Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 163-169.
ISBN 978-88-8303-295-0. EUT, 2010.

eFlora and DialGraph, tools for enhancing identification processes in plants

Fernando Sánchez Laulhé, Cecilio Cano Calonge,
Antonio Jiménez Montaño

Abstract —This document describes in some detail the eFlora web application, a powerful tool for the identification of plant species. It incorporates the corpus of Flora Iberica, a scientific description of the vascular plants living in the Iberian Peninsula, which is treated as unstructured information and therefore indexed by a full text search engine tool, in our case Lucene. eFlora also includes dichotomous keys, which are displayed using Hyperbolic Geometry. By making intelligent use of the keys, we have created two original and useful features, the comparison of arbitrarily chosen species, which is resolved by a dynamic generation of subkeys applied to these selected species, and the presentation of dichotomous keys in the form of a Virtual Assistant, or conversational robot, using our solution DialGraph, which allows to non-academic users an approach in Natural Language, such as chat, or voice recognition, Text to Speech Synthesis (TTS) or even Automatic Translation when dealing with a multilanguage context. Concerning the configuration of the Virtual Assistant, we provide a very intuitive BPM-like graphical design. This approach to dichotomous keys helps teaching biodiversity science, enhances the awareness of its importance, and makes citizens emotionally closer to science.

Index Terms — virtual assistant, hyperbolic geometry, natural language, dichotomous key, Bot, automatic translation, Flora Iberica, BPM, Text to Speech Synthesis, TTS.

1 Introduction

Nowadays there is a major challenge to properly process the large body of available botanical information, most of which lacks structure consensus, probably due to both its ancient origins and the inherent difficulty to structure it in a consistent manner. This document describes in some detail the eFlora web application, a powerful tool for the identification of plant species, which incorporates the corpus of Flora Iberica, a scientific description of the vascular plants living in the Iberian Peninsula. eFlora has been addressed

The authors are with Terasoft Consultores S.L., Parque Empresarial Punta Galea c/.Perú, 4. Loft 3 28230 Las Rozas (Madrid), Spain. E-mail of first author: fsl@terasoft.es.

with a full text type approach, and is accessible via an appropriate open source search engine, Lucene, which allows to launch free text queries, the system responding with a list of species, ranked in order of relevance. Being Flora Iberica a work aimed at the identification of species, we have also included dichotomous keys in graphic form and under hyperbolic geometry. Furthermore, the original functionality of these keys has been enhanced in the form of dynamic generation of sub-keys from any arbitrary selection of species. Another intesting feature, implemented with a dichotomous key to Iberian Conifers, is the presentation of the key in the form of a conversational robot or virtual assistant, that interacts with the user in natural language.

2 Flora Iberica

2.1 Descriptions

To give some structure to the information contained in Flora Iberica, while maintaining the characteristic of free natural language descriptions, the description part of each species has been separated into a priori-defined fields (Tab. 1).

Level 1 fields	*Level 0 fields*
Plant	Root
	Basal stem
	Stems and branches
	Spines
Leaves	Leaves
Flower	Inflorescence
	Receptacle
	Calyx
	Corolla
Gynaeceum	Gynaeceum
Androecium	Androecium
	Nectary
Fruit	Fruit
	Seed

Tab. 1 – Fields used for organising the descriptions of species in Flora Iberica.

These fields can be grouped into even greater granularity, in anticipation that some of them may belong to categories of higher order. For example, a query can be designed as generically directed to the field *Flower*, or it may be more detailed, specifying which field in the category *Flower* the query is to be launched: *Inflorescence*, *Receptacle*, *Calyx* and *Corolla*. In this way, queries can be made either under both full free text scheme, or directed, in free text as well, towards specific fields.

2.2 Dichotomous keys by hyperbolic geometry

The dichotomous keys of Flora Iberica have been digitized. These keys are specific for all taxonomic levels used in the original work i.e. there are special keys that identify species within the same genus, keys to identify genera within a family, etc.

The dichotomous keys have been ported to XML format for processing in the form of trees in hyperbolic geometry, whose root concept is that it allows to pass through a point more than one line parallel to another line that does not pass through that point. With respect to information representation - decision trees in our case – this means that the hyperbolic space can be represented as a circle in which the periphery represents infinity. In this geometry, the closer to the edge (infinity) we are, the smaller is the size of what we represent. This allows to represent graphs in hyperbolic geometry maintaining the properties of focus and context. Everything in the center is large, and if we move aside to the periphery, it becomes smaller, but still visible. For the implementation of hyperbolic trees, Treebolic, an open source solution, has been used (Fig. 1).

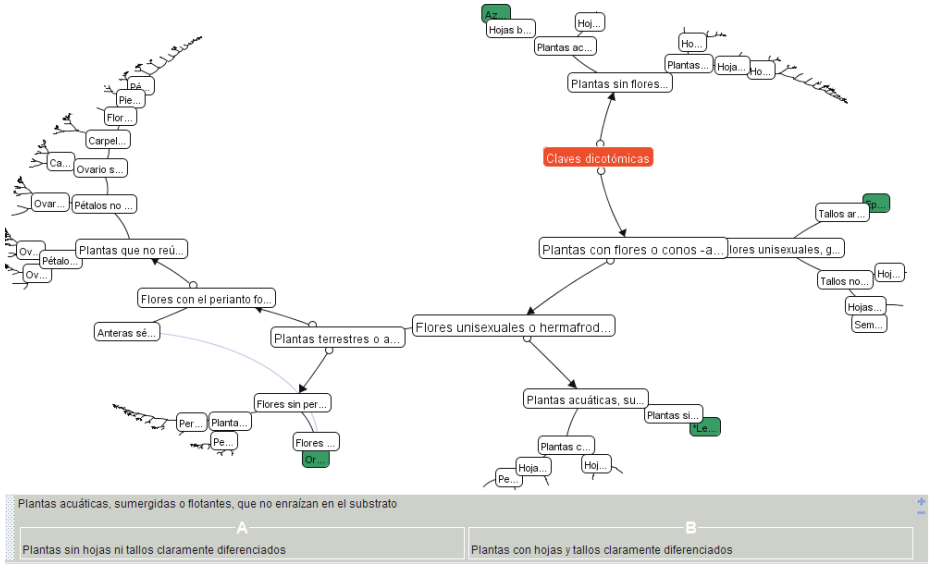

Fig. 1 – General view of dichotomous keys at initial execution from the root node, displayed using hyperbolic geometry.

3 Indexing

Lucene, an open source solution from Apache Group, was used as the main search engine. We have created a data model that includes a set of indexed fields specifically for using in queries. Other fields, such as Flowering periods, Observations or Bibliography are used only as supplementary information to be shown when displaying a full description.

As for the ability to query in English, we have added a filter that acts between

the user input string and the search engine. In this way, a user not speaking Spanish, but able to understand botanical information, can query directly in English.

4 Identification

4.1 By species comparison

While eFlora indexes the data with a full text search engine, queries have a qualitative difference in relation to those made in a classic form, such as in Google. In the latter case, some information is sought, and the result can be found in one or more points from the list of results provided by Google. It is the surfer who decides what information is to be considered as valid. This means that sometimes useful information may be scattered among a high number of returned results. In eFlora the situation is completely different. The question is: what is the species that we have in our hands? For anwering, it is not sufficient for the search system to just offer a series of results ranked by relevance. This is useful, but what is necessary is that the system can tell us what to observe for differentiating between the items included in the list of results. Otherwise the user should make a very tedious work examining the full description of every species in the list, trying to find subtle differences among similar taxa.

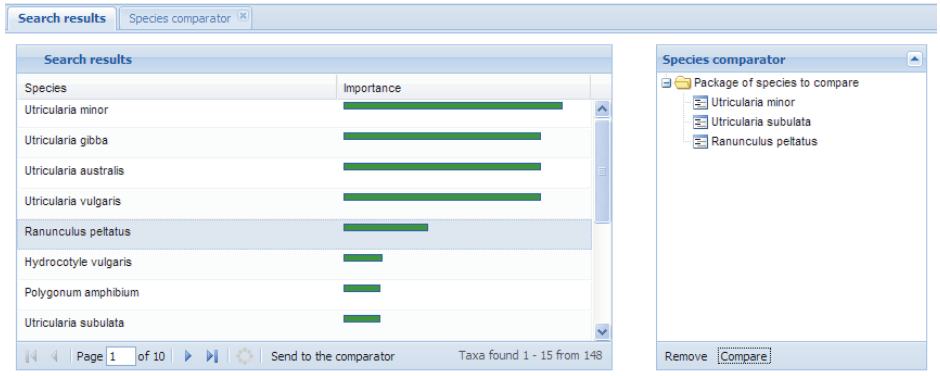

Fig. 2 – Example with a list of results ranked by relevance; on the right side, the repository filled up with three species to be compared.

With eFlora the identification task, which is originated through a query using terms describing certain observable characteristics of a certain species, begins with a set of species that have a good chance of containing the targeted specimen. To differentiate among taxa, users can select those which look more likely, and bring them to a special repository in which they can perform the function of comparison (Fig. 2). This function processes the entire tree of the dichotomous key in real time and extracts a subkey that shows the differences among the selected species. In this way, the user has a useful tool to identify the species in question (Fig. 3).

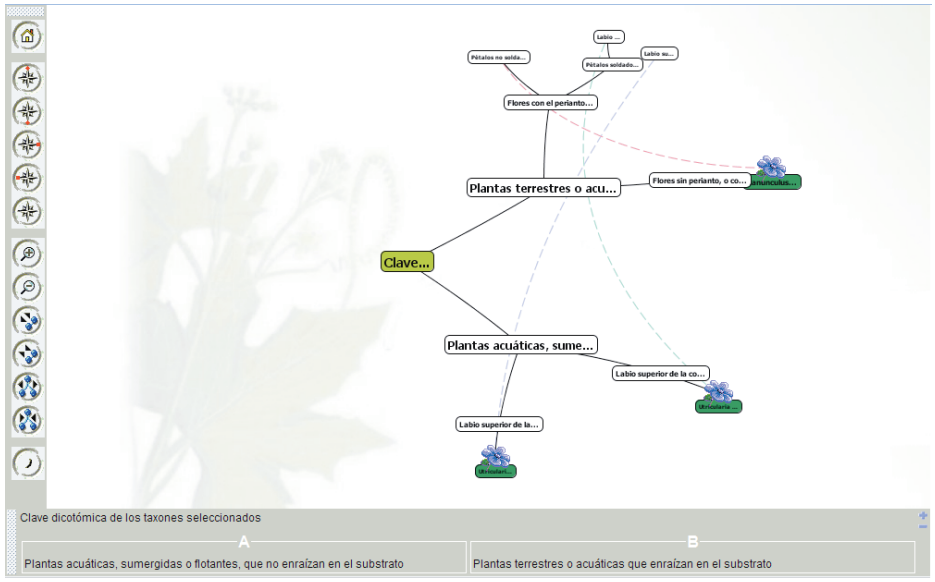

Fig. 3 – Once the compare function has been executed, a dynamic subkey is displayed in hyperbolic geometry.

4.2 The Virtual Assistant

We have developed an original tool for teaching the identification of biodiversity, based on the use of Virtual Assistants in the form of a conversational bot. A Virtual Assistant is a software application capable of maintaining a conversation with users on a specific topic in natural language. We applied DialGraph to a dichotomous key of Iberian Conifers. DialGraph allows an almost complete portability between dichotomous keys - considered as a large tree structure - and its graphical system design. It allows communication with users through Chat, materialized in a virtual character who pronounces sentences and might even recognize the voice of the Internet surfer. This system is topologically equivalent to the traditional dichotomous keys, but for its ability to chain multiple words appearing in the same sentence, and the freedom to use popular expressions and to apply automatic translation, it has important advantages for teaching Biodiversity, mainly when focused towards young users (Fig. 4).

The strongest point of DialGraph is the tool created for designing the conversation logic, which we call The Designer**.** This can be accessed through the web using standard navigators, such us IE, FF, Chrome, etc., and models the dialog by drag and drop objects within a virtual dashboard. These objects represent different assistant states and the links among states. By defining in links as parameters very simple linguistic lexical roots and appropriate messages in states, the dialog with users remains fully determined as long as the user is informed about the aim of the assistant (Fig. 5).

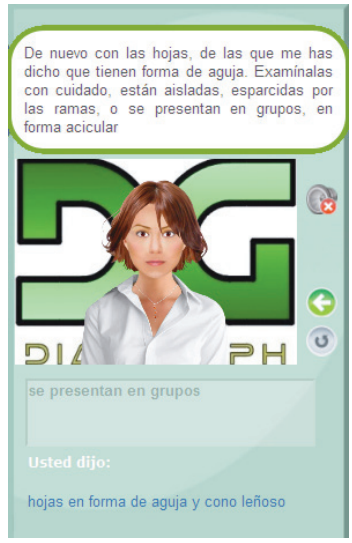

Fig. 4 – Moving avatar interacting with users in the web.

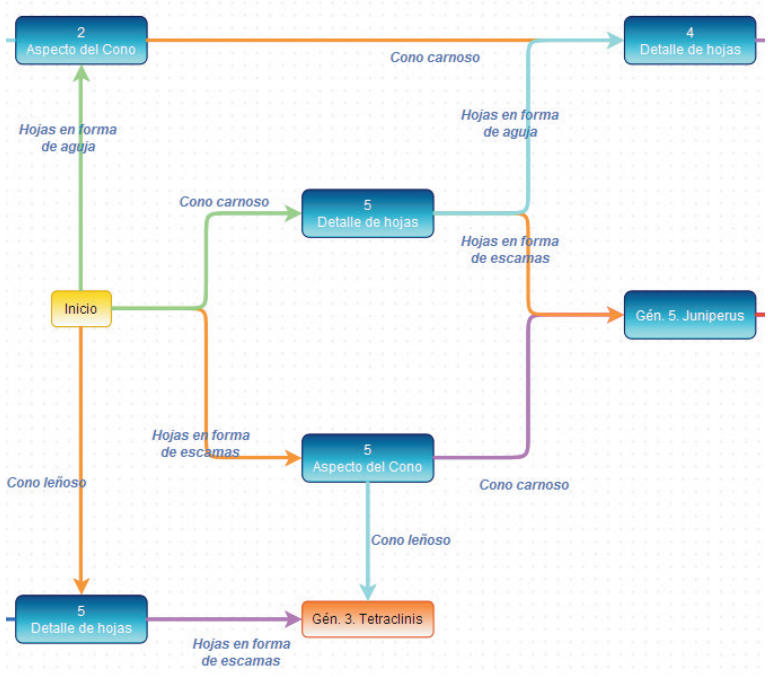

Fig. 5 – Designing the flow of dialog for a specific part of the dichotomous key using DialGraph as a BPM tool.

5 Conclusion

TeraSoft Consultores SL has created, together with the Botanic Garden of Madrid, a set of tools which enhances the identification of plant species. Based on the digitalization of botanic data “as they are” in the classic form, including dichotomous keys, it permits to approach science to citisens and to effectively teach the importance of Biodiversity.

References

[1] E-Flora Iberica: http://www.efloraiberica.es/eflora/, 2010.
[2] DialGraph: http://www.dialgraph.com, 2010.
[3] TeraSoft Consultores SL: http://www.terasoft.es, 2010.