Using preference-enriched faceted search for species identification

Yannis Tzitzikas*

Institute of Computer Science,
FORTH,
Crete, Greece
and
Computer Science Department,
University of Crete,
Crete, Greece
Email: tzitzik@ics.forth.gr
*Corresponding author

Nicolas Bailly

Hellenic Centre for Marine Research,
Crete, Greece
Email: nbailly@hcmr.gr

Panagiotis Papadakos and
Nikos Minadakis

Institute of Computer Science,
FORTH,
Crete, Greece
Email: papadako@ics.forth.gr
Email: minadakn@ics.forth.gr

George Nikitakis

Computer Science Department,
University of Crete,
Crete, Greece
Email: nikitak@csd.uoc.gr

Abstract: Species identification is essentially a decision-making process comprising steps in which the user makes a selection of characters, figures or photographs, or provides an input, that restricts other choices, until reaching one species. In some identification methods such decisions should have a specific order. Consequently, a wrong decision at the beginning of the process, could exclude a big set of options. To make this process more flexible and less vulnerable to wrong decisions, in this paper we investigate how a Preference-enriched Faceted Search (PFS) process can be used to aid the identification of species. We show how the proposed process covers and advances the existing methods and we report our experience from applying this process over data taken from FishBase. In the sequent, we elaborate on evaluation and we report the results of a task-based evaluation that shows that the PFS-based method can be used effectively by casual users.

Keywords: species identification; identification tasks; faceted search; preferences.

Reference to this paper should be made as follows: Tzitzikas, Y., Bailly, N., Papadakos, P., Minadakis, N. and Nikitakis, G. (2016) 'Using preference-enriched faceted search for species identification', *Int. J. Metadata, Semantics and Ontologies*, Vol. 11, No. 3, pp.165–179.

Biographical notes: Yannis Tzitzikas is currently Assistant Professor of Information Systems in the Computer Science Department at University of Crete (Greece) and Associate Researcher of the Information Systems Laboratory at FORTH-ICS (Greece). Before joining University of Crete and FORTH-ICS he was postdoctoral fellow at the University of Namur (Belgium) and ERCIM postdoctoral fellow at ISTI-CNR (Pisa, Italy) and at VTT Technical Research Centre of Finland. His current research focuses on exploratory searching (principles, techniques, applications) and

semantic data management (comparison functions, knowledge evolution, indexes, visualisation, integration). The results of his research have been published in more than 100 papers in refereed international conferences and journals.

Nicolas Bailly is a Biodiversity Informatics Scientist and Ichthyologist in the Institute of Marine Biology Biotechnology and Aquaculture of the Hellenic Centre for Marine Research (HCMR). Employed at WorldFish (2005–2014), he was the FishBase program manager as the Officer in Charge of the Philippines Office (2005–2010), and then seconded to the FishBase Information and Research Group (FIN), a Philippine NGO, as the Scientific Director (still acting at that position). He is Scientific Adviser for SeaLifeBase. Under WorldFish and FIN, he participated to the EC-FP7 projects D4Science I and II, BioFresh, iMarine and EU-BON, besides other smaller projects from various donors. He is member of the Global Team of the Catalogue of Life since 2008, Taxonomic Group Chair since 2009, and the vice-Chair of the Global Team since 2011. Previous member of the editorial committee of Cybium, he is member of the editorial committee of Acta Ichtyologica and Piscatoria since September 2014.

Panagiotis Papadakos is a Postdoctoral Researcher of the Information Systems Laboratory at FORTH-ICS (Greece) He is a Postdoctoral Scientist of the Information Systems Laboratory at FORTH-ICS (Greece). He got his PhD in preference-based interactive exploration of multidimensional information spaces from the Computer Science Department of the University of Crete, Greece, in 2013. His main research interests are in the areas of information retrieval, databases and the semantic web, with a focus on exploratory search and preference-based interaction over faceted and dynamic taxonomies. In the past he has worked extensively in the areas of virtual reality and 3D user interfaces. He has published more than 15 papers in peer-reviewed conferences and journals, and he has been involved in international and national projects.

Nikos Minadakis is a Research and Development Engineer in the Information Systems Laboratory (ISL) at the Foundation for Research and Technology Hellas (FORTH). He graduated from the Department of Computers Engineering and Informatics, University of Patras, and holds an MSc in Computer Science and Technology where he was mainly occupied with Internet Technologies and Databases. He has been involved in a number of projects working on automation of semantic warehouses construction process, biodiversity data modelling, data integration and reasoning.

George Nikitakis is an MSc student in the Department of Computer Science, University of Crete. He has graduated from the same department.

This paper is a revised and expanded version of a paper entitled 'Species identification through preference enriched faceted search' presented at the '9th Metadata and Semantics Research Conference (MTSR'15)', Manchester, UK.

1 Introduction

Correct identification of fish species is important in the fisheries domain for a sustainable management of stocks, balancing exploitation and conservation of biodiversity (Fischer, 2013). Species identification is actually a decision-making process comprising steps in which the user makes a selection that restricts subsequent choices. The decisions are actually selections of characters, figures or photographs. The European Project Key2Nature[1] (2007–2010) reviewed a number of identification systems used in biodiversity works and education, and published proceedings of a conference on the state of the art in collaboration with two other projects, EDIT and STERNA (Nimis and Vignes-Lebbe, 2010). Up to the development of informatics, the steps in the classic textual dichotomous identification keys were constrained by a fixed order. Consequently, if a wrong decision is made at the beginning of the identification process, the final identification is wrong.

In this paper we investigate how a particular exploratory search process, specifically the *Preference-enriched Faceted Search* (for short PFS) introduced in Tzitzikas and Papadakos (2013) and implemented by the system *Hippalus* (Papadakos and Tzitzikas, 2014), can be used for supporting the identification process in a way that is more flexible and less prone to errors. We show how the proposed process, which is based on methodologies and tools developed for *exploratory search*, covers and advances the existing methods for species identification (including the current computer-aided identification systems). The main idea is that species are modelled as objects characterised by a number of attributes. The user explores the information space by issuing two main kinds of actions: (a) the classical left-clicks of faceted search that *change* the focus (i.e. change the viewed set of objects), and (b) right-clicks that express *preference* which *rank* the focus. The order by which users issue such actions does not affect the outcome, enabling in this way the user to decide what is more convenient to him/her at each step during the interaction. We demonstrate how we have applied this approach over the data of FishBase.[2] However, the approach is generic, i.e. it can be applied over any kind of objects described by a

number of attributes (or metadata), and it could be exploited for identifying a phenomenon in general; identification tasks are important in many domains, e.g. in patent search (for investigating whether an idea is already covered by existing patents) (Fafalios et al., 2013), for identifying a rare disease (Sacco, 2008), for diagnostics of car breakdowns, and others.

The rest of this paper is organised as follows. Section 2 describes related work and background, Section 3 introduces the PFS-based approach for species identification and finally Section 4 concludes the paper.

2 Related work and background

Section 2.1 describes in brief the current species identification methods, Section 2.2 introduces the basics of PFS, and Section 2.3 introduces the system Hippalus.

2.1 Species identification

There are various methods for identifying one species. According to Bailly et al. (2010) and Fischer (2013) the following four methods[3] can be identified (see Pankhurst [1991] for a more detailed overview):

1 'Eyeballing' drawings and key features by decreasing taxonomic level from class downward;

2 Display of all pictures available for a given geographic area and/or a given family with possible restriction on fin ray meristics;

3 Dichotomous keys: these keys can be classically implemented as in printed textual documents; however, computer-aided identification systems such as XPER,[4] LucId[5] and others allow users to select steps in the order they prefer, which is a first step for the process we describe below; and

4 Polythetic keys, such as simple morphometric ratios measured on the body of individuals (e.g. fishes), or biochemistry results as in bacteria identification.

In the identification through outlines (corresponding to category 1 above) the user restricts the search space gradually by selecting drawings in each step. Figure 1 shows the first step, where the user can select one out of six highly distinctive outlines representing different classes of fishes. After his first selection, he continues in the same manner, by selecting another figure from a set of figures with more refined groupings (decreasing taxonomic level), and so on.

In the identification through *pictures* the user selects a taxonomic group (from class down to genus), and/or a geographic area, and eyeballs the corresponding displayed pictures on one web page. Other criteria like the number of dorsal and anal spines can be used to restrict choices. In the identification through *dichotomous keys* (corresponding to category 3 above), the user answers successive questions like in a decision-tree, usually about the morphology of the species. The principle, established at the end of the 17th century (Griffin, 2011), but popularised only 100 years later, is the following: (a) the user has to answer a first question that has two possible answers (hence the qualification of dichotomous), and (b) each of the possible two answers leads via a number either to a new question, or to a species, which then finishes the identification. An example of a dichotomous key implemented in simple HTML in FishBase is given in http://www.fishbase.org/keys/description.php?keycode=2 (Renaud, 2011).

Finally, in the identification through *morphometric ratios* (e.g. the Morphometrics Tool of FishBase), ratios are computed from body measurements provided by the user, e.g. Total Length (TL), Head Length (HL), Eye Diameter (ED), Body Depth (BD), etc., and other information about the area and family for reducing the number or possible species. Figure 2 shows the form for identification through morphometric ratios.

2.2 The preference-enriched faceted search (PFS)

Faceted Search: A highly prevalent model for *exploratory search* is the interaction of *Faceted and Dynamic Taxonomies* (FDT) (Sacco and Tzitzikas, 2009), usually called *Faceted Search*, which allows the user to get an *overview* of the information space and offers him/her various groupings of the objects of the information base based on their attributes, metadata, or other dynamically mined information. These groupings enable the user to restrict the focus *gradually* and in a simple way (through clicks, i.e. without having to formulate queries), enabling him to locate resources that would be difficult to locate otherwise. This model is currently the de facto standard in various domains: e-commerce (e.g. eBay), booking applications (e.g. booking.com), library and bibliographic portals (e.g. ACM Digital Library), museum portals (e.g. Europeana), mobile phone browsers, and many others. Recently faceted search is also used for exploring RDF data sets (see Tzitzikas et al. [2016] for a recent survey).

Preferences: Commonly, preferences are not hard constraints, but wishes, simple or complicated ones (covering one or more aspects), which might or might not be satisfied. Such wishes might be independent, or might affect each other, even in conflicting ways. A survey of major questions and approaches for preference handling in applications such as recommender systems, personal assistant agents and personalised user interfaces is given at Peintner et al. (2009), while Pu and Chen (2009) propose guidelines and reports examples for product search and recommender systems. In general, preferences can be defined either using a *qualitative* approach (Kießling, 2002; Chomicki et al., 2003; Georgiadis et al., 2008) or a *quantitative* approach (Agrawal and Wimmers, 2000; Balke and Güntzer, 2004; Koutrika and Ioannidis, 2005).

According to the former, preferences are described directly, using a preference relation $>_{Pref}$ (comprising pairs of the form $x >_{Pref} y$), while according to the latter, preferences are described indirectly by defining scoring functions (i.e. $Score(x) > Score(y)$). The qualitative approach is more powerful and expressive than the quantitative approach, since not every preference can be modelled using scoring functions (for more, see Chomicki et al., 2003; Fishburn, 1970). There are also approaches that support a mixture of qualitative and quantitative preferences (Rossi et al., 2009). For a recent survey about preferences and databases systems, see Stefanidis et al. (2011).

Figure 1 Example of the 'Eyeballing' drawings by decreasing taxonomic level from class downwards

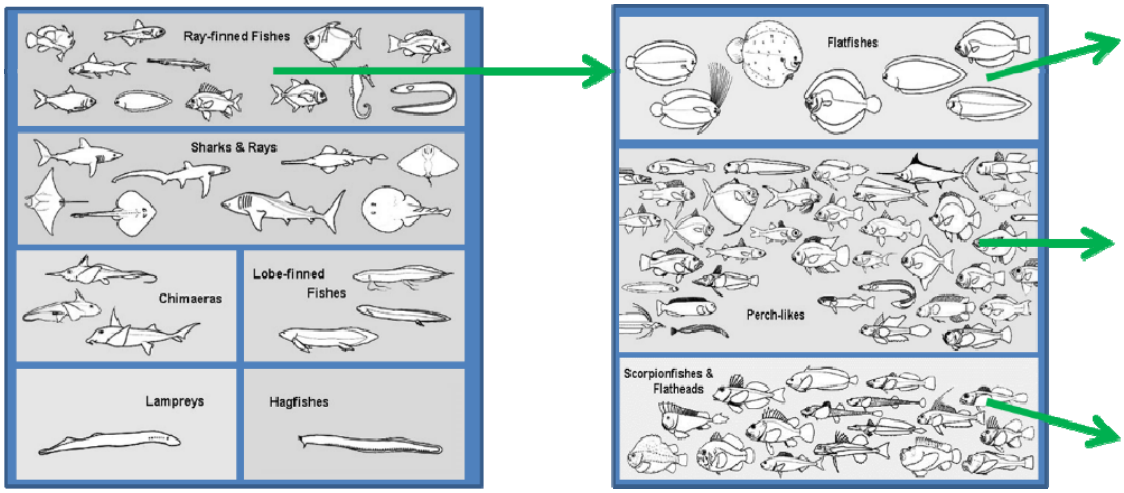
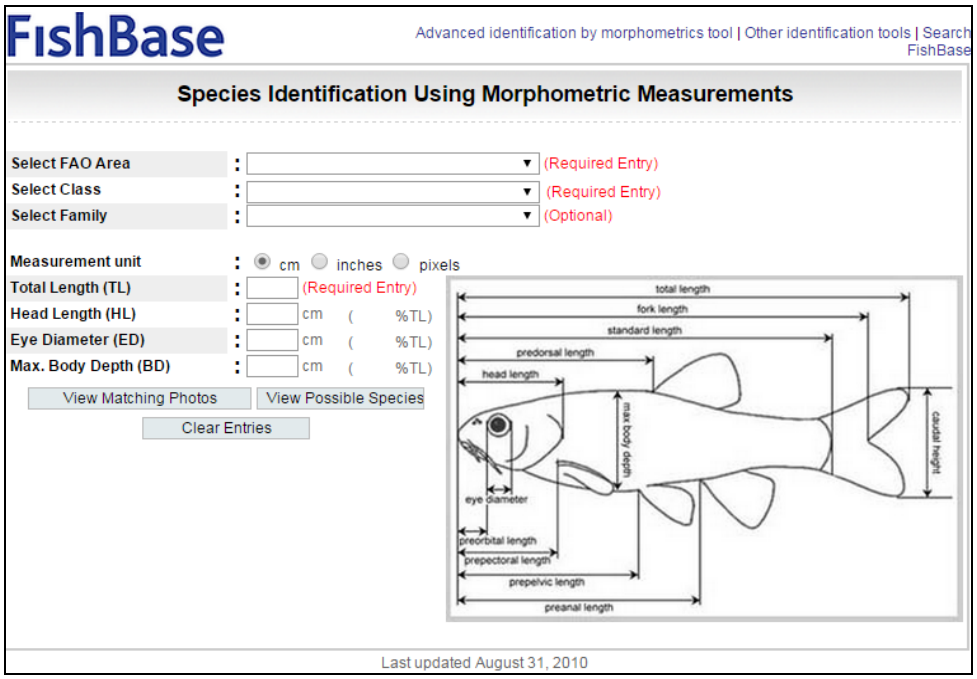


Figure 2 Form for identification through morphometric ratios in FishBase (http://www.fishbase.org/Identification/Morphometrics/centimeters/Index.php)



PFS: Most FDT systems order facets and zoom-points in lexicographical order, or according to the number of indexed objects. Other systems, like eBay, present only a manually chosen subset of facets to the users, and the zoom-points are again ranked based on the number of indexed objects. There are several works that propose methods for identifying and ranking facets and zoom-points according to various criteria (e.g. Dakka et al., 2005; Dash et al., 2008; Koren et al., 2008; Wagner et al., 2011). However, and to the best of our knowledge, the only model that allows users to define explicitly the desired *preference structure* in a gradual and flexible manner, i.e. aligned with the principles of faceted search), also anticipating attributes with hierarchically organised values and possibly set-valued, is the one proposed in Tzitzikas and Papadakos (2013). In brief, in that work the interaction model of faceted search is enriched with actions that allow the user to express *preferences* for ranking the facets, their values, and the objects. The preference actions allow expressing *best values*, *worst values*, *relative preferences* (e.g. I prefer A to B), and *around to* preference. Moreover, the framework offers actions for *composing* preferences using *Priority*, *Pareto* (Kießling, 2002), *Pareto Optimal* (i.e. skyline; Chomicki et al., 2003), and others. We shall hereafter refer to this interaction PFS with the term *Preference-enriched Faceted Search*. One distinctive feature of PFS is that it allows preferences over attributes whose values are hierarchically organised (and/or multi-valued) and preference inheritance is supported. For this reason PFS is equipped with scope-based rules that resolve automatically the conflicts that may arise. Overall, with PFS a user is able to restrict the focus by using the faceted interaction scheme (hard restrictions) that lead to non-empty results, and also rank it according to the preferences that he has expressed.

PFS has been proven useful for recall-oriented information needs, because such needs involve decision-making that can benefit from the gradual interaction and expression of preferences (Papadakos and Tzitzikas, 2015).

A relevant work to PFS approach is discussed in Qarabaqi and Riedewald (2014). That work ranks the results based on a probabilistic framework that does not consider explicit users' preferences and assumes a data model that, on contrast to PFS, does not exploit hierarchically organised and/or set-valued attributes.

2.3 The Hippalus system

Hippalus (Papadakos and Tzitzikas, 2014) is a publicly accessible web system that implements the PFS interaction model that was just described. Consequently, it offers actions that allow the user to order facets, values, and objects using *best*, *worst*, *prefer to* actions (i.e. relative preferences), *around to* actions (over a specific value), or actions that order them lexicographically, or based on their values or count values. Furthermore, the user is able to *compose* object-related preference actions. The information base that feeds Hippalus is represented in RDF/S[6] (using a schema adequate for representing objects described according to dimensions with hierarchically organised values). In more details, each object (fish species in our case) indexed by the Hippalus system is an instance of the hippalus[7]: Hippalus_Id class. The attributes of each species (i.e. the displayed facets) are properties of the Hippalus namespace, with instances of the Hippalus_Id class as their domain and literals (i.e. strings, float, int, etc.) as their range. The schema also supports hierarchically organised values, by exploiting *rdfs:subClass* relations. For example, regarding the attribute *Country*, a species can be an instance of the class *Greece*, which is a subClass of *Europe*, etc. Notice that we have implemented a tool that is able to automatically translate simple CSV files, RDF files, the results of a SPARQL query, etc., to the corresponding schema used by the Hippalus system.

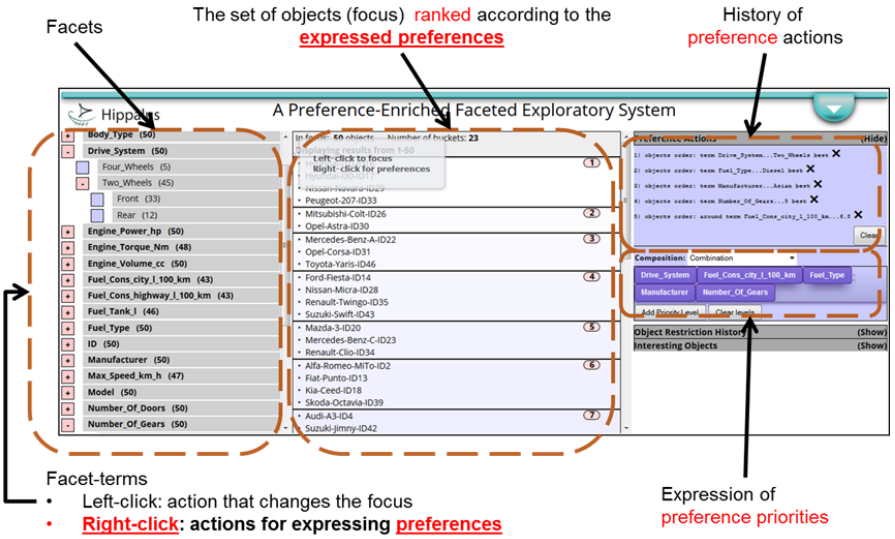
As an example below we can see how we can represent information about one particular species, namely 'Alburnus Macedonicus'. In this example we use the Turtle format (one of the several serialisation formats of RDF). The first row defines the URI for this species, the second states that the type of this URI is Hippalus_Id, and each of the rest lines describes one attribute of this species. More information about the input format of Hippalus is given in the website of Hippalus. Notice that in this example the *Country* facet is not hierarchically organised and is represented by a property.

Hippalus[8]:Alburnus_macedonicus
a hippalus:Hippalus_Id ;
hippalus:BodyShape "elongated"^^xsd:string ;
hippalus:Country "Greece"^^xsd:string ;
hippalus:Family "Cyprinidae"^^xsd:string ;
hippalus:Genus "Alburnus"^^xsd:string ;

For loading and querying such information, Hippalus uses Jena,[9] a Java framework for building Semantic Web applications. Hippalus offers a web interface for Faceted Search enriched with preference actions. The latter are offered through HTML5 context menus.[10] The performed actions are internally translated to statements of the preference language described in Tzitzikas and Papadakos (2013), and are then sent to the server through HTTP requests. The server analyses them, using the language's parser, and checks their validity. If valid, they are passed to the appropriate preference algorithm. Finally, the respective preference bucket order is computed and the ranked list of objects, according to preference, is sent to the user's browser.

Hippalus displays the preference ranked list of objects in the central part of the screen, while the right part is occupied by information that relates to the information thinning (object restrictions), preference actions history and preference composition. Figure 3 shows a screenshot of Hippalus over a data set that contains car descriptions. The preference-related actions are offered through right-click activated pop-up menus (through HTML5 context menus). The interaction is demonstrated in the next section.

Figure 3 The first screen of Hippalus



3 A PFS-based approach for species identification

Section 3.1 describes how preference-based ranking is related to species identification, and Section 3.2 describes the PFS-based fish identification process through an example. Section 3.3 provides details about the final data set, while Section 3.4 compares the various species identification approaches. Section 3.5 reports the results of an evaluation with users. Finally, Section 3.6 describes other identification tasks (apart from species identification) that could benefit from PFS.

3.1 Preferences and species identification

Let $S = \{s1, \ldots, sN\}$ be the set of N options, fish species in our case, each having a name (scientific name in our case). We have K facets $F = \{F1, \ldots, FK\}$ each associated with a taxonomy $(Ti, \leq i)$ where Ti is a set of terms, or values, while $\leq i$ is a (possibly empty) partial order over Ti enabling to organise the values of T hierarchically. Each fish species si is described by associating to it one or more values from each facet. Let **si** denote the description of si in that space, and $\boldsymbol{S}$ be the set of descriptions of all species in S. A user u explores the information space and expresses gradually a set of actions: restrictions or preferences. As regards the latter, the user expresses qualitative (i.e. relative) preferences over the terms of each facet. Let *actions(u)* denote the set of *preference actions* expressed by u. These actions define a preference relation (a binary relation) over each Ti, denoted by $\succ(i, u)$, and then these binary relations are composed to define a preference relation over the elements of the information space, i.e. over $V = T1 \times \ldots \times TK$. Note that since a species can be associated with more than one values from a facet, it is more precise to define V as the Cartesian product of the power sets of all Ti. Since the descriptions of the species $\boldsymbol{S}$ is a subset of V, the actions in *actions(u)* define a preference relation over S denoted as $(S, \succ u)$. From $(S, \succ u)$ a *bucket order* of S, i.e. a linear order $<b1, \ldots, bZ>$ of subsets of S is produced though topological sorting where $b1$ contains the most preferred species, while bZ the least preferred. All bi $(1 \leq i \leq Z)$ form a *partition* of S (i.e. they are pairwise disjoint and their union is S). The number of blocks Z ranges between 1 and N. Obviously, if $Z = N$ then the ranking forms a linear order of S, while if $Z = 1$ then all parties are equally ranked (this is true at the beginning of the interaction).

3.2 The interaction by example

We shall describe the PFS-based fish identification process through an example. We use the data set of the pilot phase containing only 720 species (mainly coming from Greece) where each species is described by six attributes: *Body Shape*, *Country*, *Genus*, *Length*, *Maximum Depth*, and *Weight*. We used the Hippalus Data Translator (HDT) tool for transforming the original FishBase data to the corresponding RDFS multidimensional schema that is supported by Hippalus. This resulted to 3254 RDF triples. The system with those 720 species is web accessible.[11]

Scenario: Suppose that someone catches a fish while fishing in a boat close to a coach of Cephalonia (a Greek island in the Ionian Sea). The caught fish is a rather flattened fish of around 3 kg and its length is between 60 and 65 cm. The fish looks like the one in Figure 4. Below we discuss how that fisherman could use the system Hippalus for the identification of that particular fish.

Figure 4 The caught fish

The first screen of Hippalus is shown in Figure 5. The left frame shows the facets and the middle frame shows the list of species names that the system is aware of. The right frames show the history of user in terms of focus restriction actions, preference expression actions, and also allow the user to prioritise his/her preferences.

Since the caught fish is a rather flattened fish, the user expands the facet Body Shape and on the value flattened he selects through right-click the value ***Best***. We can see (in Figure 6) that now the list of species has been divided into two blocks: the first contains the flattened fishes, while the second those having a different shape.

Since the weight of the fish is around 3 kg the user expands the facet Weight and on the value 3000 g he selects through right-click the value ***Around***. We can see (in Figure 7) that now the two blocks have been refined and a series or smaller blocks are shown. The first block contains a single species, namely *Torpedo marmorata*, meaning that this is the more probable one.

The user now expands the facet Country and since Greece is close to Italy, and therefore it could be a species native in either Greece or Italy, or both, he selects Spain and through right-click he selects the value ***Worst***, for expressing that Spain is the worst option, i.e. the less probable. We observe (in Figure 8) that the first block did not change.

At that point the user questions himself whether the fish is indeed flattened or compressed, because he is not sure about the semantics of this terminology. For this reason he goes again to the facet Body Shape he finds the value compressed and through right-click he selects the value ***Best***. Now the first block contains two species (Figure 9): *Torpedo marmorata* and *Solea solea* (the former due to its past ***Best*** action on flattened, the latter due to the current Best action on compressed).

The user now expands the facet Length and selects the value ***Around*** over the value 63. Now the first block contains three species, namely *S. solea*, *Scomber colias*, and *Carassius carassius* (Figure 10).

Since the user is more certain about the body shape of the fish, in comparison to weight and length, he decides to express, through the *Priorities* frame (middle right of the screen), that his preference on Body Shape should have more priority than his preferences over the other facets. Figure 11 shows the new result. The first block contains one species, *S. colias*. This is probably the right species. To confirm, the user should check the species account in FishBase,[12] and compare the various characteristics he can observe on the individual with those reported in FishBase.

Figure 5 The first screen of Hippalus

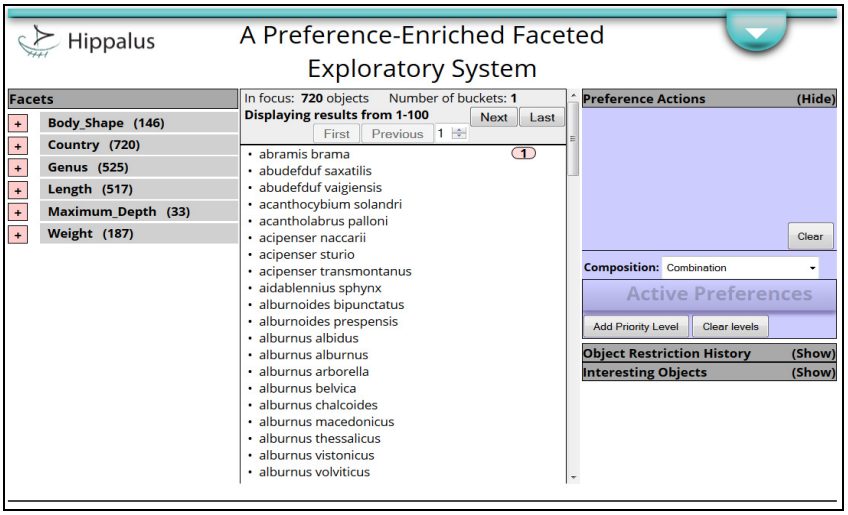


Figure 6 After the preference action *Best(BodyShape:Flattened)*

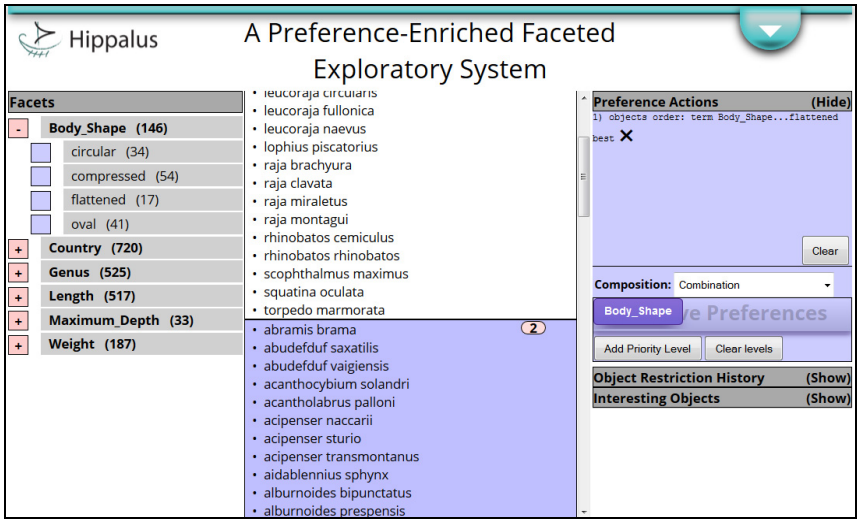


Figure 7 After the preference action *Around(Weight:3000)*

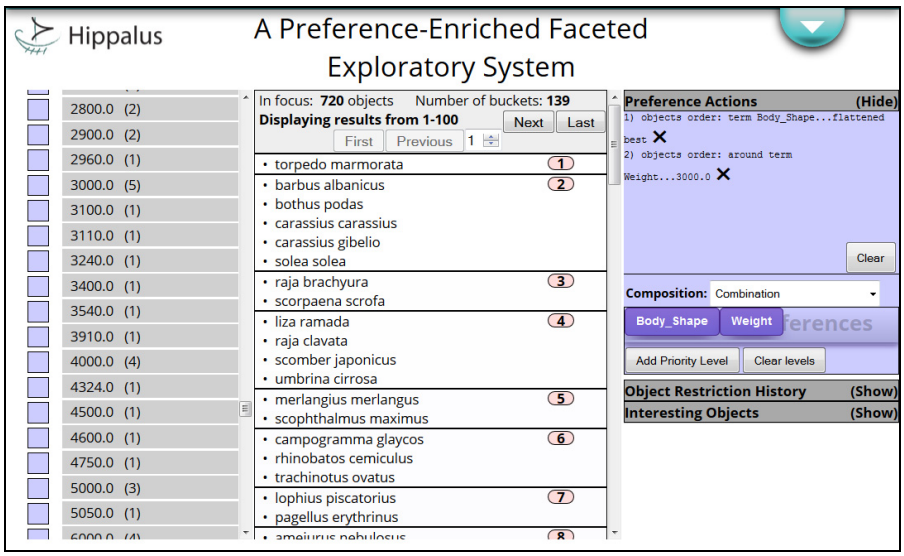


Figure 8 After the preference action *Worst(Country:Spain)*

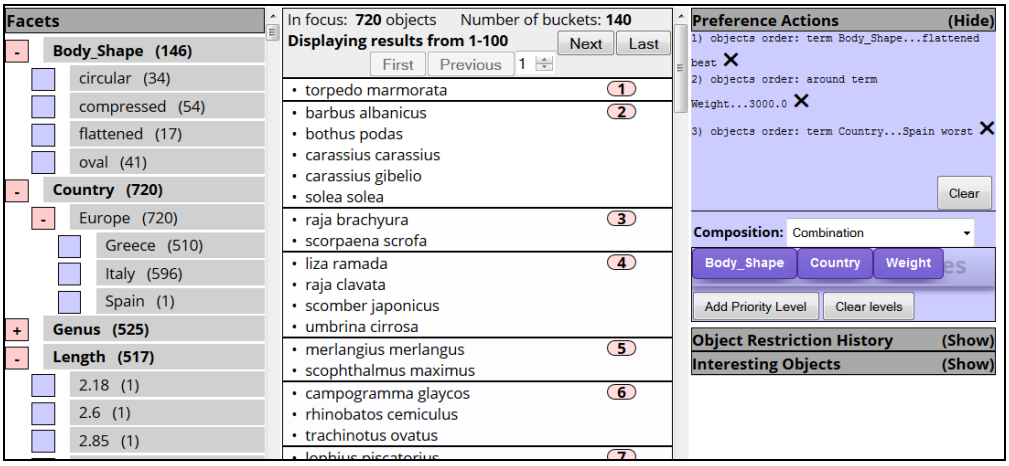


Figure 9 After the preference action *Best(BodyShape:Compressed)*

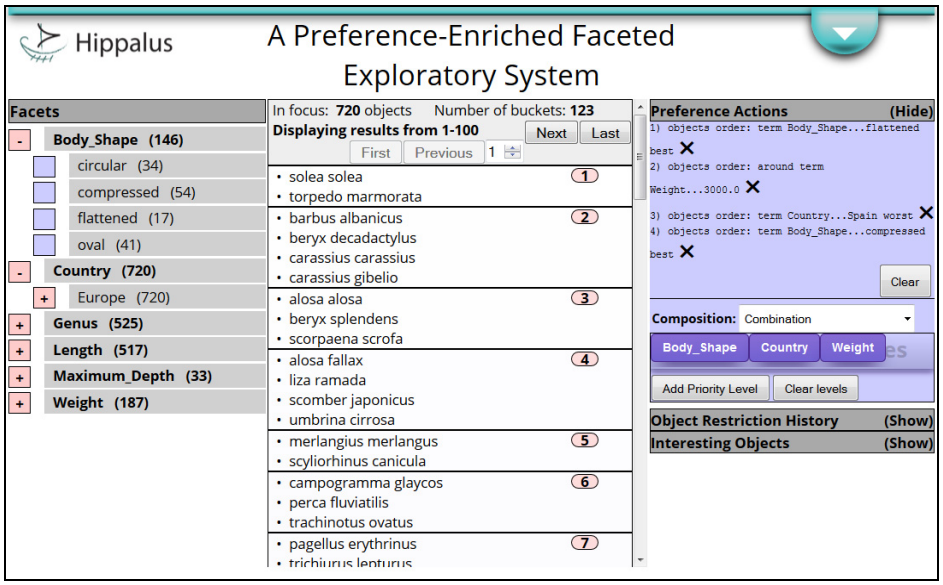


Figure 10 After the preference action *Around(Length:63)*

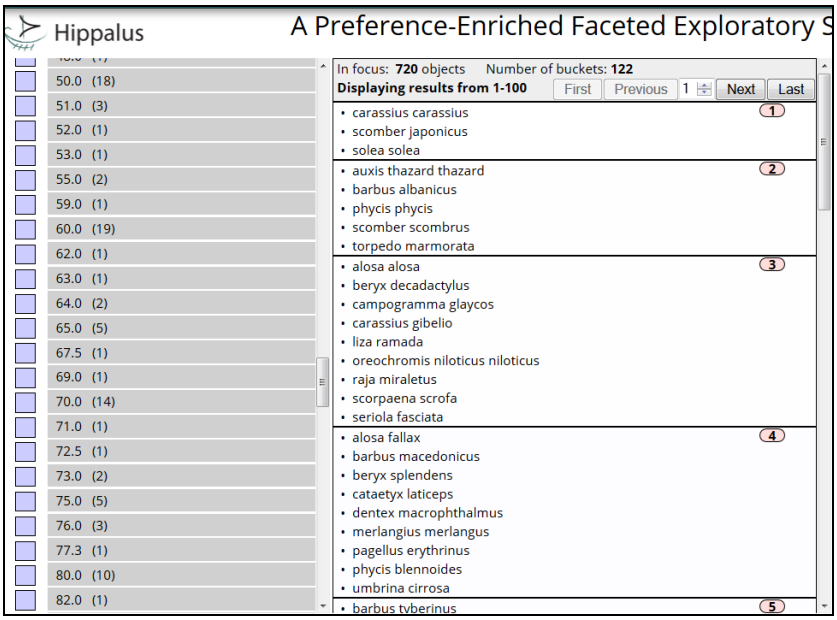
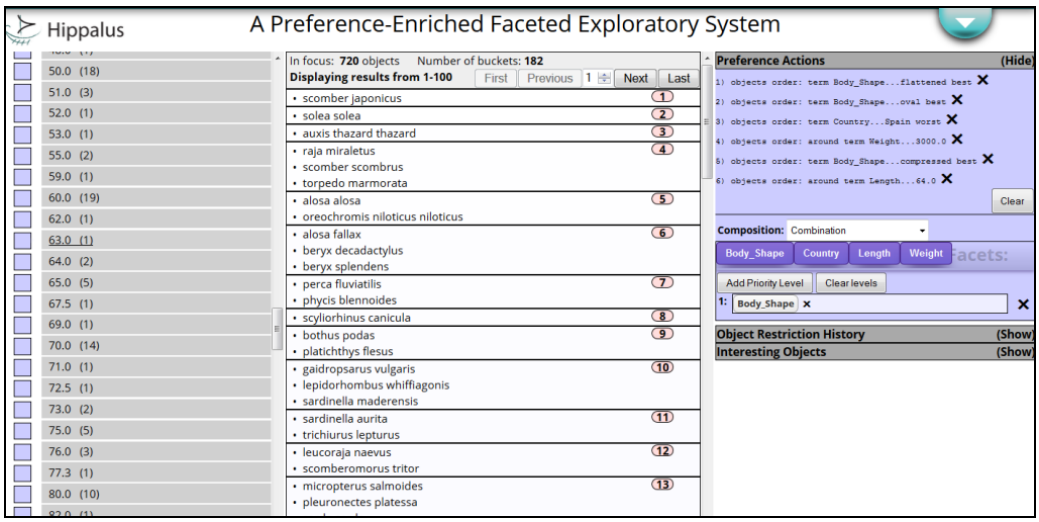


Figure 11 After prioritising first the preferences about Body Shape



3.3 *Larger and richer data set*

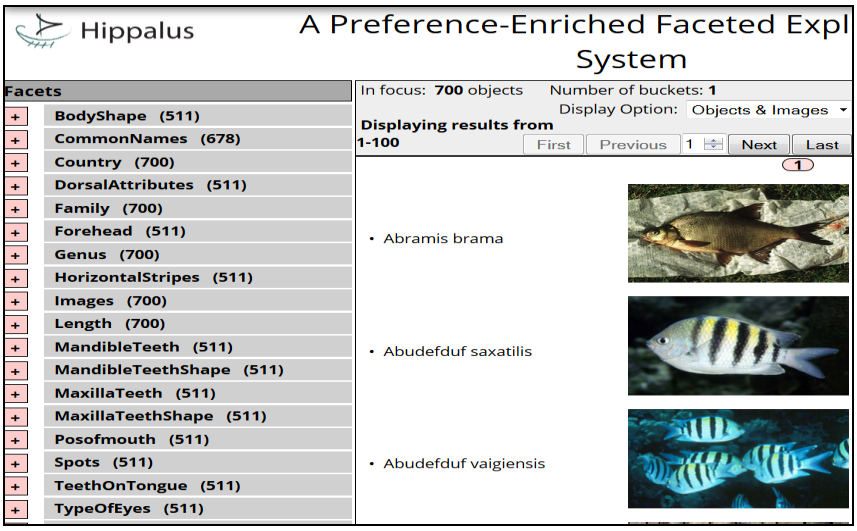
Based on the pilot phase, we decided to create and use a larger and richer data set that contains all species from FishBase and apart from the attributes described earlier, it also contains: (a) the '*preferred picture*' attribute of FishBase, (b) the family that each species belongs to, (c) more types of body shapes (e.g. '*oval*'), (d) the common names of each species, (e) dorsal fin attributes for each species (e.g. '*continuous with caudal fin*'), (f) forehead information (e.g. '*clearly convex*'), (g) absence or presence of horizontal stripes on body side, (h) absence or presence of vertical stripes on body side, (i) information about spots on body side, such as '*one spot only*', (j) type of mouth (e.g. '*clearly protrusible*'), (k) position of mouth (e.g. '*terminal*'), (l) absence or presence of mandible (lower jaw) teeth, (m) absence or presence of maxilla (upper jaw) teeth, (n) absence or presence of teeth on tongue, (o) mandible teeth shape, (p) maxilla teeth shape, (q) type of eyes such as '*eyes with fixed fatty (adipose) tissue/eyelids*', (r) type of scales (e.g. '*ctenoid scales*'), and (s) life zone (type of water) such as '*saltwater*'.

In total, this data set contains 32,471 species, described by 600,194 RDF triples and 23 facets.

Various optimisations of Hippalus were developed (specifically those described in Tzitzikas and Papadakos, 2013) for offering efficient interaction over this data set which is the biggest real data set that Hippalus has ever loaded. In brief it takes 5.7 s to load the data set, while actions that restrict the focus of the user are almost instant (e.g. restriction to species that belong to the Myctophidae family takes only 93 ms). On the other hand the computation of preference actions is more expensive, but can take advantage of the almost instant focus restriction actions. For example, lexicographically ordering species according to family (i.e. an expensive preference action due to the large number of preference relations) takes 17 seconds for 10,000 species, while it takes only 438 ms in the restricted focus of 720 species.

Moreover, and for supporting the identification through pictures, we have extended Hippalus with images, i.e. if one attribute value is a URL corresponding to an image, that image is displayed in the object list in the middle frame. A screenshot of the current prototype is shown in Figure 12.

Figure 12 Screenshot from the new version of Hippalus



3.4 Comparison of species identification methods

In this section we compare the PFS-based method with other fish identification methods, specifically with Morphometric Ratios, Dichotomous Keys and Eyeballing. As regards the support of images, only Eyeballing and PFS support images. The identification process is gradual in all methods except from the Morphometric Ratios. However, only in PFS (a) the sequence of choices is not predefined and (b) the user can express soft constraints. These features (support of images, gradual process, order-independent process, hard and soft constraints) are summarised in Table 1.

Since all methods (apart from Morphometric Ratios) support a process, it is worth to analyse it more. Roughly, each identification process is a series of selection or elimination of options as shown in the activity diagram in Figure 13 (left). Figure 13 (right) illustrates the process that is supported by PFS. Here, the process is essentially a series of *selection* statements, *ranking* statements and *inspection* statements.

For making more clear the commonalities and the differences, Figure 14 provides a taxonomy of *user actions* and *system responses*. *User actions* include *selections* of options based on attribute values, *elimination* of options based on attribute values and *ranking* of options based on attribute values. The latter can be expressed through *positive* statements, *negative* statements, or *relative* statements. The final category of user actions concerns *inspection* and includes actions for inspecting (i) the details (e.g. all attributes) of one option, (ii) the distribution of attribute values of the current set of options, and (iii) actions comparing two or more options. As regards *system responses*, they can be characterised as responsive if each user action immediately affects the user interface, or not responsive where only at the end of the process the user can see the results of his/her input. Another category of system response is whether a ranking of options is supported and if yes whether it is based on predefined knowledge or on user's confidence. We should mention at this point that there are computer-aided information systems, such as XPER3[13] and LucId[14] among others that offer the no predefined order and they may also provide weighing about easiness of observation for each characters, or guide the process by attributing frequency of occurrence in species in a given geographic area, but it has to be predefined by the creator of the knowledge database. Rather, PFS allows the user to express his confidence in his own decision, which is a noticeable progress.

At the right side of each feature of the taxonomy of Figure 14 there is one legend that indicates which of the four methods (referred here by A, B, C and D) supports that feature. It is evident that the PFS method (referred by D) is the method that supports most of the features. With 'Comparison of 2 or more options' we refer to the functionality of parallel display of a few options (like in systems for product selection, e.g. in Sacco [2003]). This feature is marked with '[D]' in the sense that although not implemented in Hippalus it is one of the classical features of Faceted Systems (this is called 'end-game' in Sacco [2003], and therefore it could be straightforwardly be supported by PFS. As regards the feature 'Ranking …/by Predefined background knowledge', PFS does not currently supports it. Overall we believe that PFS is a promising method for aiding species identification.

Table 1 Features of each identification method

Features \ *Identification method*	*Morphometric ratios*	*Dichotomous keys*	*Eyeballing*	*PFS-based method*
Support of images	No	No	Yes	Yes
Existence of gradual process	No	Yes	Yes	Yes
Flexibility of the process (order independent)	Non applicable (no process)	No	No	Yes
Soft constraints/preferences	No	No	No	Yes
Hard constraints/restrictions	Yes	Yes	Yes	Yes

Figure 13 The identification process

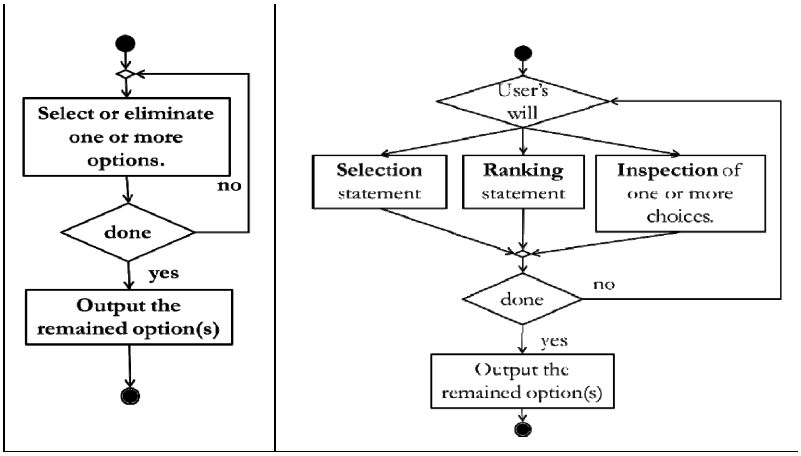
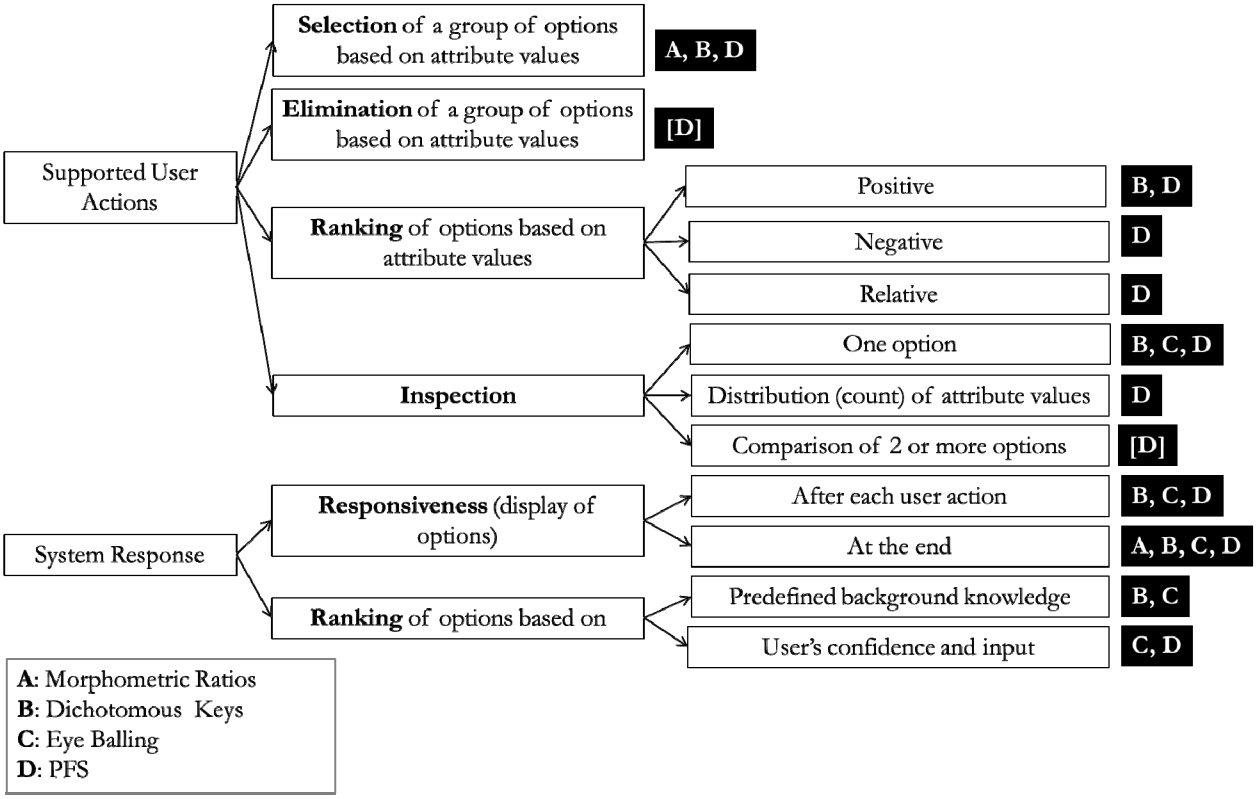


Figure 14 A taxonomy of user actions and system responses



3.5 Evaluation with users

As regards evaluation with users, Papadakos and Tzitzikas (2014) reported the results of a user study with a number of tasks related to a 'car selection' scenario. The results of the comparative evaluation, with and without the preference actions, were impressive: with the preference-enriched FDT, all users completed all the tasks successfully in 1/3 of the time, performing 1/3 of the actions compared to the plain FDT. Moreover, all users (either plain or expert) preferred the preference-enriched interface. The benefits were also evident through various other metrics.

We thought that it would be interesting to conduct a comparative task-based evaluation where casual users, i.e. non-marine biologist, would first receive a task description, actually a species description, and then they would have to use all the systems for identifying the sought species. Below we report our experiences from this attempt.

At first we realised that the systems are not very stable and mature, in the sense that they have several usability problems. This is an important issue in the sense that we would like to comparatively evaluate the methods, not the particular deployed systems or their underlying data sets. Nevertheless, we prepared a task description, specifically the one shown in Figure 15.

Figure 15 The task description that was used in the evaluation

Scenario: Suppose that you catch a fish while fishing in your boat close to south coast of Australia (Indian Ocean, Eastern, FAO Water Area). The caught fish is **around 760 gram**, its **length is between 38 – 41 cm**. The fish looks like the one in image bellow.

Important: Keep the time you spent even if did not manage to identify the fish species.

Then we performed a preliminary evaluation with a few users (including the authors). From this preliminary evaluation we realised that:

(a) As regards dichotomous keys, the user should start the identification process by selecting either a FAO water area, or an order or a family of the sought fish. This would be nearly impossible for casual users (i.e. users who are not marine biologists), since they should know the FAO water area that corresponds to every country, let alone the corresponding order or family of the fish. As a result most, if not all, of the users would give up the process at the first step.

(b) As regards eyeballing, although at first sight this method seemed to be very user friendly, when we tried it with a concrete task we realised that after a few clicks the user has to read a long list of descriptions (each comprised a drawing and a textual description). Moreover, after the second step of the identification process the (colourless) drawings of the different categories were almost identical, making it very difficult to come to a final answer (species).

In numbers, all persons that tried to carry out the task, using eyeballing and dichotomous keys, failed, i.e. they could not identify the sought species. Actually they quit the process. Based on this experience we decided that it would be meaningless to proceed to a comparative user study with a higher number of casual users. Since eyeballing and dichotomous keys could not be used with casual users (non-marine biologists), we decided to carry out a task-based evaluation of PFS with casual users also for collecting feedback about the system.

Apart from the aforementioned task description we prepared a short video tutorial of the system of 5.5 m, a web-based questionnaire and then by email we invited various persons to participate to the evaluation voluntary. The evaluation started on April and ended on May 2016. No face-to-face training took place, and no assistance was provided. The data set that was used contained 10,000 species described according to 23 facets. In numbers, 36 persons participated, ten (27.8%) female and 26 (72.2%) male, with ages ranging from 18 to 53 years. As regards occupation and skills, 14 (38.9%) of them were undergraduate student, 16 (44.4%) of them graduate students and six (16.7%) of them computer engineers or researchers.

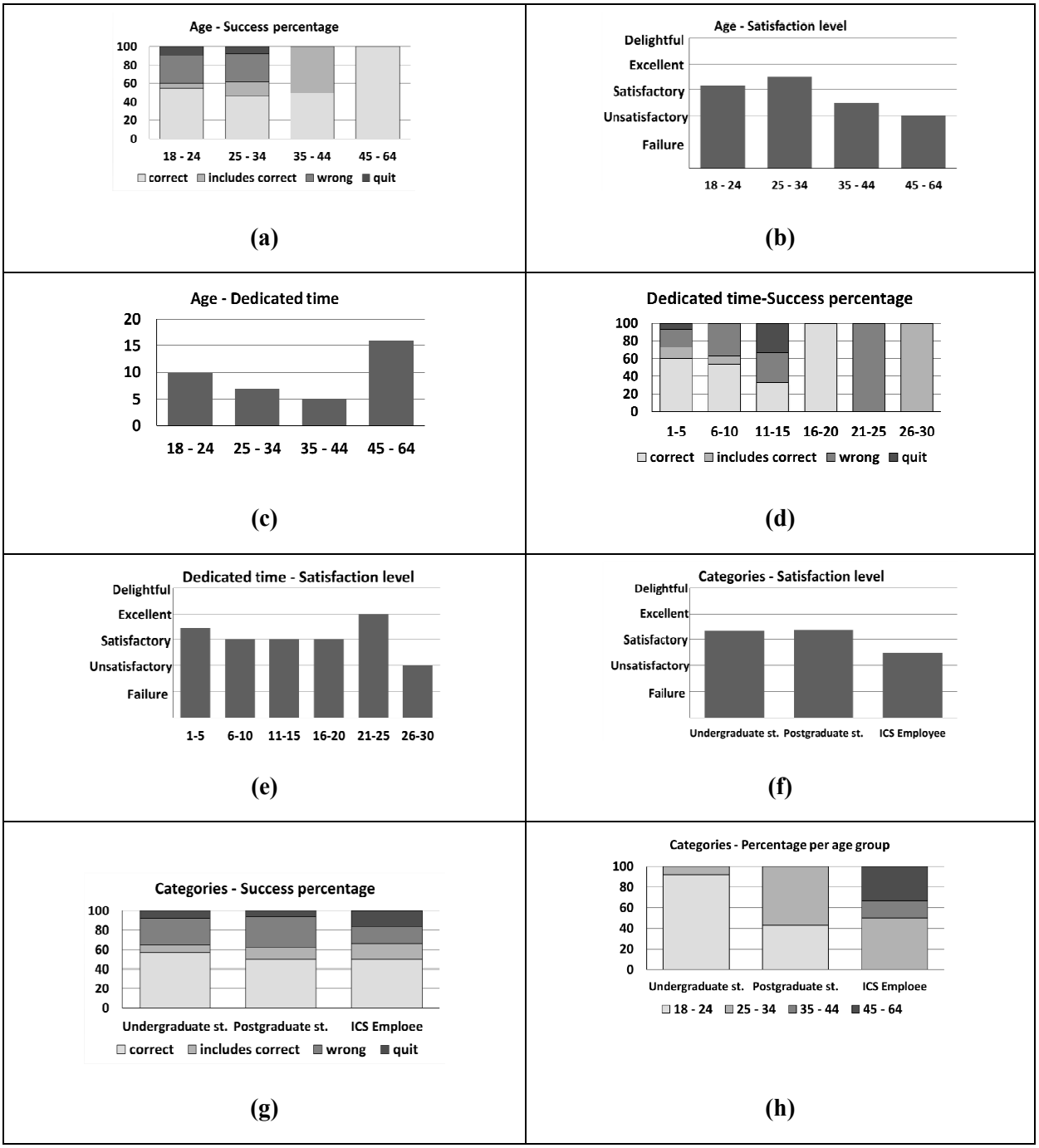
Evaluation results: Recall that with the previous methods (eyeballing and dichotomous keys) all users failed and quitted the process. With PFS only three users (8.3%) quitted the process, which is a significant improvement as regards the usability of the system and the process, i.e. 91.7% managed to complete the task. As regards the outcome of the task, 19 (52.8%) found the correct species, four (11.1%) provided more than one species (in the range from two and five, on average of 2.75) but one of them was the correct, and ten (27.8 %) provided wrong answer(s). As regards the time that each user dedicated, it ranged from 1 to 30 m, and the average was 9 m. As regards the satisfaction level, the questionnaire contained the values {Delightful, Excellent, Satisfactory, Unsatisfactory, Failure}, and the results were: Delightful: one person (2.8%), Excellent: 14 persons (38.9%), Satisfactory: 13 persons (36.1%), Unsatisfactory: eight persons (22.2%), Failure: none (0%). Figure 16 contains plots that provide additional information also useful for further analysing the results. For instance, one can see (from plot (a) and plot (c)) that the older in age participants although they dedicated more time and they declared smaller satisfaction level, they achieved the highest success percentage**.** From plot (d) we can see that those users that spent 16–20 m were the most successful. Those who spent more time were not successful; probably they followed a wrong path or they had difficulties with the system. In contrast, those who spend only 1–5 m had 50% success.

Overall the results are satisfying in the sense that although such tasks are difficult for casual users, very few quitted and a significant percentage of them managed to find the correct answer. Furthermore, the fish species in the task description was intentionally not ‘easy’ in the sense that it looked like an ordinary fish and did not have any discriminating feature. Moreover, we should mention that the video tutorial was quite short and not very detailed. According to our opinion, if the task contained a more rich description, or if the sought species had a more discriminating feature or appearance, then the success rate would be much higher.

In future we would like to conduct a comparative evaluation also with marine biologists. This is not currently possible in the sense that various aspects of the existing systems should be improved before doing that; moreover, the underlying data sets of these systems are not the same. In addition we should stress that, as Bailly et al. (2010) concluded, even for marine biologists the combination of more than one identification methods is suggested, e.g. eyeballing for finding the family and then dichotomous keys for finding the species of the individual. Therefore, based on this investigation we have concluded that a comparative evaluation with marine biologist should be organised in the form of a series of international events in the context of other symposia and conferences of marine biology. Nevertheless, we carried out a preliminary evaluation in the Hellenic Centre of Marine Research (HCMR) with six participants: two marine biologists, one person with biodiversity background, and three persons with no biodiversity background (employees of HCMR).

Three tasks were given corresponding to three different species (namely *Beryx decadactylus, Lobotes surinamensis* and *Elegatis bipinnulata*) each one described only by one image (no textual descriptions were given). The results were satisfying since, as regards the first species, all of the users identified it correctly, as regards the second species five/six users identified the correct species (the only wrong answer was submitted by a biologist) and as regards the third species, three/six users found it correctly (the three wrong answers came from the two biologists and the one user with no biological background). The comments of the users were generally positive, and their main criticism was about the lack of more images in the PFS interface, but this concerns the used data set (not the PFS method).

Figure 16 Analysis of the task-based evaluation with users



3.6 Applicability to other tasks

The PFS-based approach for identification is useful not only for species. It is generic in the sense that it can be applied over any kind of objects described by a number of attributes (or metadata), and it could be exploited for identifying a phenomenon in general. Also from a technical (implementation) perspective, Hippalus can load any data set expressed in CSV; consequently it can be used for identification tasks where each option is described by a set of attributes. Indeed identification tasks are important in many domains. Below we describe in brief a few identification tasks which could be benefited from PFS.

3.6.1 Identification of rare deceases

Faceted search can be used for the identification of rare deceases. Various facets could be useful for this task including age, sex, medical history, symptoms organised in categories, results of various kinds of examinations, living place, profession, etc. Faceted search has already been used in this context, specifically a computer-aided guided interactive diagnosis of rare diseases based of Faceted Search is described in Sacco (2008) and Sacco (2012). In that work the problem of the diagnosis of pathologies is recasted as a problem of exploring and thinning out candidate pathologies on the basis of clinical signs and other observable features.

Consequently, the preferences of PFS could further improve the process.

3.6.2 Identification of malware

The identification of malware is quite often a painful process. Faceted search could be used, and the list of possible facets could include: symptoms organised in categories (e.g. slowdowns, pop-ups, reboots, disabled security controls and antivirus), entries in the explorer of running processes, installed software in program files, registry entries, and others. As regards related works, a method of identifying malware activities using ontologies and rules is described in Jasiul et al. (2014). The method supports detection of malware at host level by observing its behaviour. The authors proposed and developed PRONTO which is a behaviour-oriented malware hunting tool that tracks suspicious activities. An approach for identifying malware behaviours based on operation and target using dynamic analysis technique is described in Zolkipli and Jantan (2011). This approach applies two ways of analysis, which are run time analysis, a malware-oriented analysis to identify malware activities, and resource monitoring, which monitors specific location of target attacks as host-oriented analysis.

In general faceted search and the preferences of PFS could aid the identification of malware.

3.6.3 Identification of precious stones

Methods to identify faceted gemstones are described in Devouard and Notari (2009). The proposed methods are based on the measurements of optical and physical properties, combined with acute observation using various illumination techniques. A simple guide on identifying gemstones has also been included in the Wikihow online community.[15]

Faceted search could be used for precious stones identification. The set of facets could include: optical properties (colour, relief, interference colours, and conoscopic observation), physical properties (measurement of specific gravity, magnetism, thermal conductivity), observation properties. Consequently, the preferences of PFS could further improve the process.

4 Concluding remarks

Species identification is usually carried out through methods that are based on morphometric ratios, dichotomous keys and eyeballing. Species identification is essentially a decision-making process comprising steps in which the user makes a selection that restricts subsequent choices. In this paper we have introduced a decision-making process for species identification which is not based on a strictly ordered set of 'absolute' decisions, but as any series (i.e. without a predefined order) of soft constraints (expressed as preferences). These constraints are expressed interactively and through simple clicks. The interaction model, called PFS, not only offers processes that are more flexible in the sense that the user can start from the easy to make decisions (or whatever decision seems easier), but also allows the decisions to be treated as preferences (i.e. not as hard constraints), therefore the outcome of the process is less vulnerable to errors since soft constraints do not exclude options; they affect only their ranking. Moreover, we have seen that by selecting values from facets that correspond to physical dimensions, the user can obtain what he could obtain with the Simple Morphometric Ratios-based identification approach.

The experimental investigation, performed using real data from FishBase, has demonstrated the feasibility of the approach and allowed us to identify, and subsequently implement, useful extensions of the system Hippalus (a system that supports PFS) based on the requirements of species identification. Subsequently we analysed the processes of the various identification methods for revealing their commonalities and differences and then we reported the results of a task-based evaluation with users. All casual users (not domain experts) that tried eyeballing and dichotomous keys failed to complete the task. However, with PFS 91.7% of the users completed the task, 63.9% of the users found the correct answer (the right species) or provided answers that included the correct one, while the average time was only 9 min. Finally we discussed other identification tasks that could benefit from PFS.

The current deployment is web accessible and in future we plan to release an improved version of the system. In future we would like to apply and evaluate the PFS-based identification method also to other kinds of identification tasks.

References

Agrawal, R. and Wimmers, E.L. (2000) 'A framework for expressing and combining preferences', *ACM SIGMOD Record*, Vol. 29, No. 2, pp.297–306.

Bailly, N., Reyes, R., Atanacio, R. and Froese, R. (2010) 'Simple identification tools in FishBase', in Nimis, P.L. and Vignes-Lebbe, R. (Eds): *Tools for Identifying Biodiversity: Progress and Problems. Proceedings of the International Congress*, 20–22 September, EUT Edizioni Università di Trieste, Trieste, Paris, pp.31–36.

Balke, W-T. and Güntzer, U. (2004) 'Multi-objective query processing for database systems', *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, Vol. 30, pp.936–947.

Chomicki, J., Godfrey, P., Gryz, J. and Liang, D. (2003) 'Skyline with presorting', *ICDE*, Vol. 3, pp.717–719.

Dakka, W., Ipeirotis, P.G. and Wood, K.R. (2005) 'Automatic construction of multifaceted browsing interfaces', *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ACM, New York, pp.768–775.

Dash, D., Rao, J., Megiddo, N., Ailamaki, A. and Lohman, G. (2008) 'Dynamic faceted search for discovery-driven analysis', *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, New York, pp.3–12.

Devouard, B. and Notari, F. (2009) 'The identification of faceted gemstones: from the naked eye to laboratory techniques', *Elements*, Vol. 5, No. 3, pp.163–168.

Fafalios, P., Salampasis, M. and Tzitzikas, Y. (2013) 'Exploratory patent search with faceted search and configurable entity mining', *Proceedings of the 1st International Workshop on Integrating IR Technologies for Professional Search, in Conjunction with the 35th European Conference on Information Retrieval (ECIR'13)*, Moscow, Russia.

Fischer, J. (Ed.) (2013) *Fish Identification Tools for Biodiversity and Fisheries Assessments: Review and Guidance for Decision-Makers*, Fisheries and Aquaculture Technical Paper No. 585, FAO, Rome, 107pp.

Fishburn, P.C. (1970) *Utility Theory for Decision Making*, No. RAC-R-105, Research Analysis Corp., McLean, VA.

Georgiadis, P., Kapantaidakis, I., Christophides, V., Nguer, E. and Spyratos, N. (2008) 'Efficient rewriting algorithms for preference queries', *IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008*, IEEE, Washington, DC, pp.1101–1110.

Griffin, L.R. (2011) 'Who invented the dichotomous key? Richard Waller's watercolors of the herbs of Britain', *American Journal of Botany*, Vol. 98, No. 12, pp.1911–1923.

Jasiul, B., Sliwa, J., Gleba, K. and Szpyrka, M. (2014) 'Identification of malware activities with rules', *2014 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, Warsaw, Poland.

Kießling, W. (2002) 'Foundations of preferences in database systems', *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB Endowment, Hong Kong, China.

Koren, J., Zhang, Y. and Liu, X. (2008) 'Personalized interactive faceted search', *Proceedings of the 17th International Conference on World Wide Web*, ACM, New York, pp.477–486.

Koutrika, G. and Ioannidis, Y. (2005) 'Personalized queries under a generalized preference model', *Proceedings of the 21st International Conference on Data Engineering, 2005. ICDE 2005*, IEEE, Washington, DC, pp.841–852.

Nimis, P.L. and Vignes-Lebbe, R. (Eds) (2010) 'Tools for identifying biodiversity: progress and problems', *Proceedings of the International Congress*, 20–22 September, EUT Edizioni Università di Trieste, Trieste, Paris, 455pp.

Pankhurst, R.J. (1991) *Practical Taxonomic Computing*, Cambridge University Press, Cambridge, 202pp.

Papadakos, P. and Tzitzikas, Y. (2014) 'Hippalus: preference-enriched faceted exploration', *ExploreDB 2014: Proceedings of the 1st International Workshop on Exploratory Search in Databases and the Web Co-located with EDBT/ICDT 2014*, Athens, Greece.

Papadakos, P. and Tzitzikas, Y. (2015) 'Comparing the effectiveness of intentional preferences versus preferences over specific choices: a user study', *International Journal of Information and Decision Sciences*, Vol. 8, No. 4, doi:10.1504/IJIDS. 2016. 10001391.

Peintner, B., Viappiani, P. and Yorke-Smith, N. (2009) 'Preferences in interactive systems: technical challenges and case studies', *AI Magazine*, Vol. 29, No. 4, pp.13–24.

Pu, P. and Chen, L. (2009) 'User-involved preference elicitation for product search and recommender systems', *AI Magazine*, Vol. 29, No. 4, pp.93–103.

Qarabaqi, B. and Riedewald, M. (2014) 'User-driven refinement of imprecise queries', *Proceedings of 30th International Conference of Data Engineering ICDE'14*, Chicago, IL, pp.916–927.

Renaud, C.B. (2011) *Lampreys of the World: An Annotated and Illustrated Catalogue of Lamprey Species Known to Date*, FAO Species Catalogue for Fishery Purposes No. 5, FAO, Rome, 109pp.

Rossi, F., Venable, K.B. and Walsh, T. (2009) 'Preferences in constraint satisfaction and optimization', *AI Magazine*, Vol. 29, No. 4, pp.58–68.

Sacco, G.M. (2003) 'The intelligent e-sales clerk: the basic ideas', *Human–Computer Interaction – INTERACT'03*, IOS Press, The Netherlands, pp.876–879.

Sacco, G.M. (2008) 'E-Rare: interactive diagnostic assistance for rare diseases through dynamic taxonomies', *19th International Workshop on Database and Expert Systems Application, DEXA'08 Workshops*, IEEE, Washington, DC.

Sacco, G.M. (2012) 'Global guided interactive diagnosis through dynamic taxonomies', *25th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE Press, Washington, DC, pp.1–6.

Sacco, G.M. and Tzitzikas, Y. (Eds) (2009) *Dynamic Taxonomies and Faceted Search: Theory, Practice and Experience*, Springer, Berlin.

Stefanidis, K., Koutrika, G. and Pitoura, E. (2011) 'A survey on representation, composition and application of preferences in database systems', *ACM Transactions on Database Systems (TODS)*, Vol. 36, No. 3, Article No. 19.

Tzitzikas, Y., Bailly, N., Papadakos, P., Minadakis, N. and Nikitakis, G. (2015) 'Species identification through preference-enriched faceted search', *Proceedings of the 9th Metadata and Semantics Research Conference (MTSR '15)*, Manchester, UK.

Tzitzikas, Y., Manolis, N. and Papadakos, P. (2016) 'Faceted exploration of RDF/S datasets: a survey', *Journal of Intelligent Information Systems (JIIS)*, pp.1–36.

Tzitzikas, Y. and Papadakos, P. (2013) 'Interactive exploration of multidimensional and hierarchical information spaces with real-time preference elicitation', *Journal Fundamenta Informaticae*, Vol. 122, No. 4, pp.357–399.

Wagner, A., Ladwig, G. and Tran, T. (2011) 'Browsing-oriented semantic faceted search', *Database and Expert Systems Applications*, Springer Berlin Heidelberg, pp.303–319.