Assisted generation of interactive species identification keys from structured morphological descriptions and other sources of information for plants of Costa Rica

Gloriana Zamora-Villalobos
Escuela de Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
Email: glorianazv@gmail.com

José E. Araya-Monge
Centro de Investigaciones en Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
Email: jaraya@itcr.ac.cr

***Abstract*—This article describes how structured morphological descriptions and other pieces of information can be exploited to automatically generate a base interactive key for plants identification. The article explains the usual process of creating an interactive key and the steps needed to generate it in a automated way from existing sources of information, taking advantage of algorithms of information extraction and feature selection.**

***Keywords*—interactive keys, species identification, semantic annotation, morphological descriptions, feature selection**

I. Introduction

Species identification is the process of finding the taxon name of a specimen or entity in question[8]. It is an important process not only because it generates entries to the identified species database, but also because of the information you get from identification: habitat, benefits, ecology, etc. This data is used to a better understanding of the world's biodiversity and how to take advantage of it in other areas like: biology, ecology, medicine, agriculture, etc.

Identification keys are one of the most popular tools for species identification. There are two major types of them: dichotomous keys and interactive keys. The keys in the first group are similar to a decision tree, while the second ones, are essentially a matrix between species and species attributes, also known as characters (color, size, texture, etc.). By selecting the characters a specimen meets, it discriminates between the species that satisfies the attributes selected (candidate species) and those that doesn't (discarded species); leading to a final answer or a few final possibilities[8].

Interactive identification keys, are a way to synthesize existing diagnostic taxonomic information or full taxonomic descriptions, from different sources. Taxonomists and experts in the area create these interactive identification keys usually manually, or with the help of computer aids for data entry and management.

Many initiatives have emerged around building tools to assist the process of creating identification keys. Some of them, like LUCID[4], facilitate data entry and key execution; while others even allows the user to create structured databases and automatically generate keys with them, which is the case of DELTA INTKEY[7]. However, none of these initiatives really avoid the fact that someone has to manually enter the data in the first place. Moreover, there has been other initiatives focused on extracting information from existent literature through different means; for example The Biodiversity Heritage Library[2]. This opportunity has led to the ability to feed systems like the Encyclopedia of Life[1] with valuable information; and make it computer-readable for a more effective use of biodiversity knowledge and better support for biodiversity research.

The process of converting free text morphological descriptions of whole organisms into a computer-readable representation is a conversion task which is commonly called "semantic annotation". Hong Cui[5] has contributed greatly in the subject with techniques for automated annotation or with less human effort. Mora[10] also, achieved to extract semantic annotated data from morphological descriptions from the Manual de Plantas de Costa Rica[9]; that publication, is specially important because it describes species summarising its major taxonomic and diagnostic characters. The semantic annotation process clearly gives a chance to take advantage of the information acquired and use it to create a more tangible product, as an identification key.

This project consists of implementing an algorithm that extracts and selects ideal species attributes and values, from semantically annotated morphological descriptions to generate an interactive key automatically. We already mentioned some of the shortcommings this process has: actual tools for generating keys does not prescinds of the human factor to enter information, this job is usually done by an expert in the area or the target group of species; but it is also worth mentioning that it also takes some effort to extract information, order and standardize the data, and eventually select and refine the characters that will form the identification key. It will be a great opportunity to be able to generate automatically a key that can be used as a base for taxonomists to work from.

In the next sections we will discuss the proposed steps to address those problems and show some preliminary results.

II. Methodology

Manually creating an interactive identification key follows four main steps: data recollection, information fusion, feature selection, and key evaluation. The idea is to apply informatics' techniques to assist the process of generating a base key.

A. Data Recollection

First, a target group of species needs to be selected. An identification key normally targets species of the same family or genre. For this project we chose species from the Passiflora family not only because it is fully documented on the Manual de Plantas de Costa Rica, from which we can easily be extract data using María Mora's algorithm; but also because it will make feasible the evaluation process described later on. Secondly, the extracted semantic annotated morphological descriptions data will be used as a source to pull a series of triplets that consists of: structure name, attribute name and value. This triplets are the input data of the identification key.

The morphological descriptions were chosen as the main source for the construction of the key, but in order to have some supplemental information for feature selection, dichotomous keys and diagnostic description sections from the Manual will be used as well. It is obvious, that dichotomous keys will have important information on species identification, but on the other hand, diagnostic descriptions are small pieces of information that contains useful and specific details on how to recognize them, for example: *"Passiflora arbelaezii is easily recognizable by its petioles with three or four tiny apical glands and small, whole leaf blades with a tiny gland at the end of the middle nerve."*[9]. Morphological descriptions texts have an special writing structure for which the extraction algorithm was developed. Extracted words from both dichotomous keys and diagnostics can be used as parameters for the triplets selection.

B. Information Fusion

In this stage of the process the most important thing is to have clear and concise information. Wei[13] in his doctoral thesis, explains this process more deeply, considering: cleaning, generalizing and normalization; similar to data mining. For this to happen first the data must be cleaned, which means: removing stop words, eliminating duplicated information, performing linguistic (plurals, synonyms, etc) and numeric normalization. For example, gathering up information about flowers coloration we could end up with several variations, for example, for yellow color, like: "amarillo", "amarillas", "amarillentas", etc.; and they all should be converted into just one variation,"amarillo". The same apply to numeric values, they all must have the same unit of measurement. These are some of the problems Hong Cui[6] mentions that appears when dealing with annotated information, and that it must be addressed to generate a useful identification key.

Therefore, the most important tasks in this step are the following:

- Numeric normalization. Generate appropriate measurement scales with standardized units.
- Linguistic normalization. Remove stop words, duplicated information.
- Generate an appropriate classification for similar colors into a scale that makes sense for the user to choose from. The same apply for shapes (leaves shapes).

C. Features Selection

Once data has been cleaned and normalized then the next step is to select those triplets that will go into the identification key. As mentioned earlier the plan is to gather a list of words from both the dichotomous keys and diagnostics, free from stop words and duplicates; and classify them into structures and attributes/values. These will be ranked based on different techniques of dimensional reduction[3], such as: information gain, chi square, and simple frequency count. An algorithm will be implemented in order to assign weights to the list of triplets extracted from the morphological descriptions, based on the ranked list of words mentioned above. After evaluating the results, a criteria will be set up, those triplets that satisfies it will remain part of the identification key, and those that doesn't will be set aside as non-significant. However, it is important to have them available for later evaluation.

D. Key Evaluation

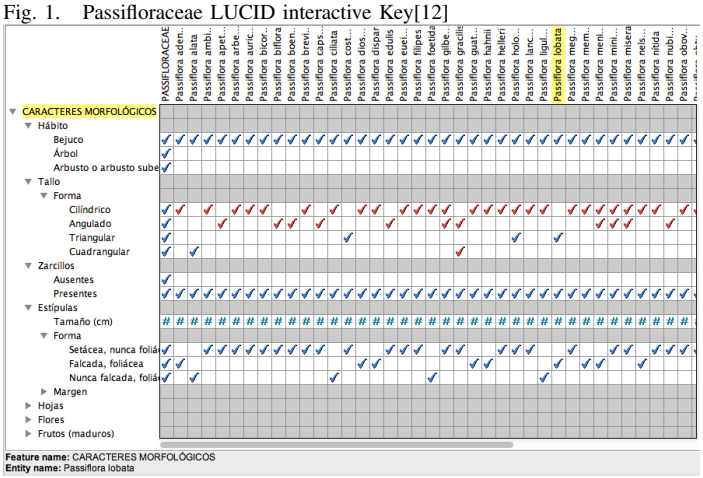
When the final list of triplets are ready, the matrix can be set up. The matrix will consist of a list of species name versus the triplets or properly said, the characters. LUCID system, provides the ability of importing CSV files structured in an specific way and convert them into an interactive key. Therefore, next step is to implement an algorithm to create a compatible matrix with species names and triplets and save it into a CSV file that can be imported into LUCID. This will allow the key be evaluated and executed, as well. The evaluation plan consists of two main tasks:

1) Share the generated interactive key with 2 or 3 taxonomists who can play with it and evaluate its performance. They will be granted a formulary with test cases, space to write down results and some other commentaries regarding their experience with the generated key.
2) Compare the automatically generated interactive key, with another one manually generated. The reason why Passifloraceae family was chosen is because researchers from the Departmento de Historia Natural del Museo Nacional de Costa Rica, recently created an interactive LUCID key of the genus Passiflora (Fig. 1). This key, apart from having information from morphological descriptions it also contains data for: distribution, ecology, etc.; but, only the morphological section will be used.

III. Results

This section presents the preliminary results obtained so far. The phase of data recollection has been completed, while the information fusion and feature selection phases are still in progress. Some data cleaning has been done, like: removing stop words, but there's still work to do, in terms of normalization and standardization of terms. The implementation of dimension reduction functions that calculates the importance of a term in a document, are ready but, the implementation of

Fig. 1. Passifloraceae LUCID interactive Key[12]



the algorithm that will select the features according to these metrics is still in the course of being done.

In the first stage of implementation the group of species chosen to work with are the ones that comprise the Passiflora genus in Costa Rica. Costa Rica has a considerable variety of passion flowers with 51 native species[11]. The morphological descriptions of this family and each of the species was retrieved from the Manual de Plantas de Costa Rica[9] and put through the semantic annotation algorithm (Listing 1).

Listing 1. Example of semantic annotated description for "frutos"

```
<statement id="T717L13" text="frutos␣
    purpura,␣3-4.5␣centimetros,␣ovoides;">
<biological_entity id="T717L13S1-230395"
    name="frutos" type="structure">
<character name="coloration" value="purpura
    " notes="Caracter␣repetido"/>
<character name="density" value="purpura"
    notes="Caracter␣repetido"/>
<character name="size_or_quantity" value="3
    -4.5" char_type="count" from="3"
    from_unit="centimetros" to="4.5"
    to_unit="centimetros"/>
<character name="shape" value="ovoides"/>
</biological_entity>
</statement>
```

Listing 2. Example of extracted triplets

```
[(frutos, coloration, purpura), (frutos,
    density, purpura), (frutos,
    size_or_quantity, 3cm-4.5cm), (frutos,
    shape, ovoides)]
```

The semantic annotation algorithm, successfully extracts all structures names alongside, with their adjectives (future state values of the key in this case) and assigns the character name it believes to match the best. Though in some cases there are more than one probable character and so it marks them as a repeated character (*"caracter repetido"*). We can see in the last example how the value *"púrpura"* was classified as *coloration* and *density*, as well; the correct one should be coloration. We need to check, if this can be easily adjusted in the algorithm, or if, we can select the correct character in an automated way using a thesaurus for example, or mark it, as a pending manual check for later.

TABLE I. TRIPLETS FOR PASSIFLORA FAMILY FLOWER PETALS

structure	character	value
petalos	density	púrpura, violeta
	prominence	intenso
	arrangement	imbricados
	fusion	separados
	quantity	ausentes
	coloration	azules, blanco verdoso, rosados, azul claro, lila rojo, escarlata, liláceo, teñidos, manchados, azulado, verdosos, blancos, blanco liláceo, purpúreo, verde amarillento, púrpura, violeta
	size_or_quantity	3, 2, 5

TABLE II. MOST FREQUENT WORDS FROM DICHOTOMOUS KEYS AND DIAGNOSTICS

dichotomous keys	diagnostics
laminas,281	laminas,74
foliares,278	foliares,67
glandulas,202	passiflora,57
peciolos,185	peciolos,51
centimetros,168	glandulas,42
bracteas,115	distingue,41
florales,108	ademas,37
lobulos,106	caracteriza,30
setaceas,95	flores,30
nunca,93	bracteas,30
foliaceas,83	frutos,27
tallos,82	rica,24
estipulas,78	costa,24
pedunculo,77	trilobuladas,24
lobuladas,72	pedunculos,22
sepalos,71	florales,21
base,67	enteras,21
ausentes,66	estipulas,20
frecuente,61	eglandulares,19
tricomas,60	foliaceas,19

This data was run through the conversion algorithm which extracts the feature triplets: (structure name, character name, value). Listing 2 shows the triplets extracted from Listing 1. There are a total of 84 triplets and 185 characters, but they still need normalization. Table I shows aggregated triplets obtained for the structure *"pétalos"*. In total, 7 characters were found in the descriptions of petals of all *Passiflora* species. The column value shows, the different values these characters can have. It can be noticed that for some reason density character is being assigned in some cases values of colors, which is not correct as we mentioned earlier. It can also be seen that colors need some normalization in order to prevent duplicates, for example, the terms [*"azules"*, *"azul"*, *"azulado"*] should be normalized as just *"azul"*. The removal of plurals and normalization of words variants are some of tasks that are still pending from implementation, but it will be resolved by applying an stemming algorithm. In addition, since the information from the Manual de Plantas de Costa Rica is on spanish, we need to translate the given character names by the semantic annotation algorithm from english to spanish. The numeric values need to be normalized as well, since they show up in different units of measurement, in range format, decimals, etc.; the unit of measurement varies when describing a seed or a stem.

Now, regarding the information of the dichotomous keys and diagnostics. We ran a simple tokenizer algorithm to extract all the words from them; removing stop words, numeric values, and accents. A total of 374 words were collected from the dichotomous keys and a total of 513 from the diagnostics. Table II, shows the first 20 words with the most frequency among both sections of information. From this group of 20 words each, there are 8 words in common, and from the whole

TABLE III. TOP RANKED WORDS

TDF	TD-IDF Weights	MI
peciolos	humedo	mitad
glandulas	bilobuladas	falcadas
laminas	flores	semillas
foliares	falcadas	profundamente
centimetros	profundamente	transversalmente
lobulos	apice	medio
florales	trilobuladas	plantas
bracteas	semillas	enves
foliaceas	metros	levemente
setaceas	oblatas	flores
basifijas	estipitadas	cordadas
tricomas	presentes	bosque
nunca	1elipticas	humedales
estipulas	mitad	inconspicuas
tallos	corona	oblongo-lanceoladas
sepalos	medio	globosos
entonces	sesiles	subigual
frutos	cortamente	foliar
base	corniculados	estigmas
lineares	foliar	prominentes

TABLE IV. TOP RANKED WORDS

IG	X2
milimetros	frecuenciatraslapados
laceradas	concoloras
medio	angostamente
flores	rugosas
pedunculo	rudimentarios
apice	protuberancias
laterales	inmaduros
hojas	triangular-alados
mas	opuberulentos
oceladas	connatasbasalmente
enteras	corto
trilobuladas	manchasblanquecinas
ausentes	distribuida
presentes	setifero-dentadas
lobulo	azulado
lobuladas	reducida
corniculados	laminasglabros
menos	largamente
metros	rojos
lineares	ferrugineo-tomentulosos

list there are 172 words in common. It can be noticed that words from diagnostics are less punctual and more broad, words like: *"usualmente"*, *"caracteriza"*, *"distingue"*; are too colloquial to find them on a key.

Next, functions for dimension reduction where applied to the list of words. We applied the functions: term document frequency, TF-IDF weights, mutual information, information gain and chi square. Tables III and IV show the top 20 ranked words from the dichotomous keys using those functions. The same process was applied to the diagnostics list of words. The most part of the top ranked words in TDF results are morphological structures, like: *"peciolos"*, *"glandulas"*, *"laminas"*, etc; while TD-IDF weights and information gain both shows structures and attribute values (*"bilobuladas"*, *"falcadas"*, *"oceladas"*, etc) as well. Mutual information seems to be less appealing since most of it's top ranked words are too general, like: *"plantas"*, *"bosque"*, *"humedales"*; and, they are not structures neither attributes values. Chi square on the other hand, most of it's words are attribute values.

IV. CONCLUSION

With the abundant available literature that content so much information about species studies, not only as a source of reference, but also, as a great source of input for creating other meaningful tools as species identification keys, guides, etc.; taking advantage of technology and accomplish it in a automated way to aid researchers, is a key issue.

The extracted annotated semantic morphological descriptions are a great source for gathering species attributes and values. Though, they need a lot of work to get them consistently and in shape for an interactive identification key. Also, there are some other characters issues that we described before on the annotated semantic data that needs to be tackled to avoid confusion on the user of the key.

The lists of words taken from both, the dichotomous keys and the diagnostics have a large intersection, which leads to assume that they certainly refer to structures and attributes thet are important for an identification of a plant species. The results obtained by applying the dimensional reduction functions to the list of words, can be used by the selection algorithm for different purposes, such as: selecting structures, selecting attributes and ranking the triplets; since, some of the them showed more top ranked words of structures and others of attributes values.

REFERENCES

[1] About Encyclopedia of Life. URL http://eol.org/about.

[2] About Biodiversity Heritage Library, 2017. URL http://biodivlib.wikispaces.com/About.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Addison Wesley, England, 2nd edition, 2011. ISBN 978-0-321-41691-9.

[4] Centre for Biological Information Technology. Lucidcentral.org, 2016. URL http://www.lucidcentral.com/.

[5] H Cui. Semantic annotation of morphological descriptions: an overal strategy. *BMC Bioinformatics*, 11:278, 2010.

[6] Hong Cui, Alex Dusenbery, James Macklin, Fengqiong Huang, Robert (Bob) Morris, and Heather Cole. Semantic Annotation, Ontology Building, and Interactive Key Generation from Morphological Descriptions. In *TDWG 2012 Annual Conference*, 2012. URL http://www.tdwg.org/fileadmin/2012conference/slides/TDWG2012-Cui.pdf.

[7] M J Dallwitz. Overview of the DELTA System, 2010. URL http://delta-intkey.com/www/overview.htm.

[8] M J Dallwitz, T A Paine, and E J Zurcher. Principles of Interactive Keys. pages 1–20, 2013.

[9] B.E. Hammel, M.H. Grayum, C. Herrera, and N. Zamora, editors. *Manual de Plantas de Costa Rica*. Missouri Botanical Garden Press, St. Louis, Missouri, 2003. ISBN 1-930723-22-9.

[10] María Auxiliadora Mora. Extracción semiautomática de atributos morfológicos de especies a partir de descripciones taxonómicas, 2016.

[11] Alexander Rodríguez G. and Armando Estrada Ch. *Flores de Pasión de Costa Rica*. INBio, Costa Rica, 1 edition, 2009.

[12] Alexander Rodríguez G., Armando Estrada Ch., Marianela Mata R., and Tatiana Gutiérrez R. Passifloras de Costa Rica, 2015.

[13] Qin Wei. *INFORMATION FUSION IN TAXONOMIC DESCRIPTIONS*. Doctoral, University of Illinois at Urbana-Champaign, 2011. URL https://www.ideals.illinois.edu/bitstream/handle/2142/26070/Wei{_}Qin.pdf.