

Generating Knowledge Networks from Phenotypic Descriptions

Fagner Leal

Patrícia Cavoto, Julio dos Reis, André Santanchè
pantoja.ti@gmail.com

Laboratory of Information Systems
University of Campinas
Campinas - São Paulo - Brazil

October 24, 2016

Research Scenario

Phenotype Descriptions

Research Scenario

Phenotype Descriptions

- ▶ Morphological structures
- ▶ Behavior traits
- ▶ Life cycles; *etc.*

Research Scenario

Phenotype Descriptions

- ▶ Morphological structures
- ▶ Behavior traits
- ▶ Life cycles; *etc.*

Examples:

1. *No dark longitudinal stripes on head and body.*

Research Scenario

Phenotype Descriptions

- ▶ Morphological structures
- ▶ Behavior traits
- ▶ Life cycles; etc.

Examples:

1. *No dark longitudinal stripes on head and body.*
2. *Scattered breast melanophores (Fuiman et al., 1983).
Pteronotropis hubbsi can also be distinguished from Notropis chalybaeus by the presence of two caudal spots, one large spot centered at the base of the caudal fin below the flexed notochord and a smaller spot located dorsally above it, and by the presence of 9 dorsal rays in late metalarvae. Notropis chalybaeus has a single caudal spot in which no part extends above the notochord and 8 dorsal rays (Marshall, 1947).*

Research Scenario

- ▶ Biology Knowledge Bases

¹<http://www.fishbase.org>

Research Scenario

- ▶ Biology Knowledge Bases
e.g., FishBase: knowledge base about fishes¹

¹<http://www.fishbase.org>

Research Scenario

- ▶ Biology Knowledge Bases
e.g., FishBase: knowledge base about fishes¹
- ▶ Identification Keys (IK)s

¹<http://www.fishbase.org>

Research Scenario

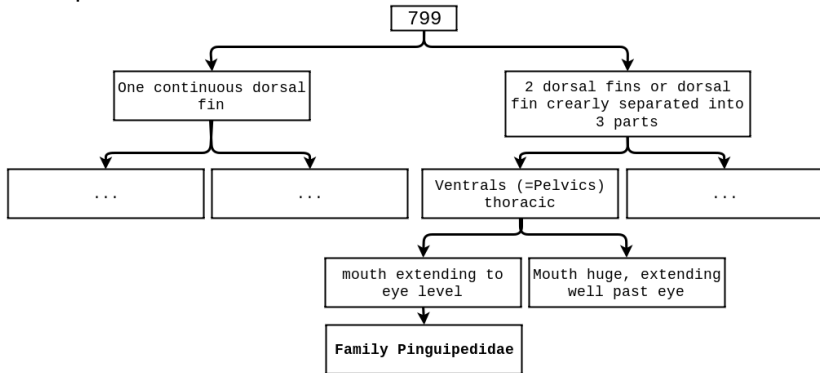
- ▶ Biology Knowledge Bases
e.g., FishBase: knowledge base about fishes¹
- ▶ Identification Keys (IK)s
 - ▶ Artifacts to identify specimens
 - ▶ Observable characteristics

¹<http://www.fishbase.org>

Research Scenario

Identification Keys

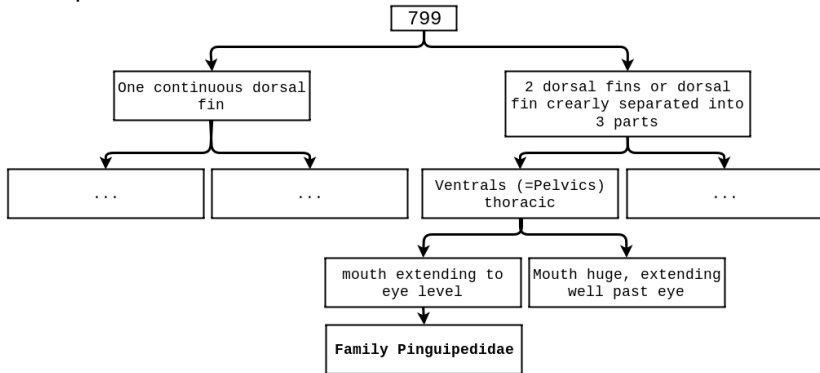
Example of IK to Teleostean families



Research Scenario

Identification Keys

Example of IK to Teleostean families



Drawbacks:

- ▶ Need **previous knowledge**
- ▶ Need to **follow the flow**

Goal

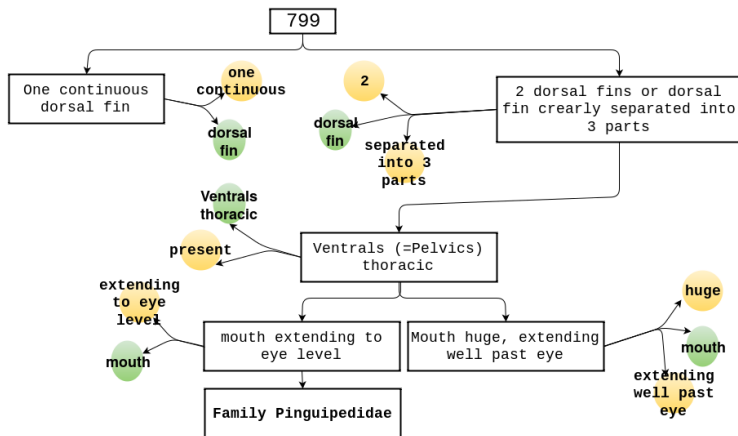
To recognize and explicit phenotype elements locked in the Identification Keys. Using the Entity-Quality (EQ) representation:

- ▶ **Entity**: morphological structure
- ▶ **Quality**: qualifier state of the *Entity*

Goal

To recognize and explicit phenotype elements locked in the Identification Keys. Using the Entity-Quality (EQ) representation:

- ▶ **Entity**: morphological structure
- ▶ **Quality**: qualifier state of the *Entity*



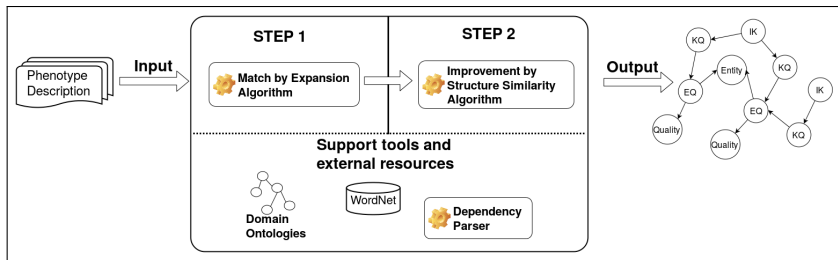
Related Work

Information Extraction

Reference	Context	Approach
Ciaramita <i>et al.</i> , 2005	Interactions in molecular biology	Unsupervised Learning and Rules over Dependency Trees
Song <i>et al.</i> , 2015	Biomedical anatomic entities	Dictionary-based
Pyysalo and Ananiadou, 2014	Biomedical Anatomic entities	Supervised learning
Ramakrishnan <i>et al.</i> , 2008	Biomedical Anatomical entities	Dictionary-based, Rules over Dependencies Trees and Statistical Learning
Fundel <i>et al.</i> , 2007	Gene and Protein Interaction	Rules over Dependency Trees
Cui, 2012	Morphological structures of organisms	Unsupervised Learning

Method

General View



Step 1:

It explores isolated sentences

Step 2:

It explores the sentence correlations

Method

Step 1 - General View

Assumption:

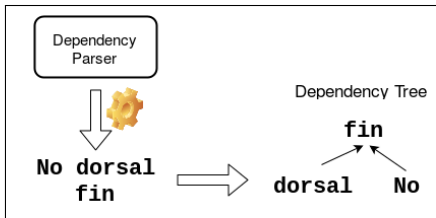
The typical way in which phenotype descriptions are written can guide the extraction of EQ elements.

Method

Step 1 - General View

Assumption:

The typical way in which phenotype descriptions are written can guide the extraction of EQ elements.

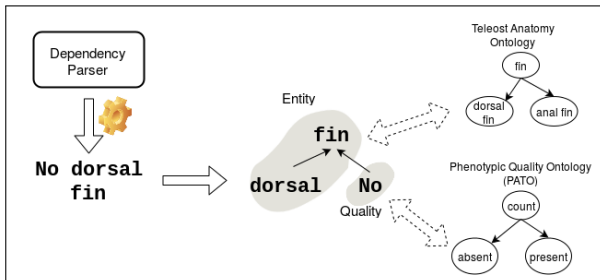


Method

Step 1 - General View

Assumption:

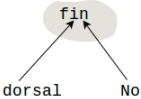
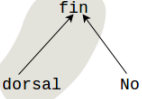
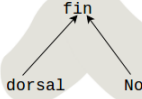
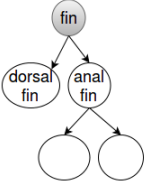
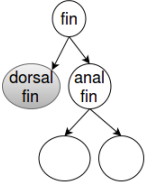
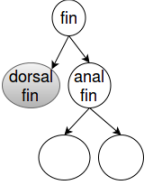
The typical way in which phenotype descriptions are written can guide the extraction of EQ elements.



Method

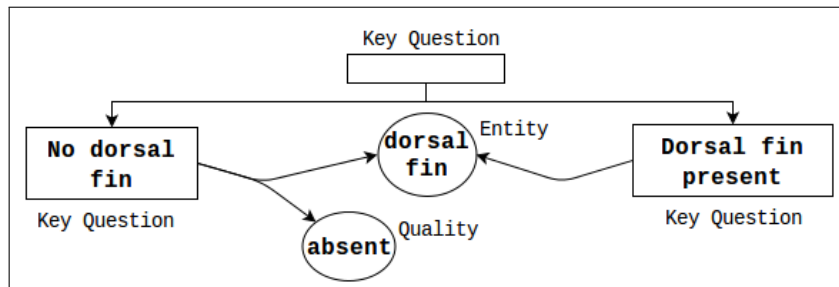
Step 1 - Match Algorithm

Identifying **Entities** and **Qualities**:

	Iteration 1 Vertexes: fin Concept: fin Similarity: 1	Iteration 2 Vertexes: dorsal fin Concept: dorsal fin Similarity: 1	Iteration 3 Vertexes: no dorsal fin Concept: dorsal fin Similarity: 0.76
Dependency Tree			
Teleost Anatomy Ontology (TAO)			

Method

Step 1 - Output



Method

Step 2 - General View

Assumption:

The structure of Identification Keys holds correlations that can be exploited to improve the extraction of EQ statements.

Method

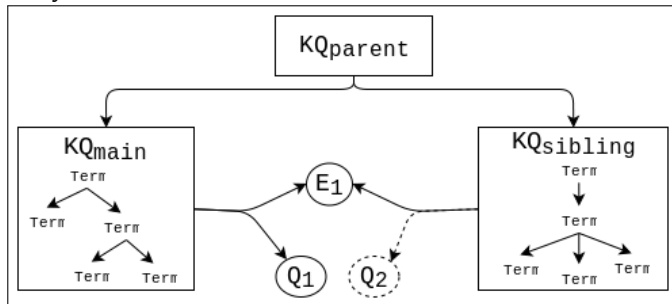
Step 2 - General View

Assumption:

The structure of Identification Keys holds correlations that can be exploited to improve the extraction of EQ statements.

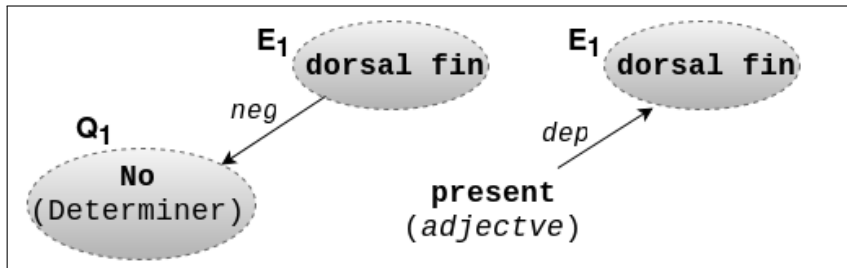
Generally, in phenotype descriptions:

1. Alternative sentences refer to the same *Entities*.
2. Alternative sentences assign complementary *Qualities* to *Entity*.



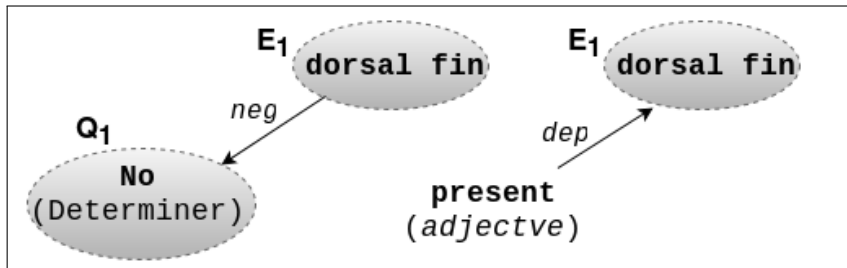
Method

Step 2 - Algorithm



Method

Step 2 - Algorithm

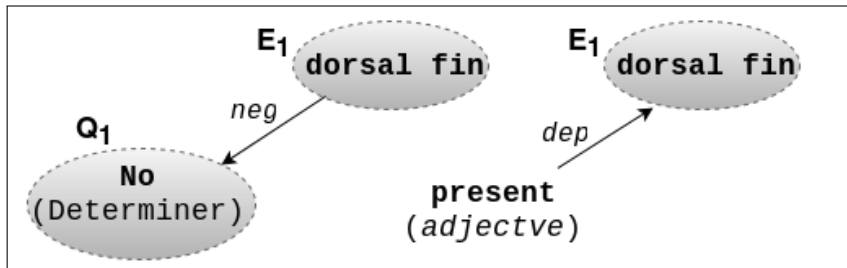


Compare the two relations, based on:

- (a) Existence of antonymy between the quality parts
- (b) Relation Type
- (c) Grammatical classes of quality parts
- (d) Relation Directions

Method

Step 2 - Algorithm



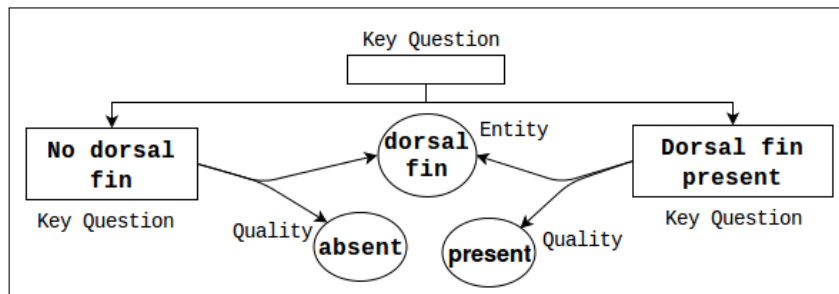
Compare the two relations, based on:

- (a) Existence of antonymy between the quality parts
- (b) Relation Type
- (c) Grammatical classes of quality parts
- (d) Relation Directions

$$Similarity = \sum_{i=a}^d v_i$$

Method

Step 2 - Output



Evaluation - Numerical Assessment

Gold Standard-based Assessment

Gold standard set: 100 phenotype descriptions (randomly selected) were manually annotated

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

Evaluation - Application Experiments

EQ sharing through taxons

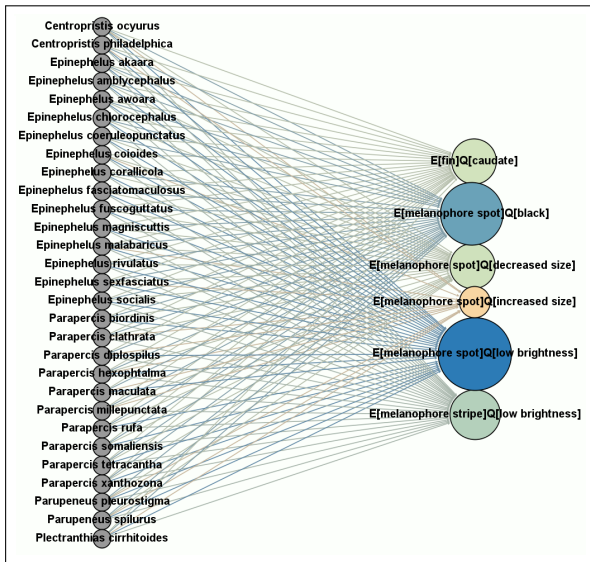


Figure 1: Bipartite network of Species and EQs

Evaluation - Application Experiments

EQ sharing through taxons

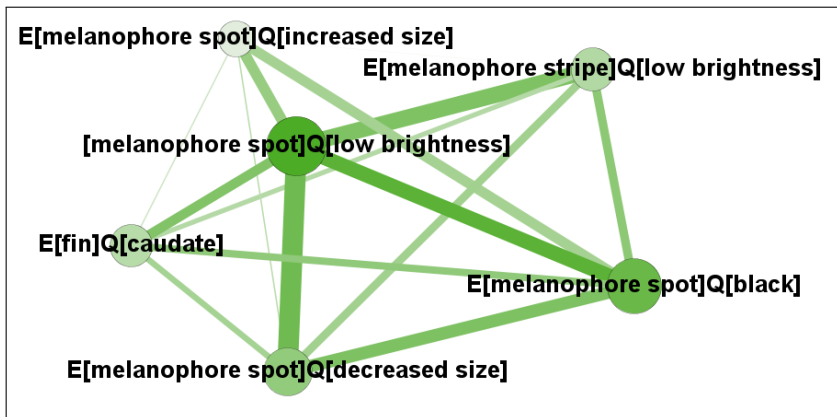


Figure 2: Projection of bipartide network

Conclusion

Original approach to automatically recognize *Entities* and *Qualities*, exploring :

- ▶ Writing characteristics of phenotype descriptions
- ▶ Organizational structure of IKs

Future Work

- ▶ To compare against other approaches
- ▶ To recognize complete EQs in Step 2 (not only the quality part)
- ▶ To calibrate the parameters and thresholds

Thank you!

Classical Measures

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

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

Examples of:

- ▶ True Positive:
 - ▶ expected: *E[lips]Q[notfringed]*
 - ▶ recognized: *E[lips]Q[notfringed]*
- ▶ False Positive:
 - ▶ expected *E[vertebrae]Q[119 to 132]*
 - ▶ recognized: *E[vertebrae]Q[132]*
- ▶ False Negative:
 - ▶ recognized *E[breastmelanophores]Q[Scattered]*

Considering Partial Matches

- ▶ **Complete Miss (CM)**: false negative
- ▶ **Wrong Hit (WH)**: false Positive
- ▶ **Full Match (FM)**: true Positive

$$\text{Partial Precision} = \frac{\text{Partial Match}}{\text{Full Match} + \text{Partial Match} + \text{Wrong Hit}} \quad (4)$$

$$\text{Full Precision} = \frac{\text{Full Match}}{\text{Full Match} + \text{Partial Match} + \text{Wrong Hit}} \quad (5)$$

$$\text{Partial Recall} = \frac{\text{Partial Match}}{\text{Full Match} + \text{Partial Match} + \text{Complete Miss}} \quad (6)$$

$$\text{Full Recall} = \frac{\text{Full Match}}{\text{Full Match} + \text{Partial Match} + \text{Complete Miss}} \quad (7)$$

Considering Partial Matches

Total Precision = Partial Precision + Full Precision

Total Recall = Partial Recall + Full Recall

Measures \ Elements	EQ pair	Entity
Partial-Recall	0.05	0.08
Full-Recall	0.39	0.67
Partial-Precision	0.11	0.1
Full-Precision	0.75	0.84

Table 1: Results concerning Perfect and also Partial Matches

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