Semantic Interpretation of Biological Identification Keys

Fagner L. Pantoja, Julio Cesar Dos Reis, André Santanchè

[1]Institute of Computing – University of Campinas (UNICAMP)
Campinas, SP – Brazil

pantoja.ti@gmail.com, {julio.dosreis,santanche}@ic.unicamp.br

Abstract. *In biological data, Identification Keys (IKs) are central artifacts used by biologists to identify the taxonomic group of an observed specimen, such as family, order, species, etc. Despite their relevance, IKs are usually defined in a semistructured textual format, which does not favor easily retrieval and deep analysis over their data. This article aims to present a method to formally structure and extract semantic facts from IKs relying on graphs and domain ontologies. The approach explores classical extraction and matching procedures combined with the specific characteristics of IKs. Initial experiments reveal the feasibility of the approach.*

1. Introduction

FishBase[1] is a global information system recording dozens of information of almost all fishes known to science [Froese and Pauly 2000]. Among several types of data managed by *FishBase*, *Identification Keys* (IKs) consist of artifacts created by biologists to identify the species or any other taxonomic group (called taxon) of an observed specimen [Pyysalo and Ananiadou 2014]. For identifying a specimen via IKs, users might navigate through a series of multiple choice questions. According to the picked answers, the path lead to the respective taxon. Currently, *FishBase* contains 1,668 IKs of fishes. They are one of the most relevant artifacts to support biological research.

Despite their richness, they remain in free text format and are suitable only to be consumed by humans. Therefore, computers making automatic analysis are unable to interpret the underlying concepts and their semantics, *e.g.*, distinguish characters and their states inside a text. In this article, we propose an original method to extract structured and semantically enriched information from IKs. The method detects the concepts and transforms them in a semantic-based description with the support of domain ontologies. Furthermore, this research explores the intrinsic characteristics of IKs, such as the fact that IKs of similar taxons frequently share some characters. We assume that these peculiarities can be helpful to aid the extraction of the phenotype descriptions carried by them.

This article is organized as follows: Section 2 states the problem and describes the related work. Section 3 presents our method. Section 4 shows our experimental evaluation and Section 5 presents our conclusions and future work.

2. Problem Definition and Related Work

As an example of IK usage, Figure 1 presents an IK that identifies the sub-order *Trachinoidei of Teleostean family* from East Africa. The process of identification begins with

[1]www.fishbase.org

the question 1, that has the couple of answers 1a and 1b with their descriptive texts. Depending on the picked answer, the user might navigate to question either 2 or 4, in the column *Next*. Each descriptive text is called of Key Question (KQ). This step is repeated until one reaches a row that does not lead to another question. At this stage, the specimen was identified and the group name found is at the column *Link*.

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

Figure 1. Fragment of Identification Key to the *Teleostean families* from East Africa (sub-order Trachinoidei). Source: www.fishbase.org

Currently, IKs are described as a list of observable characters and their states in a free-form text. They have been inserted into the system without a pattern, in such way that one character can be written in different ways. For example, "*median fin skeleton*" has the variations: "*unpaired fin skeleton*" and "*axial fin skeleton*".

This representation hampers the interpretation, retrieval and correlation of data related to IKs. Even though there are formalisms to represent the characters and their states with explicit and interoperable semantics related to ontologies – as the Entity-Quality (EQ) formalism [Grand et al. 2014] – there is the challenge faced in this work of how to automatically convert descriptions in EQ sentences. An EQ representation will provide the following benefits:

- **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, like the *Teleostean* family (Figure 1). Therefore, users must know beforehand the taxon to which the specimen belongs to pick a correct IK. This process is laborious and limits the use of the system only to expert biologists. An EQ representation will enable to correlate characters of several IKs and combine them in a unified identification tree.
- **Relation of taxons and keys:** With the unified and semantic enriched EQ, it will be possible to perform analysis to understand facts including: (i) which IKs identify similar taxonomic groups; (ii) which characters are determinant to discriminate a taxon of a specimen; (iii) which characters join a specific taxon.

Despite to the recent advances in literature, existing methods are unable to completely tackle the addressed problem. [Dahdul et al. 2015] investigate ways of transforming the descriptive biology data in a format that enable large-scale computation. They advocate the real need for efficient methods to automatically extract the phenotype from descriptions to reduce the efforts of achieving such large-scale computable format. Differently, [Pyysalo and Ananiadou 2014] propose a machine learning-based method to extract the entity anatomy from a scientific paper corpus. Our approach differs in the sense that it explores the specificity of an IK structure to improve the automatic recognition.

3. Proposed Approach

We propose a technique to extract implicit semantics from semistructured IKs, in order to map their terms and relations to a more formal representation with explicit semantics, based on domain ontologies. The technique starts representing the original data as a graph and applies successive graph transformations towards the final formal representation, as illustrates Figure 2. We define a three-step method, as follows.

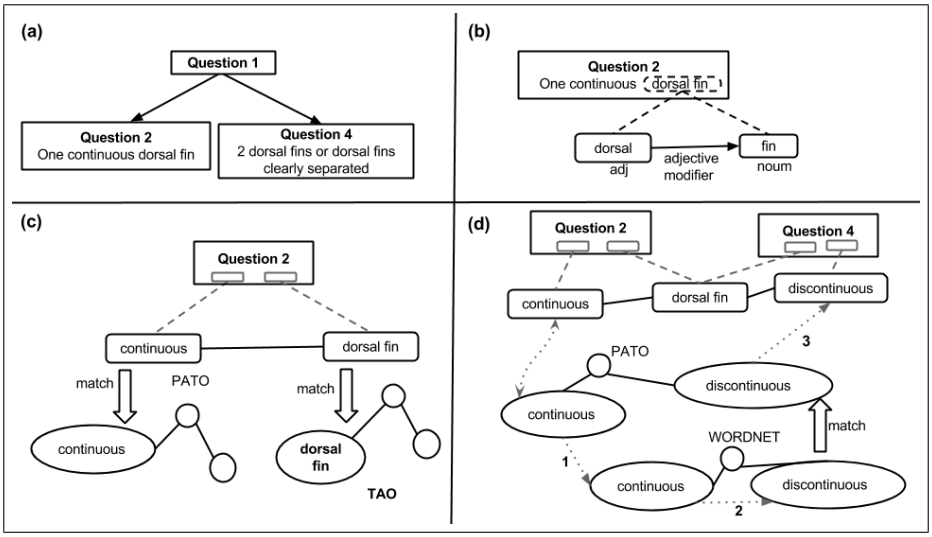

Figure 2. General view of the proposed approach.

3.1. Step 1: graph representation

This step transforms raw relational data into a graph representation. Figure 2(a) presents the graph resulting from the IK showed in Figure 1. Nodes are Key Questions (KQs) containing descriptive texts and edges link them to other questions or taxa according to answer choices.

3.2. Step 2: extraction of EQ

This step extracts an *Entity-Quality* (EQ) representation from free-text descriptions inside of each KQ. EQ is a formalism to describe organisms (further detailed), which make their semantics explicit by matching terms with domain ontologies [Grand et al. 2014]. IKs contain phenotype descriptions, which are composed by characters and their states. Characters can be anatomical structures, such as *dorsal fin*. States are qualifiers of these characters, such as *absent*. FishBase descriptions do not clearly distinguish character and states inside the text. This step is divided in 2 phases following described.

3.2.1. Step 2 – Phase 1: performing the NLP parser

The first graph transformation – from Figure 2(a) to 2(b) – involves detecting characters and character states (C/CS) within the text. It uses a natural language toolkit to extract typed dependencies [De Marneffe and Manning 2008]. We have chosen it due to its ability of capturing the relation between terms inside a sentence, reflecting the general form of terms in most of the ontologies used in this work, i.e., compound of multiple words,

instead of isolated words. Figure 2(b) shows part of the parser's output graph for the question "*One continuous dorsal fin*". The extracted dependency captures the relation between an adjectival term "dorsal" (node) modifying (edge) a noun term "fin" (node). Preliminary experiments showed that this type of relations are good candidates to be identified as EQs. This assumption was confirmed in the next phase.

3.2.2. Step 2 – Phase 2: matching with ontologies

This phase matches identified terms with ontologies. The technique explores string-based methods for the matching. An Entity-Quality representation relates an Entity – the subject described, which is usually an anatomical part of the organism – to an observed Quality. Therefore, two different ontologies are explored, which has been chosen based on their domains and acceptance in community:

1. Anatomical to match entities, *e.g.* *Teleost Anatomy Ontology* (TAO) [Dahdul et al. 2010]
2. *Phenotype Quality Ontology* (PATO) [Mungall et al. 2010] to match the qualities.

The match is enhanced using the root of the involved terms, retrieved using the same language toolkit of the previous phase. At the best case, the method might relate the entity together with its quality. In some cases, it can only relate either the entity or the quality. At the worst case, the method fails relating both in the KQ. The EQs matched to the domain-ontologies are attached to the graph generated from step 1. Figure 2(c) shows a piece of the output graph produced at this phase, where the confirmed EQ terms are added as new nodes to the graph. At the current stage of this project, our method does not handle logical connectors (e.g., *and* or *not*) contained within a KQ. That issue will be addressed in the next stages of this project.

In the Question 2 (Figure 2(a)), which contains the text "*One continuous dorsal fin*", our method extracted the quality "*continuous*" and the entity "*dorsal fin*". However, the method was not able to identify EQ terms in the Question 4, which has the text "*2 dorsal fins or dorsal fins clearly separated*". Then, the next step tries to identify those EQ that was not captured, exploring the IKs' structure.

3.3. Step 3: exploring the IKs peculiarities

The goal of step 3 is to enrich the graph output from step 2 with further elements. We aim at exploring IKs' intrinsic characteristics to increase the rate of recovered and connected terms. We found some of those aspects in an initial analysis. In this approach, we consider two of them: (a) answers of a question frequently refers to the same characters; (b) answers of a question are complementary, in the sense that they frequently mean opposite options to the question. These aspects lead to the following heuristics:

1. If an entity is identified in a node, the same entity might also be present in its sibling nodes.
2. If an EQ pair is identified in a node, then either (a) the pair might also be present in its siblings, with the quality replaced by a word that represents a complementary or opposite meaning; or (b) only a quality with a complementary or opposite meaning might be present in the sibling KQ, linking it to an already recognized entity.

3. If a quality is identified without the respective entity, a quality with a complementary or opposite meaning might be related to the sibling nodes.

In our procedure, we retrieve terms with opposite meaning using the *WordNet* lexical database [Miller 1995], getting antonymous, in the following way: for each quality already contained in our graph, we look for it at *Wordnet* (Figure 2(d) arrow 1). The retrieved term has a set of antonyms (Figure 2(d) arrow 2). For each retrieved antonym, we match it with the PATO ontology in order to discover the resource representing that term. If the term is found in PATO, then we relate it to the node (Figure 2(d) arrow 3).

As Figure 2(d) shows, this step was able to detect the quality "*discontinuous*" as opposite of "*continuous*" present in the sibling node. Note that the original term extracted from the sentence was "*separated*" instead of "*discontinuous*". However, they have the same meaning, based on explicit relations stated by ontologies.

4. Experimental Evaluation

We conducted a preliminary evaluation to assess the viability of our method. The experiment was performed over a total of 1,659 IKs containing 25,542 KQs from *FishBase*. Table 1 shows the obtained results. The metrics were divided by types of recognized elements (in rows), which are: Entity without a related Quality (*e.g.* caudal fin), Quality without a related Entity (*e.g.* pointed) or a compound of an Entity related to a Quality (*e.g.* continuous dorsal fin).

Column *Amount* shows how many elements (entities or qualities) were recognized. Column *Ratio* presents the average of extracted elements from each KQ. Column *Coverage* shows the rate of KQs containing at least one element extracted, varying from 0 to 1. All of these metrics were divided in Step 2 and Step 3. Then, it is possible to observe the difference of using only NLP (Step 2) and how much we can increase exploring IKs' characteristics (Step 3), which is our main contribution.

Table 1. Results

Elements \ Metrics	Amount		Ratio		Coverage	
	Step 2	Step 3	Step 2	Step 3	Step 2	Step 3
Entity	30,747	41,611	1.2	1.62	0.61	0.7
Quality	20,177	24,895	0.78	1.0	0.43	0.46
EQ	15,239	17,267	0.6	0.67	0.36	0.4

Results reveal that the method allows extracting EQ from IKs, although the rate is still relatively low. But it is important to observe that we exploited only a small part of the potential of our method. Research efforts will be devoted to refine mostly Steps 2 and 3. It involves improving the matching algorithms. Moreover, we are considering to explore other kind of relations returned by the NLP parser. Moreover, other IKs' characteristics can be explored in Step 3 and the search for recognized Entities and Qualities can be extended to other branches of the tree beyond the siblings. Further experiments are required to evaluate the accuracy of the approach. We plan to involve domain experts to manually annotate EQs in IKs to create a corpus of reference. The results of our method will be then compared with the annotated corpus. Traditional measures like precision and

recall will be computed to validate the effectiveness, as well as to compare our results with others in literature.

5. Conclusion

Identification keys play a central role to enable identifying and assessing biological specimens. However, in FishBase they are only available in a textual format limiting automated analysis. This paper proposed an automatic method to leverage IKs formal representation from texts, structuring data via graphs and making the semantics of concepts explicit via ontologies. A preliminary evaluation showed its feasibility. Future work involves examining further aspects that might generate additional rules to improve the accuracy of the method and thoroughly evaluate it. Even though this method have been developed inside the FishBase context, it was designed to be generalized to a wider spectrum of biological information systems.

Acknowledgements Work partially financed[2] by CNPq (134205/2015-4), FAPESP (2014/14890-0), FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual grants from CNPq.

References

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

Dahdul, W. M. et al. (2010). The teleost anatomy ontology: anatomical representation for the genomics age. *Systematic Biology*, 59(4):369–383.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Froese, R. and Pauly, D. (2000). *FishBase 2000: concepts, design and data sources*.

Grand, A., Lebbe, R. V., and Santanche, A. (2014). From phenotypes to trees of life: A metamodel-driven approach for the integration of taxonomy models. In *IEEE 10th International Conference on e-Science*, volume 1, pages 65–72.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, 11(1).

Pyysalo, S. and Ananiadou, S. (2014). Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

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