotate descriptions at the sentence level, to build a lexicon; (b) the learned lexicon, enhanced by a human user, feeds a parser that recognizes biological characters in descriptive sentences and annotates them. Our work differs since it does not require human intervention during the process, in such a way that a non-expert can use the system.

[2]http://phenoscape.org/wiki/Main Page

Alnazzawi *et al.* [3] compare several statistical learning methods against a curated corpus made by experts, called *PhenoCHF*. This corpus contains annotations about phenotypic information related to Congestive Heart Failure (CHF). One of their objectives is to demonstrate how the well-known methods perform better when a curated corpus is available. However, the creation of a corpus is a hard and expensive task. Our approach was developed to serve in contexts in which such corpus are unavailable.

IV. Extraction of Entity-Quality Terms

This section details our approach to recognize and to make explicit *Entity-Quality* (EQ) elements, which are part of textual descriptions inside semi-structured Identification Keys (IKs). The method involves mapping them to a more formal representation with explicit semantics, based on domain ontologies. The approach departs from natural language text sentences (phenotype descriptions) and produces a graph representation of the recognized EQs. Figure 2 shows the general view of our approach, which encompasses two steps:

- **Step 1:** recognizes EQ elements through an algorithm that analyses the text of the sentence;
- **Step 2:** improves the results of Step 1 recognizing more EQ elements through an algorithm that analyses the relations of sentences according to the structure of the IK.

Both steps rely on external tools and resources throughout the process. This proposal is founded on two assumptions that synthesize the principles behind our method:

Assumption 1: The typical way in which a phenotypic description is written can guide the extraction of EQ elements.

Assumption 2: The way in which a set of phenotype descriptions is organized and structured holds implicit relations that can be exploited to improve the extraction of EQ elements.

Steps 1 and 2 implement algorithms based on Assumptions 1 and 2, respectively (*cf.* Sections IV-A and IV-B). We further define a notation to be used throughout this chapter, which will support the explanation of the method.

$E[e_x]$ = an Entity
$Q[q_y]$ = a Quality
$EQ[e_x, q_1, q_2, ..., q_n]$ = an Entity-Quality
$S[s_x]$ = sentence in free-text format
$V[v_1, v_2, ..., v_n]$ = vertexes of a Dependency Tree

A. Step 1: Exploiting the Writing Characteristics of a Phenotypic Description

Following Assumption 1, in order to guide the extraction task, this step exploits the typical approach followed by biologists to write phenotypic descriptions. This principle was previously exploited by other authors like Cui [14], who listed out some writing characteristics observed in Biology description texts:

1) Generally, morphological descriptions are constituted by two elements: Characters and Character States (C/CS);
2) Omission of Function Words – it is usual the omission of words that do not carry relevant meaning, such as articles and auxiliary verbs (*e.g.*, a, an, the, is, are);
3) Characters are often not explicitly stated in the descriptions. For example, in the sentence "*Black and big eyes*", the characters color and size are not explicitly stated.

To deal with the first item, we have chose to work with Dependency Trees in order to reveal relations between sentence terms reflecting C/CS relations. Dependency Trees are produced by a Dependency Parser, which transforms the sentence in a tree of relationships between words, where each node represents a word and each edge denotes a grammatical dependency. The dependencies are all binary relations [15]. In this work, we use the *Stanford Typed Dependencies Parser*[3] (STDP), a Dependency Parser implementation which belongs to the *Stanford Core NLP* toolkit. Figure 3 shows two examples of Dependency Trees generated by the parser.

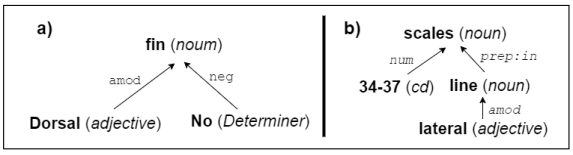

Fig. 3: Dependency trees of the sentences (a) S[*No dorsal fin*] and (b) S[*34-37 scales in lateral line*].

STDP contains approximately 50 grammatical dependencies [16]. We selected a subset which reflects the written characteristics to be analized. For example, $amod$ is a dependency that has an adjective qualifying a noun. In the Dependency Tree, Function Words are represented as edges (*e.g.*, the word "*in*" in Figure 3.b) instead of vertexes. Our match algorithm does not consider the edge labels. Therefore, these Function Words are ignored, which is conform the second writing characteristic of phenotype descriptions, observed by Cui [14] since it does not carry relevant meaning.

In order to obtain a semantic description of EQs, the Dependency Relations are matched with domain ontologies.

[3]http://nlp.stanford.edu/software/stanford-dependencies.shtml

We used ontologies widely adopted by the community: (1) *Teleost Anatomy Ontology* (TAO) [17] – an ontology that formalizes the knowledge about teleostean fishes anatomy; (2) *Phenotypic Quality Ontology* (PATO) – an ontology that defines *Qualities* to be related to *Entities* and their respective values.

We applied a recursive match algorithm that performs a search over a domain ontology (TAO or PATO: *Entity* or *Quality*, respectively). The match algorithm returns the concept in the ontology that has the highest similarity with a given subgraph of the input Dependency Tree. Similarity refers to which degree ($similarity \in [0, 1]$) an existing ontology concept is similar to the terms of the given subgraph. Firstly, the algorithm discovers the *Entities* present in the Dependency Tree. Then, it discovers *Qualities* related to the *Entities* already recognized. Figure 4 presents two elements returned by the match algorithm. The subgraph containing V[*fin, dorsal*] is matched with E[*Dorsal fin*] (a concept of TAO), while the subgraph containing V[*no*] is matched with Q[*absent*] (a concept of PATO).

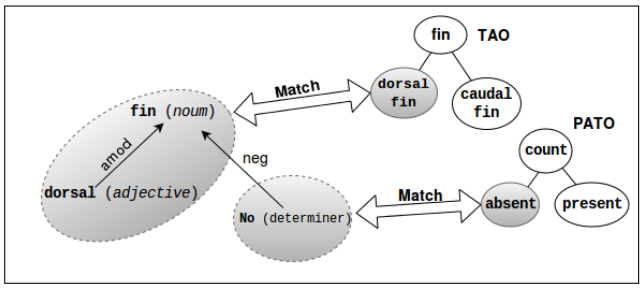

Fig. 4: *Entity* and *Quality* recognized in the sentence S[*No dorsal fin*].

Figure 5 illustrates the execution of the recursive match algorithm looking for an *Entity*. At each iteration, the algorithm expands the subgraph adding vertexes connected to the current subgraph (neighbors). After recursively traversing all the Dependency Tree, it looks for the most similar concept E[*dorsal fin*].

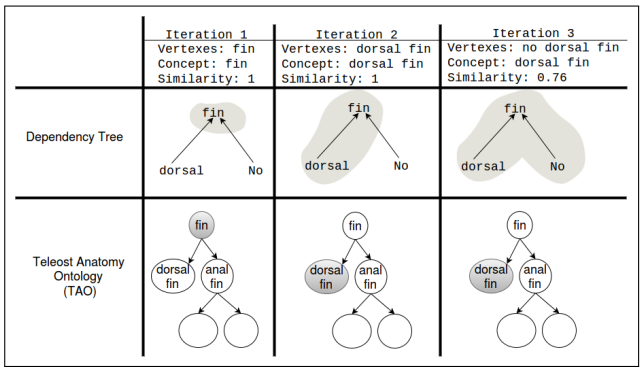

Fig. 5: Match by expansion algorithm over the sentence S[*No dorsal fin*].

Figure 6 presents the result of Step 1: a graph where each

Key Question is connected to the respective recognized *Entities* and *Qualities*, *e.g.*, the nodes E[*dorsal fin*] and Q[*absent*].

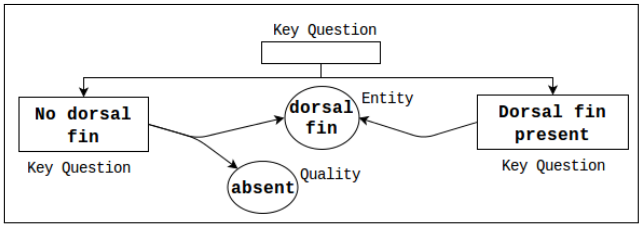

Fig. 6: Step 1 output.

It is possible to observe in Figure 6 that the method fails at recognizing $Q[present]$ in the sentence S[*Dorsal fin present*]. This failure is due to the violation of the English language rules in the sentence formulation. There are other cases where Step 1 fails, then the following Step 2 aims to treat these cases.

B. Step 2: Exploring the Structure of Identification Keys

This step explores the structure of IKs to enrich the output graph from Step 1. We assume that the correlation between distinct descriptions might be useful in the extraction of additional EQs. Such correlation is an intrinsic characteristic of IKs, as a result of their organizational structure. This step is based on the previously mentioned Assumption 2: **The way in which a set of phenotype descriptions is organized and structured holds implicit relations that can be exploited to improve the extraction of EQ statements**.

We believe that the principles behind this work could be generalized to other fields in the future. An organizational structure, as we exploit in the IKs, could also be the sessions of a technical report, the structure of legal documents with juridical rules, the layout of a Web site, *etc*. Wong *et al.* [18] indicate that such noncontent cues may be used to support information extraction tasks. This perspective opens a future wider application scenario for our technique.

IKs are structured in a tree format, in which the alternatives of a given KQ are its sibling nodes containing complementary alternative sentences. This structure offers clues about its content, from which we consider the following characteristics:

(a) Alternatives of a KQ frequently refer to the same *Entities*. In our previous example, both sibling sentences S_1[*No dorsal fin*] and S_2[*Dorsal fin present*] refer to the same anatomical character E[*dorsal fin*];

(b) Alternatives of a KQ are frequently complementary, in the sense that they assign complementary states to the described *Entity*. In the same previous example, the *Qualities* Q_1[*absent*] and Q_2[*present*], assigned to the *Entity* E[*dorsal fin*], are opposites, encompassing its possible state values.

In summary, we assume that if an EQ pair is identified in a KQ, it is very likely that the sibling KQs must refer to the same *Entity*, but potentially using complementary *Quality* terms to modify the *Entities*. The challenge here is to verify if the sibling nodes hold this property.

Therefore, we developed an algorithm that measures the similarity between two sentence pieces. It is based on the general principle of Paraphrase Recognition, which is a process to judge if two different sentences convey the same aspect or the same information. Androutsopoulos and Malakasiotis [19] present a survey regarding Paraphrase Recognition techniques. There are techniques that exploit the dependency tree to measure the similarity between the sentences. In general, they assume that if there is a value above a given threshold, the involved sentences are considered paraphrases.

Usually, Paraphrases Recognition algorithms compare the whole trees [19]. We have adapted the principle of Paraphrases Recognition to the problem of recognizing complementary sentences in an IK.

Step 2 acts in cases where Step 1 was successful in one sentence, but failed in recognizing EQ statements in its siblings. It determines if these sentences have complementary *Qualities* for the same *Entities*. It measures the similarity between the subtrees comparing each edge inside them. We aim to verify if they refer to the same *Entity* with complementary *Qualities*, based on a settled threshold.

Figure 7 illustrates the Step 2 input elements. The algorithm receives a pair of *Key Questions*: KQ_{main} and $KQ_{sibling}$. Inside each KQ node, there are the Dependency Trees of the sentences. The KQ_{main} has a link to an EQ pair E_1Q_1 and the $KQ_{sibling}$ has a link to the same E_1, but it lacks the Q_2 (dashed element), to be recognized by the algorithm.

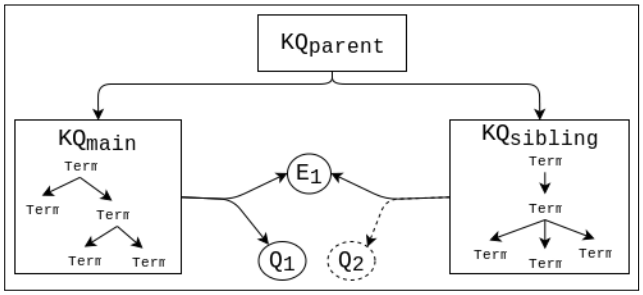

Fig. 7: A generic example of Step 2 input.

The Step 2 algorithm iterates over each EQ pair extracted from the KQ_{main} (E_1 and Q_1 in Figure 7). For each EQ pair, the algorithm gets the corresponding subtrees $Entity_{main\_subtree}$ and $Quality_{main\_subtree}$ that contain the terms that are part of the E_1 and Q_1, respectively. Figure 8 exemplifies these subtrees, highlighting $Entity_{main\_subtree}$ and $Quality_{main\_subtree}$ inside the DT_{main}.

The algorithm gets the $Edge_{main\_2}$, which links the $Entity_{main\_subtree}$ to the $Quality_{main\_subtree}$. Then, the algorithm fetches the subtrees extracted from sibling KQs ($KQ_sibling$, in this case) and compares the original edge with all edges related to the subtree $Entity_{sibling\_subtree}$ (which represents the same entity of $Entity_{main\_subtree}$): edges $\langle Edge_{sibling\_1}, Edge_{sibling\_2}, Edge_{sibling\_3}, Edge_{sibling\_4}\rangle$. This comparison looks for an edge that connects the $Entity_{sibling\_subtree}$ to the complementary Q_2. The similarity

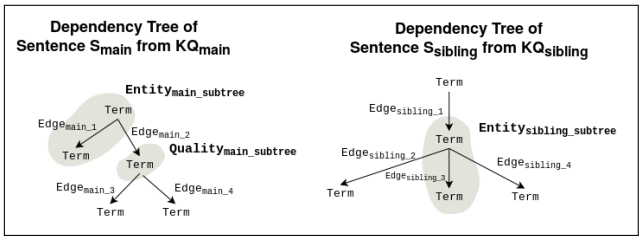

Fig. 8: Generic example of dependency trees of two sentences. *Entity* and *Qualities* recognized in the previous Step 1 are highlighted.

computation between the edges takes into account the following parameters:

(a) Directions of the dependency relations $edge_{sibling\_n}$ and $edge_{main\_2}$;
(b) Grammatical class of Q_1 and Q_2;
(c) Types of the dependency relations of $edge_{sibling\_eq}$ and $edge_{main\_eq}$;
(d) Antonymy between Q_1 and Q_2 (the algorithm explores the *WordNet* lexical database [20] to check if two words are antonyms).

These parameters represent to which extent one edge is similar to another. To calculate the degree of similarity, each parameter contributes with a pre-defined value: $v_a = 0.25$; $v_b = 0.50$; $v_c = 0.75$; $v_d = 1$.

We have chose these parameters and estimated their corresponding values based on empirical observations regarding their relevance in Dependency Tree elements (edges and vertexes) concerning phenotype description sentences. For example, we noted that a pair of edges having the same direction is important, but it is less important than the fact that the *Qualities* have antonyms terms since the algorithm is looking for opposite *Qualities*. These parameters and their values can be adapted to the execution of the algorithm in other scenarios.

The similarity between each pair of edges is calculated through a summation of those parameters. The edge with the highest similarity value is selected as the potential Q_2, if it is equal or higher than a determined $threshold$. In the conducted experiments, we assigned the $threshold = 0.75$ to avoid retrieving edges with low similarity values. Afterward, the recursive match algorithm (the same used in Step 1) performs a search over the $PATO$ ontology in order to confirm if the selected edge corresponds to a *Quality* concept.

The $threshold$ value can be modified and it affects the behaviour of the algorithm. A high $threshold$ value enables to recognize more *Qualities*, but it can increase the rate of false positives. On the other hand, a low value can decrease the number of recognized *Qualities*, but it increases the rate of correct elements. The values of each parameter and threshold have been empirically determined by experimental analyses.

Figure 9 shows the the Step 2 output for the KQs example. Compared to Figure 6, the $Q[present]$ was inserted as a new node in the graph as a result of the algorithm.

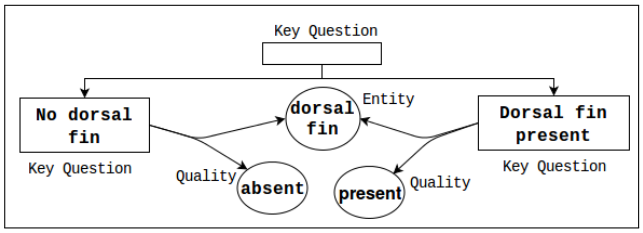

Fig. 9: Step 2 output.

V. Evaluation and Application Experiments

This section reports experimental results of this investigation. We rely on the *FishBase* database to conduct the proposed assessments. Section V-A presents an evaluation to assess the viability of our extraction method. The objective is to investigate the effectiveness of the approach considering a gold standard dataset and traditional metrics.

The initial motivation for this research was to obtain a knowledge network correlating and integrating several elements of phenotype descriptions. Therefore, Section V-B presents experiments linking *FishBase* species data through the recognized EQs.

A. Accuracy Assessment

The quality of the recognition and extraction of elements in natural language texts, *i.e.*, entities or relations, can be evaluated by several mechanisms. The most common considers a standard evaluation set generated by either a group of specialists in the domain, or an organizing committee of a competition. A standard evaluation set contains fragments of texts highlighting the elements that are supposed to be recognized. Such kind of evaluation is suitable when there is a mature developed community acting in the area of interest.

However, there is still no standard evaluation set for morphological descriptions, in the context that we are working, i.e., *Entity* and *Quality* linked in an EQ pair. Therefore, this investigation involved the creation of an evaluation dataset to assess the performance of our method. This dataset has the original sentence descriptions where the EQ elements are annotated. A set of 100 KQs have been manually annotated, from the total of 25,542 KQ from *FishBase*.

Figure 10 shows four examples of sentences in our evaluation dataset. The words in bold compose *Entities*, and words in italic compose *Qualities*, while the boxes represent EQ pairs.

1) **Lips** *not fringed*; **mouth** *horizontal*.
2) *No dark longitudinal* **stripes** on **head** and **body**.
3) Total **vertebrae** *119 to 132*.
4) *Scattered* **breast melanophores**. *One large* **spot** centered at the base of the **caudal fin**.

Fig. 10: Examples of sentences within our Gold Standard.

Several criteria were explored to create the Gold Standard. First, we considered only Simple EQs, *i.e.,* those composed

strictly by one *Entity* and one *Quality*, such as in the second sentence in Figure 10: E[*stripes*]Q[*no*], E[*stripes*]Q[*dark*] and E[*stripes*]Q[*longitudinal*]. To save space, we group them as follows: E[*stripes*]Q[*no*]Q[*dark*]Q[*longitudinal*].

We ignored complex EQs, *i.e.,* those composed by complex *Qualities*, which recursively contain *Qualities* linked to other *Entities*. For example, Sentence 4 in Figure 10 has a complex EQ formed by E[*spot*] Q[*centered at the base of*] E[*caudal fin*]. This kind of phenotype construction requires further efforts and expertise to produce annotation. In particular, complex EQs are not treated by our approach and to avoid misinterpretations in the numerical evaluation, they are not computed.

We applied our method to each annotated KQ. We compared the EQs recognized by our method with the annotations of the Gold Standard. The comparison considers four indicators:

- **True Positive (TP):** elements correctly identified. For Example: our method identified in Sentence 1 the following EQs: $E[lips]Q$[*not fringed*]; $E[mouth]Q[horizontal]$. These elements were actually annotated in Sentence 1 of the Gold Standard;
- **False Positive (FP):** An expression recognized by the method as a phenotype, which does not appear as such in the Gold Standard. Example: in Sentence 3, our approach recognized $E[vertebrae]Q$[*132*], which is a *Quality* that slightly differs from the expected one;
- **False Negative (FN):** those phenotypes annotated in the Gold standard that were not detected by the method. Example: E[*breast melanophores*]$Q[Scattered]$ should be identified in Sentence 4 and we failed in recognizing it.

The computation of TP, FP and FN allows calculating the following traditional measures:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Table I presents the obtained results of Precision, Recall and F-measure. The column "EQ pair" compute the recognition of complete *Entity-Quality* pairs and the column "Entity" computes the recognition of *Entities* alone without related *Qualities*.

TABLE I: Results concerning Perfect Matches.

Measures \ Elements	EQ pair	Entity
Recall	0,39	0,69
Precision	0,75	0,85
F-measure	0,51	0,76

The results are influenced by EQs partially recognized. These cases might not be considered a totally wrong result, then we added a more flexible partial match count:

- **Partial Matches (PM):** The cases where the recognized element contains only a part of the expected one. Example: in sentence 3, the E[*vertebrae*]Q[*132*] was recognized, but the *Quality* is not complete.

As a consequence, the measures are adapted as follows:

$$\textit{Total Precision} = Precision + \frac{PM}{TP + PM + FP} \tag{4}$$

$$\textit{Partial Recall} = Recall + \frac{PM}{TP + PM + FN} \tag{5}$$

$$F\text{-}measure = \frac{2 * \textit{Total Precision} * \textit{Total Recall}}{\textit{Total Precision} + \textit{Total Recall}} \tag{6}$$

Tables II presents the obtained results considering partial matches. As expected, in an overall analysis, results reached with partial matches overcome the results of exact matches. We note that better results yield mostly by the Recall.

TABLE II: Total Results.

Measures \ Elements	EQ pair	Entity
Total Recall	0,45	0,76
Total Precision	0,87	0,94
Total F-measure	0,59	0,84

The results are affected by many factors, among them: the range of terms in the domain covered by the ontology. In our case, the universe of *Quality* terms is more vast than those available in PATO.

A study comparing our results with related work is hampered by the unavailability of a Gold Standard set. However, it is possible to compare the proposals conceptually. Among the existing approaches, the most related is the *CharaParser* – part of the *Phenoscape* project – which has a good acceptance in the community. Our work presents more independence of human action in identifying the EQ elements, since *CharaParser* requires some steps of validation by the user over the extracted information, to feed the next steps.

Our approach demands further refinements to identify more EQ elements. Among them, we highlight the need of:

- extending the EQ formalism to handle complex EQs with compound *Entities* and *Qualities*;
- developing a method to perform an Entity Linking task to handle complex cases, *e.g.*, S[*first four dorsal spines prolonged, the second and third longest*]. This sentence requires identifying that the words *second* and *third* implicitly mention the E[*dorsal spine*];

B. Knowledge Network Analysis

In this section, we present practical applications, which are possible due to the extraction of phenotypes. The objective is demonstrating the usefulness of explicitly recognizing EQs. The knowledge network was created by correlating the detected EQs with other information elements available in *FishBase*. In particular, we correlated EQ pairs with data concerning taxonomic groups of fishes. Afterwards, we generated different information visualizations/perspectives to evaluate the obtained correlations. We selected specific cases to highlight the relevance of considering EQ statements.

Figure 11 shows a graph model derived from *FishBase* (*FishGraph*), created by a previous work of Cavoto *et al.* [21]. It highlights the node types (*class, order, family, species, genus, country, key, and ecosystem*) and relationships among them. We have added new nodes to FishGraph – *keyQuestion*, *EQ*, *Entity*, and *Quality* – and linked them to the existing ones.

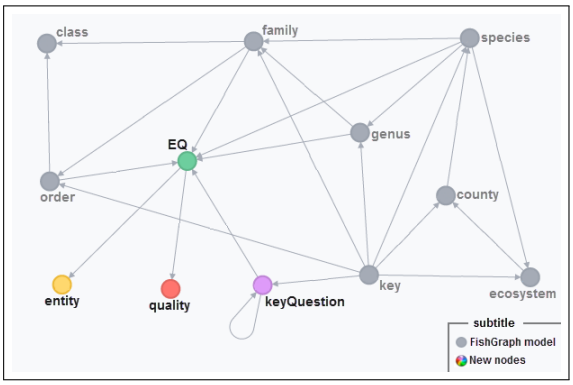

Fig. 11: New nodes added in the *FishGraph* database.

This updated *FishGraph* version models our knowledge network. It allows scientists to perform analyses making use of the new extracted information. We further present examples of possible applications to improve the system usage by the user and analyses to understand facts about living beings.

1) No need of previous knowledge: Currently, each *IK* is represented in *FishBase* as an independent tree, which hampers their usage, as one needs to know beforehand the main taxon (the root of the IK tree) to start the identification process.

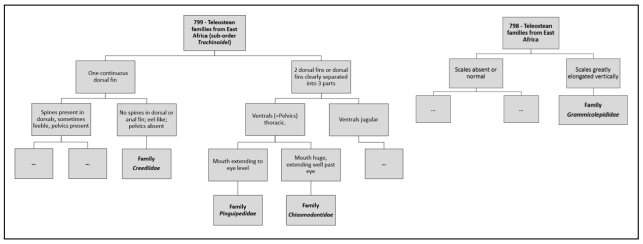

Fig. 12: Part of *Identification Keys*: 799 of Teleostean families from East Africa (sub-order *Trachinoidei*) and 798 of the Teleostean families from East Africa.

As an example, consider the IK "799 – Teleostean families from East Africa (sub-order Trachinoidei)" (*cf.* Figure 12). The identification process using this key requires the following knowledge: the specimen belongs to the teleostean family, sub-order Trachinoidei and it is found in East Africa. The identification process is hampered if the user knows only part of the root – suppose family and geographic location – since FishBase has another 6 IKs of the teleostean family distinguished mainly by the sub-order, *e.g.*, IK 798 also in Figure 12. Even with all the required knowledge, it is necessary to follow the proposed path in the IK tree. All these particularities make the identification process only possible to specialists.

The new structure allows starting the identification process from any known characteristic. For instance, we can start the identification process using a known characteristic like *dorsal fin soft*, independently of any IK or other characteristic.

2) Searching through incremental filtering: The generated knowledge graph allows searching for specific taxons by applying an incremental filtering process.

Figure 13 shows an example of this incremental filter using *Entities* and *Qualities*, which leads to a family with three specific characteristics. Figure 13.a shows the initial filtered graph with 27 families of species that have the $E[dorsal\ fin]Q[soft]$ (*Entities* and *Qualities* are collapsed in a single node, in order to simplify the view). Adding a second filter of the $E[anal\ fin]Q[soft]$ (Figure 13.b) means to select those species with edges to both EQs. The number of families with both characteristics decreases to 8. A third filter of the $E[body\ scale]$, results in only 1 family that has the 3 characteristics: *Creediidae* (Figure 13.c).

3) Relation of taxons and IKs: One taxon is referred in many IKs in *FishBase* but, since they are independent, each IK has its own set of characteristics. When we analyse IKs referring to the same taxon, there are two possible cases: (i) keys share partially or totally the characteristics of a given taxon; (ii) keys that have complementary information about the taxon.

Our unified graph structure links distinct characteristics of the same taxonomic group, coming from many independent IKs, enriching and facilitating the identification process. Returning to the previous experiment, the $E[body\ scale]$ is a characteristic that belongs to IK 324 but it does not belong to IK 799. Since they refer to the same taxonomic group, it is possible to combine them to achieve a more complete description of the taxons.

4) Phenotypes distinguishing taxons: Figure 14.a shows a fragment of the obtained knowledge network highlighting 3 classes of fishes and the EQ elements concerning the *tooth* structure. As can be seen, our approach enabled to unify the *Entities* and it is possible to verify that all 3 classes share the same EQ elements. However, if we drill down to the level of family, it is possible to verify which EQ elements distinguish the two families *Aulopiformes* and *Cetomimiformes* – the size of the tooth: the first one (Figure 14.b) is large and the second (Figure 14.c) is small.

5) Sharing EQs through taxons: We built a bipartite network consisting of two different types of nodes: species and

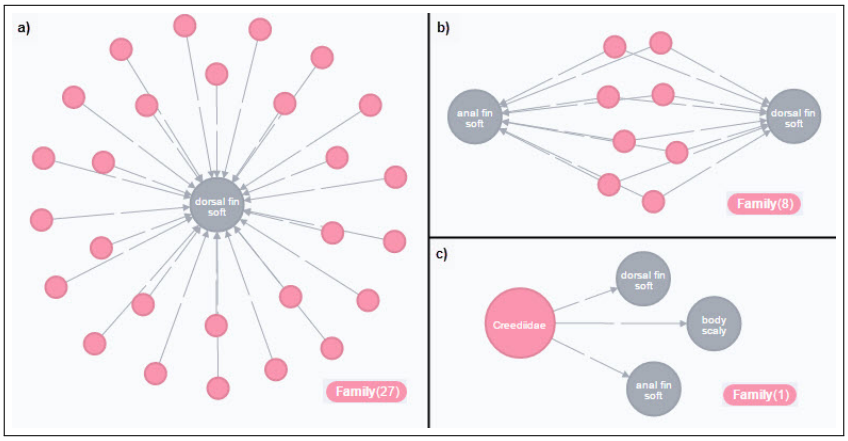

Fig. 13: Filtering families of species by EQ: a) dorsal fin soft; b) dorsal fin soft and anal fin soft; and c) dorsal fin soft, anal fin soft, and body scale.

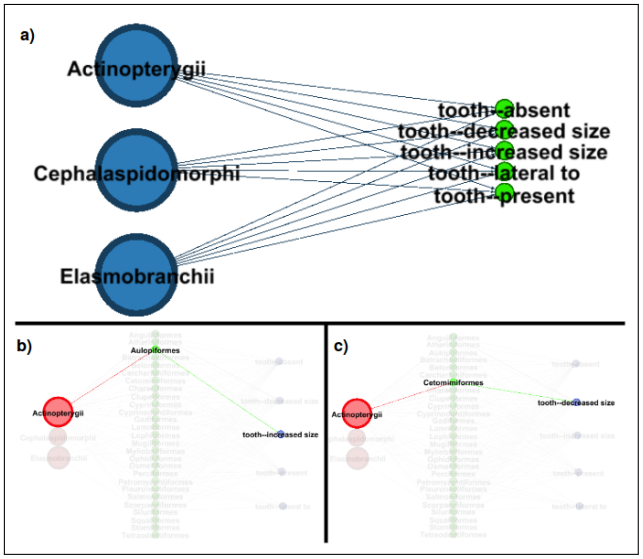

Fig. 14: (a) Relation between a set of EQ elements and classes. (b) EQ element determiner of Aulopiformes Family. (c) EQ element determiner of Cetomimiformes family.

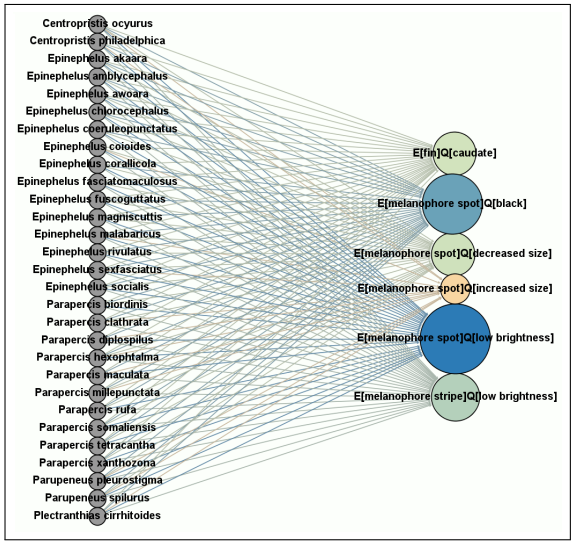

Fig. 15: Bipartite network of Species and EQs elements, showing some of the most present EQ in the species.

EQ statements. In this network, each EQ element is linked to the species that has it.

Figure 15 shows a small portion of this network, in a synthetic view. Since several EQ elements are shared by a large number of species, the resulting bipartide network is too dense for direct visualization. While 29 species are on the left side, 6 EQ pair elements are on the right side. This network enables visualizing which EQ pairs are the most shared by the species.

In the visualization aspect, the size of the EQ nodes indicates the amount of linked species, *e.g.*, the *E[melanophore spot]Q[low brightness]* is the biggest node, which means that it is an EQ pair present in many species.

Figure 16 shows a projection of the bipartite network. In this visualization, the nodes are EQs and they are connected if they are present together at least one species. The link width is proportional to the amount of shares. The size of the nodes indicates the prevalence of the EQ elements in species.

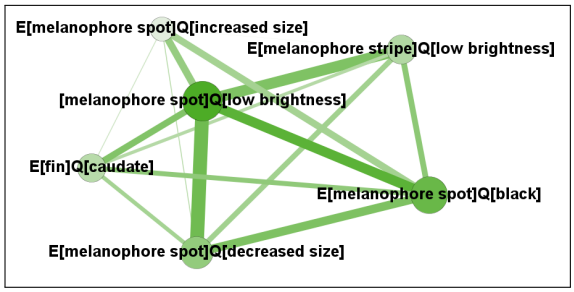

Fig. 16: Projection of the Bipartite network showing the most shared EQ elements by species.

This visualization allows to study which EQ elements frequently occur together. For example, the link width between the nodes E[*melanophore spot*]Q[*low brightness*] and E[*melanophore spot*]Q[*decreased size*] indicates that they are EQ elements present together in many species.

VI. Conclusion

Phenotype descriptions play a key role in biological knowledge bases, but most of the descriptions remain in a free-textual format, which affects machine interpretation and their applicability in network-driven analyses.

This paper proposed an original approach to recognize *Entities* and *Qualities* connecting them to concepts in ontologies to make their representation semantically interpretable by machines. Our key point, not addressed by related work found in literature, consists in our approach, which explored clues of non-textual information: from the writing characteristics of phenotype descriptions to their organizational structure.

The experimental evaluations revealed encouraging results regarding the assessment against a gold standard set. The experiments point out the contributions of each step to improve the results of the recognition process. The experiments using the EQ elements, extracted from free-text sentences applying our proposal, showed the advantages of bringing these descriptions to a common and formal language. It enables machines better consuming and interpreting the available descriptions.

Future work involves conducting further evaluations to measure and compare the efficiency concerning to other approaches. It also aims at addressing limitations of the recognition of EQ elements.

Acknowledgment

Work partially financed[4] by CNPq (134205/2015-4), FAPESP (2014/14890-0), FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), FAPESP Storage, Modeling and Analysis of Dynamical Systems for e-Science Applications project (2014/08285-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), CNPq (Descrições Complexas na Web), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from CNPq.

References

[1] S. Pyysalo and S. Ananiadou, "Anatomical entity mention recognition at literature scale," *Bioinformatics*, vol. 30, no. 6, pp. 868–875, 2014.

[2] A. Grand, R. V. Lebbe, and A. Santanche, "From Phenotypes to Trees of Life: A Metamodel-Driven Approach for the Integration of Taxonomy Models," in *IEEE 10th International Conference on e-Science*, vol. 1, 2014, pp. 65–72.

[3] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the phenochf corpus," *BMC Medical Informatics and Decision Making*, vol. 15, no. Suppl 2, p. S3, 2015.

[4] E. J. Farnsworth, M. Chu, W. J. Kress, A. K. Neill, J. H. Best, J. Pickering, R. D. Stevenson, G. W. Courtney, J. K. VanDyk, and A. M. Ellison, "Next-generation field guides," *BioScience*, vol. 63, no. 11, pp. 891–899, 2013.

[4]The opinions expressed in this work do not necessarily reflect those of the funding agencies

[5] Lucid. (2016) Lucid phoenix. [Online]. Available: http://www.lucidcentral.com/

[6] V. Ung, G. Dubus, R. Zaragüeta-Bagils, and R. Vignes-Lebbe, "Xper2: introducing e-taxonomy," *Bioinformatics*, vol. 26, no. 5, pp. 703–704, 2010.

[7] S. Martellos and P. Nimis, "Keytonature: teaching and learning biodiversity. dryades, the italian experience," in *Proceedings of the IASK International Conference Teaching and Learning*, 2008, pp. 863–868.

[8] M. J. Dallwitz, T. Paine, and E. Zurcher. (2000) Principles of interactive keys. [Online]. Available: http://biodiversity.uno.edu/delta

[9] W. Dahdul, T. A. Dececchi, N. Ibrahim, H. Lapp, and P. Mabee, "Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy," *Database*, vol. 2015, p. bav040, 2015.

[10] M. Ciaramita, A. Gangemi, E. Ratsch, J. Šaric, and I. Rojas, "Unsupervised learning of semantic relations between concepts of a molecular biology ontology," *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 659–664, 2005.

[11] M. Song, H. Yu, and W.-S. Han, "Developing a hybrid dictionary-based bio-entity recognition technique," *BMC Medical Informatics and Decision Making*, vol. 15, no. Suppl 1, p. S9, 2015.

[12] C. Ramakrishnan, P. N. Mendes, S. Wang, and A. P. Sheth, "Unsupervised discovery of compound entities for relationship extraction," in *Knowledge Engineering: Practice and Patterns*. Springer, 2008, pp. 146–155.

[13] K. Fundel, R. Kuffner, and R. Zimmer, "RelEx - Relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[14] H. Cui, "Charaparser for fine-grained semantic annotation of organism morphological descriptions," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 4, pp. 738–754, 2012.

[15] M.-C. De Marneffe and C. D. Manning, "Stanford typed dependencies manual," *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[17] W. M. Dahdul, J. G. Lundberg, P. E. Midford, J. P. Balhoff, H. Lapp, T. J. Vision, M. A. Haendel, M. Westerfield, and P. M. Mabee, "The teleost anatomy ontology: anatomical representation for the genomics age," *Systematic Biology*, vol. 59, no. 4, pp. 369–383, 2010.

[18] T.-L. Wong, W. Lam, and T.-S. Wong, "An unsupervised framework for extracting and normalizing product attributes from multiple web sites," in *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2008, pp. 35–42.

[19] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*, pp. 135–187, 2010.

[20] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[21] P. Cavoto, V. Cardoso, R. V. Lebbe, and A. Santanchè, "FishGraph: A Network-Driven Data Analysis," in *11th IEEE International Conference on e-Science*, 2015.