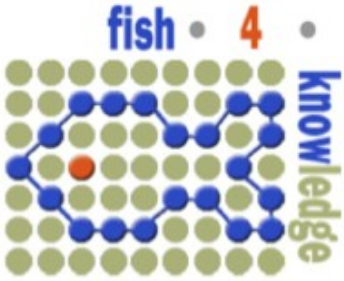


Fish4Knowledge Deliverable D2.1

User Information Needs

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document specifies the different types of information biologists would wish to have automatically extracted from videos. It also briefly describes the related biologists' research topics and data analysis tasks.

Deliverable due: Month 3

Contents

1. Introduction

1.1. About this document

To design the Fish4Knowledge tool, we will specify user needs by investigating the following questions:

1) What would biologists query our system for?
   This question investigates user information needs.
2) Why would they use our system?
   This question investigates user goals and tasks.
3) How would they use our system?
   This question investigates user interactions.

This document investigates the first two questions by identifying the basic user information needs (the "what") and the biologists' research topics that can be studied by using our video analysis tool (the "why"). We investigated these questions through a user study that is described in this document.

Following this document, the point 3) regarding the user interactions will be investigated by prototyping and evaluating user interfaces.

1.2. Content of this document

In section 2, we discuss the approach used to elicit user information needs and to conduct the user study. In section 3, we present the results of this study.

In section 4 we derive implications regarding the biologists' tasks we can support and the design of our tool.

Finally, section 5 gives an executive summary that concludes this step of the overall design process and section 6 introduces future work.

One important outcome of this first user study is the specification of basic system functionalities that bridge the gap between raw data from the database and information manipulated by users (Appendix VI "Basic user tasks and functionalities"). These functionalities are meant to facilitate the formulation of user queries.

2. Approach used to elicit user information needs

We discuss here the constraints we encountered to elicit user information needs (section 2.1.), and the approach used given these constraints (section 2.2.). Then we discuss the application of our approach to the user study conducted to elicit user information needs (section 2.3.).

2.1. Constraints

a) Novelty of video analysis in biology domain

Biologists currently use video cameras in their work, e.g. handheld cameras or Remotely Operated underwater Vehicles (ROV). Since automated analysis tools are relatively underdeveloped in the biology domain, biologists manually analyse their videos.

It is therefore difficult for us to recruit potential users, specifically biologists who are already familiar with video analysis. Further, without existing examples of video analysis tools, it could be difficult for interviewees to describe what usage of video analysis they could envision. Thus we have to explain them what tasks the system could perform and what data it can produce.

b) Time constraints

It is preferable to specify initial user needs as soon as possible because it will help our project partners to identify directions and requirements for their work.

A fully documented user-centred study would take a significant amount of time to interview a wide range of users about each particular aspect of the tool (e.g., working environment, tasks & goals, levels of experience).

Another time-consuming design task is to acquire domain knowledge about biology itself. This will allow us to design a tool that fits users' habits and mental model, and to facilitate our dialog with biologists we interview.

c) Availability of participants

We currently have contacts with 3 teams of expert users in Taiwan, Greece and the Caribbean. This is sufficient to derive initial user needs and system requirements. Recruiting more expert users faces 2 difficulties described above: the time constraint and the novelty of video analysis in biology domain.

2.2. Approach

Given the above constraints, we have to find a method that speeds up the process of discovering potential usages of our tool and eliciting the user information needs. Following a method introduced by Jim Gray, we used an approach that consists of describing the available data and ask potential users what data they would query. Principles and remarks concerning this approach are discussed in this section.

a) Application of Jim Gray's method

In 2004, Microsoft researcher and Turing award winner Jim Gray presented the

"20 Question Method" applied by his team to design a database system for astronomy researchers[1].

This methods consists of asking users what would be the 20 most important queries they would use to retrieve data. It is meant to solve similar issues we encounter in our project, and especially to bridge the gap between computer scientists knowledge and expert users knowledge.

To apply this approach, we explained to interviewees what data can be collected through video analysis and what external data about environmental conditions can be easily integrated in the system. We also investigated biology research concerns regarding the usage of these data.

b) Discussion of our approach

Our approach requires first defining global capabilities of the system before interviewing potential users about how they would use these capabilities. This contrasts with user-centred approaches that conduct user studies prior to the specification of system capabilities. A similar contrast can be observed in industrial innovation strategies of *technology push* and *demand pull*.

In principles, we can identify 2 parties involved in design processes: feasible solutions and practical purposes. Design processes makes one meet the other. Further, innovative design can be inspired by either one or the other party:

A) Existing purposes can inspire the exploration of new solutions (*demand pull*). In this case, innovation is grounded in the observation and description of practical purposes: current user goals, needs, methods and issues.

B) Existing solutions can inspire the exploration of new purposes (*technology push*). In this case, innovation is grounded in the specification of solutions' principles and capabilities.

Our European innovative project falls into the type B). Existing solutions of computer vision and video analysis are applied to biology domain and explores new purposes: the biologists' purposes. We grounded our study in the specification of video analysis capabilities we assume we will be able to support and develop during the project.

c) Possible biases

In our approach, given system functionalities are explained to interviewees in the beginning of the user study. As predefining a given system influences interviewees answers, we must be aware of possible biases.

Users may describe their tasks according to the given system functionalities, adapting their behaviour and expectations to what the system offers, instead

[1] http://research.microsoft.com/~gray/talks/SciData.ppt

of exactly describing their goals and requirements. Interviewees may not mention tasks that are not supported by the given system. Thus the study may not reveal information that could improve the overall users workflow or address yet unfeasible tasks.

This approach is likely to only partially cover the whole set of users tasks, requirements, issues and working environment. Our ambition is not to cover every relevant user task and to design a complete video analysis tool. Our goal is to explore the application of specific video analysis solutions to the biology domain. Thus these biases are not likely to compromise the design of the functionalities we will support.

2.3. Design of the preliminary user study

a) Goals of the study

To elicit user needs for meaningful information, we have to understand potential usages of the available raw data. Furthermore, this study will later be used to draft a user interface prototype. Thus the overall goals of this study are to explore:

- Relevant uses of video data in the biology domain,
- Data analysis tasks we can support,
- User information needs related to targeted data analysis tasks,
- Data manipulation performed by users.

b) Describing available data

To apply Jim Gray's method, we focused on the data that can be collected through video analysis and the external data that can be integrated in the system.

As described in the Deliverable 5.1. "Component interface and integration plan", data that can be automatically collected through video analysis are: the recognition of fish among other objects, the description of fish appearance (colour, shape), and the recognition of fish species and behaviour. We can also extract data about the quality of video images and the certainty fish recognition.

External data that can be integrated into the system are about weather (e.g., air pressure, occurrence of a typhoon) and water quality (e.g., salinity). We can also include location, date and time when specific events occurred (e.g., pollution, typhoon). These data would be collected through existing web sites and web services. If necessary, data describing environmental events would be manually entered by users.

c) Gathering interviewees

We interviewed biologists we already are in contact with. We lacked time to gather more potential users or to visit them in their university. Some of the interviewees had limited time to spend in the interview. There was also a language barrier as some of our biologists partners in Taiwan do not speak fluent English.

Thus we decided to conduct the interviews by mail or by phone, with a flexible interview structure, depending on the requirements and availability of the interviewees.

d) Content of the interview

Here is the text that was sent by mail to all participants, including those who answered the questions during a phone interview.

The Fish4Knowledge project description

This project aims at realizing a video analysis tool dedicated to the study of undersea ecosystems. Fixed underwater camera continuously record videos which are automatically analysed to detect fish species and behaviours.

---

1. Briefly, what are your scientific research goals & topics of interest?

(if relevant, please name biological patterns, processes or models implied)

---

2. What information, data or measures do you need to fulfil your goals?

---

3. How do you collect relevant data (manual methods as well as automated)? What trust or reliability issues do you encounter?

---

4. What tools do you use to process and analyse those data? What issues do you encounter while using those tools?

---

5. What would be the 20 most important queries you would ask the Fish4Knowledge tool?

For example, here are some sample queries:

- What species and numbers of fish appeared in the last N days?
- What unrecognised fish were detected? Do they cluster by appearance?
- Show me examples of fish from species X?
- Show me examples of a fish with description X?
- What other species were also present when species X was seen?
- Are the observed numbers of species X increasing in the past 3 years?

---

3. User study

This section describes the user study we conducted to elicit user information

needs. Individual reports of the interviews are available (see Appendices I, II and III).

In section 3.1. we present the interviewees, and in section 3.2. we discuss the analysis of interviewees' answers. Finally, in section 3.3. we summarise our main conclusions from the user study, and in section 3.4. we discuss feasibility issues.

3.1. Participants

a) Prof. Shao

Prof. Shao and his team from the Academia Sinica study Taiwanese coastal ecosystems. Among the ecosystems of the Island, they focus on coral fish and ecosystems' evolution correlated with human impact.

They have access to fixed underwater cameras in 4 different locations: 3 areas on the south of Taiwan, and one area on Orchid Island located around 70km away from the southern part of Taiwan island (see map). They also plan to use infrared cameras for night vision.

Figure 1 - Map of the southern part of Taiwan indicating locations of underwater cameras

An initial interview was conducted by mail in March 2011. It was completed during a project meeting organised in Taiwan in April 2011.

b) Dr. Day

Dr. Owen Day and his team study Caribbean coastal ecosystems, specifically in the context of the Buccoo Reef Trust projects (www.buccooreef.org). They are interested in launching a project using video monitoring and stereoscopic vision to evaluate the size of fish.

For Dr. Day, using cameras to collect data would be an alternative to diving observation: cameras would collect data in places where divers would spend a significant amount time collecting samples and observations. Cameras would stay fixed in sampling locations for a period of time until they are taken to another sampling location.

The interview was conducted in April 5th 2011 and consisted of a 45 minute phone call.

c) Prof. Stergiou

Prof. Stergiou and his team from the Aristotle University of Thessaloniki study Greek marine ecosystems. Prof. Stergiou is a member of the project advisory board. He currently does not have access to fixed underwater cameras.

The interview was conducted in April 12th 2011 and consisted of a 20 minute phone call.

3.2. Analysing study's results

To elicit user information needs we have to understand how raw data can be interpreted into meaningful information. We also aim at defining a consistent organisation of information regarding data analysis biologists would perform. Thus, we processed to the following analysis steps:

1. **Relevant uses of video data in biology domain**: we correlated biologists' research goals with the 20 most important queries they mentioned. This allowed us to understand how raw data collected through video analysis can be interpreted into meaningful information. This step is summarised in section 3.3.b.
2. **Data analysis tasks we can support**: after identifying topics of interest that can be studied through video analysis, we elicited those that belong to the scope of our project and that we can implement. This step is summarised in sections 3.3.c. And 3.3.d.
3. **User information needs**: we identified and organised information required by biologists, according to the supported user tasks and to the 20 most important queries. This step is summarised in paragraphs 4.1. and 4.2.
4. **Data manipulation performed by users**: we specified basic functionalities that allow users to formulate high-level queries. High level querying involve low-level data manipulation such as data selection and additional computation (e.g., rate in percentage). These functionalities facilitate interactions between the database (low-level queries) and the user interface. The related functionalities are detailed in Appendix VI "Basic user tasks and functionalities", along with examples of user queries.

3.3. Results

The main outcome of the user study is the most important information users would query the system for. We also gained insights about requirements for data to be effectively used. According to these insights, we identified user

tasks we can support.

a) 20 Questions

We collected a total of 27 important queries from the following sources:

- Prof. Shao explicitly formulated 20 most important questions (Q1-Q20).
- Dr. Owen Day expressed information needs during the phone interview. We derived 7 queries implied by his needs (Q21-Q27).

These queries are given in the following table.

Q1	How many species appear and their abundance and body size in day and night including sunrise and sunset period.
Q2	How many species appear and their abundance and body size in certain period of time (day, week, month, season or year). Species composition change within one period.
Q3	Give the rank of above species, i.e., list them according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.
Q4	Fish colour pattern change and fish behaviour in the night for diurnal fish and in daytime for nocturnal fishes.
Q5	Fish activity within one day (24 hours).
Q6	Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behaviour of various species.
Q7	Growth rate of certain species for a certain colony or group of observed fishes.
Q8	Population size change for certain species within a single period of time.
Q9	The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.
Q10	Immigration or emigration rate of one group of fish inside one monitoring station or one coral head.
Q11	Solitary, pairing or schooling behaviour of fishes.
Q12	Settle down time or recruitment season, body size and abundance for various fish.
Q13	In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic character to distinguish some similar or sibling species.
Q14	Association among different fish species or fish-invertebrates.
Q15	Short term, mid-term or long term fish assemblage fluctuation at one monitoring station or comparison between experimental and control (MPA) station.
Q16	Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.

Q17	The difference of using different camera lens and different angle width.
Q18	Is it possible to do the same monitoring in the evening time.
Q19	How to clean the lens and solve the biofouling problem.
Q20	Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.
Q21	What is the average body size for *species X?* How many percent of fish are *small*, *normal* or *big*?
Q22	What is the number of fish in *area X* for indicative species related to pollution?
Q23	What is the distribution and number of fish for indicative species of *factor X*?
Q24	What is the analysis of *factor X* impact, using *pattern of indicative data Y*?
Q25	What are the areas and periods of time of *species X* migrations?
Q26	What are the areas and periods of time of *species X* SPAGS[1]?
Q27	What are the SPAGS[1] periods in *area Y*?

Table 1 - The most important queries envisioned by potential users

b) Biologists' uses of data

Biologists described their research goals during the interviews. We observed that they would mostly use video data to study phenomena that underly the evolution of fish populations' composition (e.g., number of fish from each species, each body size range, each age) and geographical distribution.

We derived from the user study that population dynamics (i.e. the evolution of fish populations) is the main topic of interest of potential users of our tool.

We identified 4 underlying phenomena relevant for the study of population dynamics: reproduction, migration, trophic system (e.g., predation, food chain), and environmental factors (e.g., seasons, current) and events (e.g., pollution, typhoon). These topics would imply specific information needs and data analysis tasks.

c) Limitations

We identified the following practical concerns that might limit the range of data analysis tasks we can implement.

Sampling completeness: one fixed underwater camera can only collect data for a very small portion of the coastal area. Prof. Shao mentioned the "small window" effect, when too few cameras are not enough to get a representative overview of an area. Furthermore, some species may be underrepresented (e.g., small, nocturnal or cryptic species may never appear on the videos).

1 SPAGS: SPawning Agregation Sites, where fish gather to reproduce.

Sampling completeness is an issue in many biologists' research and they have methods to deal with it. But these methods may not be applicable in our case, as the underrepresented species and behaviours may not be the same.

Fish body size: this measure is necessary to evaluate the age of fish and their fertility. Dr. Day explained that when a fish size doubles, the number of eggs it produces is multiplied by more that 9.
Evaluating fish body size requires additional equipments (e.g., stereoscopic camera, infrared sensor) that are not in the scope of our project. With regards to our hardware resources, we could only evaluate changes of the average sizes of fish over time, using pixels as size unit, without being able to evaluate real fish size in centimetres. We do not know if this measurement would be valuable for biologists.

Trust and video analysis reliability: users expressed concerns about how complete and certain extracted data would be. We identified the following sources of uncertainty:

- **Recognition certainty**: some species and behaviour may be difficult or complex to identify (e.g., need to combine both shape and colour pattern). Thus identified species and behaviours may contain errors (e.g., false positive or false negative).
- **Need for scientific evidence**: to assess scientific research, biologists need to control how data are collected and processed. They would need to access and understand the layers of computation that produced the information they use. They are not concerned by technical computing details (e.g., performance and used CPUs), but by data transformations, possible biases, and replicability of data analyses.

Given these limitations, we may not be able to support data analysis tasks that requires:

- A extensive sampling coverage of the studied area (i.e. numerous or movable cameras are needed).
  This limitation particularly concerns the study of migration.
- The accurate detection of small, nocturnal or cryptic species, as well as any species and behaviour we might not accurately recognise.
  This limitation particularly concerns the study of reproduction and trophic system.
- The evaluation of fish size, and thus the evaluation of fish age or fertility that can be derived from fish size.
  This limitation particularly concerns the study of reproduction.

d) Biology topics and related tasks

We identified 5 biology topics that might be studied using video analysis data. Due to the limitations mentioned above (section 3.3.c.), we excluded the study

of 3 topics: trophic system, fish reproduction and fish migration. Their related information needs are described in Appendix V "Unsupported tasks".

As described in the Deliverable 5.1. "Component interface and integration plan", we expect to be able to automatically extract the following data from videos: fish contour, colour pattern, species and behaviour. On the basis of these raw video data, we assume we can support data analysis tasks related to the study of **population dynamics** (i.e. monitoring the various species living in an area) and **impact of environmental events**. The related tasks and information needs are described in Section 4.

3.4. Discussion on unsupported tasks and additional data

We identified the following unsupported hardware and software resources that would be required for biologists to be able to study fish migration, reproduction and trophic systems:

- **Domain knowledge**: integrate prior biology knowledge that can improve automated video analysis, or that can help biologists to interpret video analysis data.
  It is feasible to integrate domain knowledge by using our current software resources, but it would need extra research to investigate how our tool could benefit from domain knowledge, and how useful knowledge could be automatically extracted from existing resources. This would require reallocation of the project's human resources.
- **Extensive species and behaviour recognition**: provide the accurate detection of a consistent set of behaviours for a targeted set of species (e.g., the accurate recognition of all reproduction behaviours of species X).
  By the end of the project, we aim at being able to detect a wide range of species and behaviours. Currently we are not sure if we will be able to recognise the complete sets of behaviours needed to study reproduction or trophic system. Furthermore, some behaviours may not be recordable on video because they occur in specific timeframes and locations (e.g., night time, holes in rock).
- **Fish body size**: provide the size of fish and from this derive their age and their fertility. This would require extra equipments such as stereoscopic cameras.
  However, our current resources allow to evaluate the evolution of fish size (e.g., growth of young fish, ageing of population) calculated in pixels, without evaluation real sizes in centimetres.
- **Extensive camera coverage**: provide a sufficient number of cameras to cover the area of study. This would require extra equipments, such as cameras, that can be integrated to our tool without requiring specific or additional development.

- **Night vision**: provide equipment (e.g., infrared cameras) to record videos during night time. This might require specific development regarding the automated analysis of night vision images that have different characteristics than daytime images (e.g., light, colour, quality of images).

Some of these additional resources could also improve the study of population dynamics and environmental events. The table below summarises the impact of unsupported resources on data analysis tasks.

	Population dynamics	Environmental events	Trophic systems	Reproduction	Migration
Extensive camera coverage	+	+	+	++	**+++**
Night vision	++	+	+	+	++
Fish body size	+	+	+	**+++**	++
Extensive species and behaviour recognition	+	+	**+++**	++	+
Domain knowledge	+	+	++	+	+

Legend
+ *: the resource would generally improve the data analysis task*
++ *: the resource would dramatically improve the data analysis task*
+++ *: the resource is required for the data analysis task*

Table 2 - Impact of unsupported resources on data analysis tasks

Regarding these limitations on data that can be collected using video analysis (e.g., regarding fish size, nocturnal and cryptic fish), we could consider that manually collected data would provide additional missing data. But we assume that biologists do not merge video data and manually collected data into the same study, because it would create biases to merge data collected through different sampling processes.

Biologists may rather compare results from video data analyses with results from analyses conducted with other data collection methods. These comparison tasks are not in the scope of our project, our tool will only support video data analysis tasks.

The table below gives examples of additional data that can be manually collected. More details can be found in Appendix IV "Existing data collection and data analysis practices".

	Dead fish	Living fish		
	Caught by fishing	Caught by fishing	Observed by diving	Observed by video analysis
Taxonomic classification	+++	++	++	+
Behavioural data	-	-	++	+
Fish body size	+++	+++	+	-
Fish age	+++ *(derived from fish dissection)*	++ *(derived from fish size)*	+	-
Fish fertility	+++ *(derived from fish dissection)*	++ *(derived from fish size)*	+	-
Fish diet	+++	-	++	+
Chemicals in fish organisms	+++	+	-	-
Data regarding hiding fish (cryptic)	-	-	++	-

Legend
- *: this type of data can not be collected*
+ *: this type of data can partially be collected*
++ *: this type of data can be collected*
+++ *: this type of data can be very precisely collected*

Table 3 - Types of data that can be collected w.r.t. data collection technique

4. Targeted data analysis

In this section we discuss user information needs regarding the 2 biology topics that can be studied by using our tool: population dynamics (section 4.1.) and environmental events (section 4.2.).

4.1. Population dynamics

We derived from the user study that population dynamics is of general interest for biologists. It concerns the demographical study of populations of fish (i.e. fish of the same species living in the same location) or communities of fish (i.e.

fish from different species living in the same location). To support biologists, an important goal is to supply the count of fish from specific species observed at a specific time in a specific location.

We concluded that biologists would need the following information:

- Taxonomic classification of detected fish,
- Specific measurements based on counting fish and species,
- Evolution of data and the derived measurements over time,
- Statistics regarding detected behaviours occurrences.

a) Classifying fish

Taxonomic classification of observed fish is essential for biologists. Among the taxonomical ranks they are likely to use only **family**, **genus** and **species** (ranks above family are domain>kingdom>phylum>class>order>family).

b) Counting fish and species

Biologists use the following measurements, mentioned in the 20 most important queries:

- **Overall abundance**: also called community size, it consists of counting the total number of fish regardless of the species.
- **Abundance**: also called population size, it consists of counting the number of fish of a specific species. We assume that biologists would sometimes evaluate the abundance of a family or a genus.
- **Relative abundance**: the proportion of a population size (i.e., abundance) compared to the community size (i.e., overall abundance), expressed as a percentage (e.g., *100 * abundance of species X / overall abundance).*
- **Abundance level**: the range of values to which a species' relative abundance belongs. The levels could be the AFCOR scale (Abundant, Frequent, Common, Occasional and Rare species) or a simplified set of levels (Abundant, Common, Occasional, Rare).
- **Distribution in abundance levels**: the number and percentage of species belonging to each abundance level (Abundant, Common, Occasional, Rare).
- **Species richness**: the number of species.
- **Species composition**: biologists often use a set of data consisting of the species that are gathered at a specific location, during a specific period, and the fish count for each species. On this basis any metric mentioned above can be calculated. We assume that the main

measurements of interest are overall abundance, species richness and relative abundance of each species.
This data set has many names including: species composition, fish assemblage, community structure, community composition, composition of fishes, composition of fish assemblage, structure of fish assemblage, association among species.

c) Comparing measurements over time

Biologists are interested in the evolution over time of the measurements described above. The evolution of the above measurements can be studied w.r.t. different time units such as hour, day, month or year. For instance, the evolution of abundance (called *growth rate*) can be calculated from one month to next month, comparing fish counted each month:

100 * ("*number of fish in April*" - "*number of fish in March*") / "*number of fish in March*"

The evolution of overall abundance (i.e., *overall growth rate*), abundance (i.e., *growth rate*), and species richness (i.e., *species richness evolution*) are studied using similar percentage calculations:

100 * ("*measurement for period 2*" - "*measurement for period 1*") / "*measurement for period 1*"

The evolution of species composition (mentioned by Prof. Shao as fish assemblage fluctuation or **species composition change**) is studied through the evolution of overall abundance, species richness and abundance for each species.

d) Behaviours

Behaviours can give additional insights on interactions between populations, particularly regarding reproduction or trophic systems. Thus, we support 2 measurements: the **number of occurrences** of a specific behaviour, and the **percentage increase of behaviour occurrences**. For instance, we can calculate the percentage increase of feeding behaviours observed for species X between today and yesterday:

100 * ("*number of behaviour today*" - "*number of behaviour yesterday*") / "*number of behaviour yesterday*"

Multiple fish might be implied in behaviours, and even fish from different species. For instance preying imply a "predator" species and a "prey" species.

e) Selecting data to calculate measurements

To calculate the above measurements, biologists choose data to use by selecting the period of time and location of interest. They also choose the species to take into account by selecting specific families, genus or species of

interest.

We derived from one of Prof Shao's publications[2] that very sparse species (e.g., that are observed less than 5 times) can bias the representativeness of samples (e.g., the video samples, the detected fish) as the real population might actually not contain the same number of fish from these sparse species. Furthermore, measurements and their variations among samples can be over-exaggerated as a one fish difference can greatly impact calculated values.

To balance this uncertainty regarding proportions of species and significance of measurements, it may be necessary for biologists to exclude very sparse species (i.e., species identified a too small number of times) from measurements (e.g., overall abundance, species richness, species composition). Our tool should allow biologists to define an abundance threshold to select species to be taken into account.

4.2. Ecological impact of events

In addition to studying gradual, long-term population dynamics, biologists are also concerned with the impact of shorter-term phenomena such as environmental events (e.g., typhoon, extreme temperature, pollution).

a) Describing environmental events

We will consider different **types of event** because biologists will need to analyse certain types of data depending on the nature of events. The targeted types of event are either natural (e.g., typhoon, extreme temperature, current, pressure) and artificial (e.g., chemical spills, modification of habitats, noise disturbance).

Events have **starting and ending dates**, as well as **areas of occurrence**. Some of them may be cyclic (e.g., typhoon, currents) with expected starting or ending dates.

Data describing environmental events can be supplied by external databases (e.g., weather information, sensor networks) or can eventually be manually entered when no data sources are available. A large variety of environmental data can be collected about: **current**; **turbidity**; **salinity**; **temperature**; **pressure**; **wave**; oxygen concentration; acoustic levels; phytoplankton, algae or chlorophyll concentration; sunlight absorption; conductivity; phosphoric acid and other chemicals presence. The first six measurements (in bold) appear to be the most commonly used as a basis to describe environmental conditions. Depending of the type of environmental event, other specific measurements can be used (e.g., certain pollutions can be analysed using chemical concentration, and others with algae concentration).

[2]*Effects of habitat modification on coastal fish assemblages* – Journal of Fish Biology - 2010

b) Impact of environmental event

Biologists can correlate environmental data with changes in measurements we can supply (e.g., metrics mentioned in section 4.1.). Timeframe and location of environmental events can be used to select video data, and to compare the situation before, during and after an event.

Data visualisation displaying event timeframes would help biologists to identify the impact of an event. Data visualisations can also include graphic display of environmental data.

c) Indicative patterns of data

Dr. Day mentioned that some species, called **indicative species**, are especially sensitive to specific environmental conditions (e.g. species X is never observed in a polluted environment). Data about their abundance or behaviour can be used as an indicator of the occurrence or impact of an event.

Depending on events' types, biologists might wish to gather measurements regarding specific species, genus or families, or specific behaviours. Biologists would use specific datasets to investigate specific types of event. For instance, one type of event can be investigated by analysing the abundance of species X, Y and Z, along with the water temperature.

5. Summary and implications for system design

As an executive summary, this section summarises the user tasks to support and their information needs (section 5.1.), and presents the related implications regarding the design and the development of our tool (section 5.2.).

5.1. Supported biologists' tasks

We support two data analysis tasks, analysing **population dynamics** and analysing **impacts of events**, and one video search task. We summarise below the information needs for each task.

We also derived basic functionalities that allow to extract the needed data from the database in order to supply users with the needed information. These functionalities are specified in Appendix VI "Basic user tasks and functionalities", which also contains examples of user queries.

a) Population dynamics

To study population dynamics, we will support the following measurements that evaluate the evolution of number of fish and number of species:

- **Overall abundance**: the total count of fish regardless of species.

- **Overall growth rate**: the evolution of overall abundance over time, expressed as a percentage.
- **Abundance**: the count of fish from a single species.
- **Growth rate**: the evolution of abundance over time, expressed as a percentage.
- **Relative abundance** and **abundance level**: the proportion of a fish species in the overall fish community, expressed as a percentage or as ranges of values.
- **Distribution in abundance levels**: the number and percentage of species belonging to each abundance level.
- **Species richness**: the number of species.
- **Species richness evolution**: the evolution of species richness over time, expressed as a percentage.
- **Species composition**: a set of measurements composed of the overall abundance, the species richness and the relative abundance of each species.
- **Species composition change**: a set of measurements composed of the overall growth rate, the species richness evolution and the growth rate of each species.
- **Behaviour occurrence**: the number of detected occurrences of a specific behaviour.
- **Behaviour occurrence evolution**: the evolution over time of a behaviour occurrence, expressed as a percentage.

All these metrics need to be calculated using a specific locations, timeframes and time units (e.g., daily growth rates of species X calculated for month Y). Instead of using species, biologists may calculate measurements using genus or family.

Biologists might also discard species that do not occur a sufficient number of time (i.e., use an abundance threshold) to evaluate the following measurements: overall abundance, overall growth rate, relative abundance, abundance level, species richness, species richness evolution, species composition, and species composition change.

For some behaviours, it is necessary to indicate the multiple fishes implied, their species and their role (e.g., species X is preying on species Y).

b) Environmental events

The study of impact of events consists of analysing the changes in population dynamics occurring before, during and after events. Users will also need to

manually describe events and consult the related weather data. Thus, the needed information are:

- **Event definition**: name, time, location, type and descriptive text (optional) of an event.
- **Weather information**: weather data extracted from external databases (e.g., pressure, temperature).
- **Event impact**: a set of custom metrics selected among those used to study population dynamics, and calculated for periods before and/or after the event.

c) Video search

Consulting video excerpts seems to be less important than studying population dynamics or environmental events. However, this allows users to manually indicate species and behaviours that have not been consistently recognised by video analysis. Moreover, users will verify videos when suspecting a technical problem with camera (e.g., cleanliness of the lens) or visibility (e.g., water turbidity, object obstructing the view). Thus, we identified the following information needs:

- **Search video**: list of videos retrieved using a free text query, or by indicating timeframe, location, event, species, genus, family, or behaviour of interest.
- **Video tags**: list of fish detected in a video along with the recognised species, genus, family, and behaviours.
- **Modify tag**: users can add, remove or modify species, genus, families or behaviours that have not been correctly recognised by video analysis. Modifications and their authors would be logged.

5.2. Implications for system design

From the user study, we derived the following implications regarding the design and the implementation of our tool.

a) Data

Taxonomic classification: In addition to species recognition, family and genus recognition is also valuable to biologists, especially when a fish species cannot be recognised but its genus or family is recognisable. The Fish Recognition component should not only identify fish species but also genus and family.

Data schema: The current database schema (described in deliverable 5.2.) supports all data types needed by software components to process the videos. However, it does not support all those needed by users. We will need to add

external data used to study environmental factors (e.g., weather data), or links to external resources that can supply these data.
We can also investigate simple ways to integrate basic existing domain knowledge (e.g., species X is nocturnal).

b) Algorithm and workflow

Behaviour recognition: For some species, colour pattern is an indicator of recent or present activities (it can be seen as an expression of fish "mood"). This is an important feature for detecting fish behaviours, even if behaviours are not happening in front of the camera but happened a few moments before the fish is detected.

Workflow: The above point on behaviour recognition impacts the Workflow Component (CF Deliverable 5.1 - Component Interface and Integration Plan). Once the species is recognised (whether using colour patterns or not), we can then use colour patterns to recognise behaviours. The workflow should be able to derive a behaviour from the association of one colour pattern occurring for one specific species.

Computing biologists' measurements: Counting fish and species, and other measurements and statistical data have to be calculated. Some of them may be frequent and resource consuming. We will study how to optimise those measurements, for instance by precomputing them.

Fish clustering: The system will be able to cluster fish by using their visual appearance (e.g., colour pattern, fins, tail or body shape) or any other characteristics (e.g., similar periods of predation behaviours). We will investigate potential functionalities that can be supported by the clustering technique.
For instance, it might improve taxonomic classification when there is uncertainty in fish recognition. Biologists might also be interested in clustering fish regarding periods of growth (increase of a population size), periods of occurrence of specific behaviours, or regarding reactions to environmental changes.

c) User interface

Querying video excerpts: Biologists appeared to be more interested in consulting measurements and statistical data extracted from video analysis rather than watching video excerpts themselves. We anticipate that video excerpt will be most useful for biologists to check where uncertain classifications have been classified correctly.

Using morphological features: low level morphological features, such as type of fins or tails, are not likely to be used in high level user interactions, to calculate measurements (e.g., abundance of fish with tail of type X) or to retrieve videos (e.g., show me example of fish with tail of type X). We initially

envisioned these kind of user interactions, but they are not a priority.
Biologists would rather use family or genus, or other characteristics that refer to existing domain knowledge about fish species such as nocturnal, cryptic or migratory species.
One goal of the analysis of fish shape is to determine fish species. So far no other goal of fish shape analysis was mentioned by our interviewees.

Domain oriented concepts: We will investigate what human understandable data structure could bridge the semantic gap between raw data and information manipulated by users. A user ontology would provide a human understandable labelling system (this *behaviour* is called *spawning,* this type of *tail* is called a *lunate tail,* this *location* is called *NPP-3*), as well as an organised representation of relationships between ecosystem entities and characteristics (e.g., this *camera* is located in *rocky habitat*, *spawning* and *mating behaviours* can be recognised and studied for this *species*). The user ontology might also integrate basic domain knowledge (e.g., species X is nocturnal, species Y feeds on plankton).

6. Future work

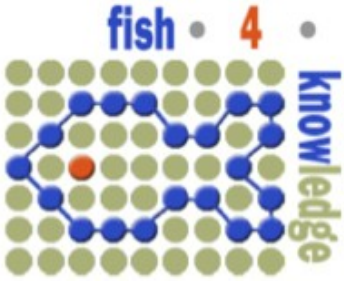
To explore further the design and implementation of our tool, we will now build a prototype of our tool. This will allow us to:

- Identify possible engineering issues.
- Confirm the identified user tasks and information needs.
- Study the needed user interactions.

We will follow 6 design steps:

1. Identify user scenarios and describe them in the Deliverable 2.2.
2. Draft a prototype of the user interface that allow users to interact with data from the actual database, and identify engineering issues.
3. Conduct a user study to explore user feedback regarding the prototype.
4. Derive refinements of user tasks and information needs.
5. Implement targeted refinements and proceed to a new user study.
   We will eventually conduct studies that are dedicated to a specific aspect of user interactions (e.g., to evaluate different forms of uncertainty visualisation).
6. Finally, we will specify and implement the user interface that integrates the video analysis components.
   This will be done once additional studies allowed us to confirm the information needs we identified and the feasibility of the system functionalities, and to derive the needed user interactions.

Appendices



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix I
Report on the interview of Prof. Shao

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document reports essential information exchanged with Prof. Shao in the context of our user study.

In March and April 2011, we conducted a user study to elicit user information needs related to the data analysis tasks we can support w.r.t. the data we can collect from videos. This work is discussed in the Deliverable 2.1 "User information needs".

Prof. Shao and his team from the Academia Sinica participated in the user study. They study Taiwanese coastal ecosystems., and have access to 10 cameras fixed in the seabed in 3 different locations: 2 areas on the south of Taiwan, and one area on Orchid Island (located around 70km away from the southern part of Taiwan). They also plan to use infrared cameras.

This document reports the information exchanged by mail with Prof. Shao during March 2011, i.e. the original text of the answers given by him and his team to the questions of the interview.

Question 1 - Briefly, what are your scientific research goals & topics of interest?

My academic research goals are to understand the fish diversity in Taiwan and in the world including how many of them, how they are distributed, how and why their composition and population change as well as how to restore their resources or how to decrease the rate of biodiversity loss. My study fields are quite broad, from taxonomy, ecology, evolution and conservation to database integration.

Research interests: (Red colour may relate to your project)

- Systematics of marine fish especially of deep-sea fish in recent years.
- Identification of fish eggs using electronic microscope and DNA sequencing.
- Barcode of life project regarding fishes and its application for fish biodiversity studies and sustainable fisheries.
- Phylogeography of marine fish in West-Pacific.
- Community ecology of marine fishes and its application on marine environmental and fishery resources assessment.
- Ecosystem trophic modelling for various coastal areas.
- Fish database and biodiversity informatics.

Completed or ongoing projects in 2008-2010: (Red colour may relate to your project)

- Deep-sea fish diversity studies—systematics and life history. (NSC 2007.8- 2010.7)
- Systematic studies on Anguilliformes, Pleuronectiformes and Callyonymidae. (NSC 2010.8-2013.7)
- Coordinate Taiwan GBIF (TaiBIF) & Taiwan Biodiversity Information Network (TaiBNET) (NSC and COA, 2001-)
- EIA monitoring of fishes at NPP I, Ⅱ& Ⅳ (Taiwan Power Company, 2001-)
- Cryobanking and database establishment of wildlife in Taiwanese fishes (COA, 2004-)
- TELDAP (Taiwan E-learning and Digital Archive Program) fish project and international portal project (NSC grant, 2006-2012)
- Marine biodiversity census at a global biodiversity hotspot (Philippines)—fishes (AS grant, 2009-2012)
- Benthic and deep-sea fish diversity studies in South China Sea and its database (NSC 2007.8-2010.7)

- Feasibility studies for establishing Taiping Island, Spratly Island as a National Park. (MOI, 2008.5-2009.4)
- Fish egg and larval fish studies at Tungsha Island, South China Sea. (MOI, 2011.3-2011.2)
- Underwater real-time video monitoring for coral reef system at the intake area in the 3rd Nuclear Power plant in Taiwan (Taipower Company, since 2004)

For academic services, I help the government to do the following tasks:

- Promote the establishment and enforcement of marine protected areas and marine conservation in Taiwan. (including promoting movies "The End of the Line" & "Seafood Guide in Taiwan")
- Promote sustainable use and management in coastal and offshore fisheries, control the import of aquatic organisms, and assist ecosystem-based fishery management.
- Integrate biodiversity databases or networks in Taiwan and link to global databases, such as FishBase, GBIF, OBIS, WoRMS, CoL.
- Publish the first edition of "National Species Checklist in Taiwan" in hard copy, CD-ROM and electronic version in 2008 which includes 50,000+ native species. The 2nd edition was published in 2010 and the species number increases to 53,000+.
- Organizing the IUCN-SSC, Global Marine Species Assessment meeting in Asian-Pacific region, for tuna, billfish, sea bream, and lobster in November 2009.

2. What information, data or measures do you need to fulfil your goals? And what trust or reliability issue do you encounter?

For ecological studies, certainly how to get more accurate data or real time data is very important and is a future trend. It will be even better if the massive data can be analysed by some data mining techniques from accumulated database. Data collected include species, individual number, body size observed. According to my own research experiences and interests, I can list the following projects which I have done or am conducting at the moment:

- Natural coral reef or artificial reef fish monitoring project, including their community structure and population fluctuation.
- Malformed fish (caused by thermal plume) monitoring project for the 2nd Nuclear Power Plant at the outlet bay.
- Use acoustic method to repel fish to swim into the intake area for decreasing the impingement possibility at Nuclear Power Plants.

- Recruitment, larval fish settlement, behaviour or species interaction studies for reef fish.

3. What tools can provide useful information? And what usability issues do you encounter while using those tools?

I do not understand what kind of information you like to know about this question. If *tools* means sampling tools, then my answer includes: hand collecting from fish market, and field trip by using hand net, various fishing net, angling, light trap, larval fish net or dredge either by diving or taking the boat, or research vessel. Studied marine habitats include intertidal zone, coral reef, estuaries, coastal areas, deep-sea or open ocean.

If we use video monitoring to collect observational data, the problem will be:

- No cryptic species, small size fishes could be observed;
- Not easy to get nocturnal species or observations in evening time even if using infrared camera;
- Low water visibility and visibility changes all the time which will limit the observation distance;
- Quality and resolution power of camera and image transfer;
- Many fish species could not be recognized or distinguished based on images if their colour pattern are similar each other;
- Duplicated counts of individual number when fish swim in and out of the observing range. So only resident species are much easier to be observed rather than migratory or semi-resident species.

4. What would be the 20 most important queries you would ask the Fish4Knowledge tool?

1. How many species appear and their abundance and body size in day and night including sunrise and sunset period.

2. How many species appear and their abundance and body size in certain period of time (day, week, month, season or year). Species composition change within one period.

3. Give the rank of above species, i.e., list them according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.

4. Fish colour pattern change and fish behaviour in the night for diurnal fish and in daytime for nocturnal fishes.

5. Fish activity within one day (24 hours).

6. Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behaviour of various species.

7. Growth rate of certain species for a certain colony or group of observed fishes.

8. Population size change for certain species within a single period of time.

9. The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.

10. Immigration or emigration rate of one group of fishes inside one monitoring station or one coral head.

11. Solitary, pairing or schooling behaviour of fishes.

12. Settle down time or recruitment season, body size and abundance for various fish.

13. In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic character to distinguish some similar or sibling species.

14. Association among different fish species or fish-invertebrates.

15. Short term, mid-term or long term fish assemblage fluctuation at one monitoring station or comparison between experimental and control (MPA) station.

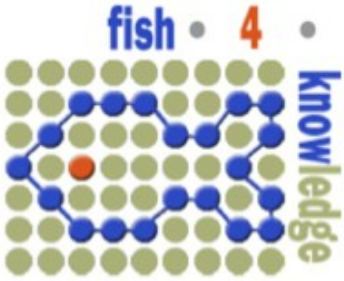
16. Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.

17. The difference of using different camera lens and different angle width.

18. Is it possible to do the same monitoring in the evening time.

19. How to clean the lens and solve the biofouling problem.

20. Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix II
Report on the interview of Dr. Day

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document reports essential information exchanged with Dr. Day in the context of our user study.

In March and April 2011, we conducted a user study to elicit user information needs related to the data analysis tasks we can support w.r.t. the data we can collect from videos. This work is discussed in the Deliverable 2.1 "User information needs".

This document reports essential information exchanged with Dr. Day during a 45 minute phone call on April 5th 2011.

Dr. Day and his colleagues' needs are quite similar to those of Prof. Shao's team. Based on additional information exchanged in this interview, we derived 7 additional queries to express complementary user needs.

1. Introduction

- Owen Day and his colleagues study the Caribbean ecosystem, which is

quite similar as the Taiwanese ecosystem (similar species of fish & coral).
- They mainly study population dynamics, migration, reproduction and the health of Marine Protected Area (MPA).
- One of their main interests are methods to determine the best boundaries of Marine Protected Areas.
- The main external factors they monitor are the impact of pollution, fishery, and tourism.
- They are interested in video monitoring tools because some fishes flee divers as they could be spear fishers.
- Video monitoring would also facilitate their study by avoiding biologists to spend long hours in diving to observe fish.

2. General needs

- To study population dynamics, the most needed features are fish species, number of fish and body size of fish.
- Behaviour analysis is not the most important feature they expect.
- They would like the system to be deployed on performant but standard computers, without having to invest on very high performance facilities.
- Therefore, they are interested in the workflow capacity to manage load balancing.

3. Fish body size

- Studying fish body size is very important for their research.
- A major reason is because it is directly related to fish fertility (e.g., if a fish size is doubled, the amount of eggs it can produce is multiplied by more that 9).
- A minor reason is that body size could help to detect fish species.
- They are interested in using sets of 2 cameras for stereoscopic vision, as this technique is far more accurate than using a single camera.
- The biggest species they would need to detect is 1 meter long on average, and between 1.5 to 2 meter long as a maximum. They assume they could use the same cameras and lenses (angle width) to detect smallest and biggest fishes.

4. Indicative species and standard patterns

- Some species are especially sensitive to ecological factors (e.g., some species are never found in polluted area). Biologists consider those species as indicators of specific factors such as pollution or overfishing, and call them *indicative species*.
- Beside studying one single species, biologist can also use patterns of data including several species (e.g., distribution of herbivorous species).

5. Migrations & reproduction

- Some species are sedentary, and other are migratory.
- Massive migrations can occur, especially for feeding or reproduction matters.
- Some species reproduce on the same areas and the same periods of time, that are called SPAGS for Spawning Aggregation Sites, and that particularly need to be monitored and protected (e.g., by defining Marine Protected Areas).

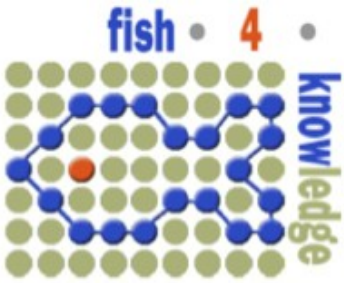
6. Movable cameras

- They would move static cameras on several places, for example by leaving the camera in a location for one day and moving it in another location it each day.
- By doing so, they could also get measurements at different locations of a area of study.
- Covering large area with such a method would allow them to follow migrations.

7. Derived queries

We derived from this interview that Dr. Day and his colleagues would need the F4K tool to answer the same queries as those mentioned by Prof. Shao, plus the following queries:

- What is the average body size for *species X?* How many percent of fish are *small*, *normal* or *big*?
- What is the number of fish in *area X* for indicative species related to pollution?
- What is the distribution and number of fish for indicative species of *factor X*?
- What is the analysis of *factor X* impact, using *pattern of indicative data Y*?
- What are the areas and periods of time of *species X* migrations?
- What are the areas and periods of time of *species X* SPAGS?
- What are the SPAGS periods in *area Y*?



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix III
Report on the interview of Prof. Stergiou

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document reports essential information exchanged with Prof. Stergiou in the context of our user study.

In March and April 2011, we conducted a user study to elicit user information needs related to the data analysis tasks we can support w.r.t. the data we can collect from videos. This work is discussed in the Deliverable 2.1 "User information needs".

This document reports essential information exchanged with Prof. Stergiou during a 20 minute phone call on April 12th 2011. The interview was short, and additional information collected from Prof. Stergiou publications & websites are added in the introduction chapter.

Prof. Stergiou has not yet envisioned to integrate the Fish4Knowledge tool in his research. Therefore, he did not express any typical queries we could add to our "20 Questions" but he gave information about how his research is done without the Fish4Knowledge tool. Prof. Stergiou is interested in the project

potential usages, as a biologist and as a project advisor.

1. Introduction

- Prof. Stergiou and his colleagues study the Greek marine ecosystem at the Aristotle University of Thessaloniki.
- They mainly study life history aspects (e.g., age & fecundity of fish), fisheries and fish biology of the Aegan Sea.
- The Aegan Sea ecosystem is quite different than the Taiwanese & Caribbean ecosystems (studied by the other interviewees - Prof. Shao & Dr. Day).
- For optimisation reason, the Fish4Knowledge tool may only target a subset of detectable the species, instead of comparing a fish contour to all existing species. Therefore a dedicated subset of detectable species must be used for studying the Aegan Sea ecosystem.
- Specific species of coral live in the area they study, but they may live in more deep water along with completely different species & behaviours of fish.
- We assume that even when studying different ecosystems, biologists' data collection methods and data analysis methods are similar.

2. Using the Fish4Knowledge tool

- Prof. Stergiou does not currently envision the usage his team could make of the Fish4Knowledge tool.
- This may also be due to uncertainty about how video analysis tools may be compatible with their existing methods, studied environments and research questions.
- For example, they may wish to study intermediate water rather than shallow water, and this implies technical issues (e.g., less enlightenment, fixation & maintenance of cameras).
- Another example (explained below) is that they need some data that could only be extracted from dead fish (e.g., skeleton or intestine content)

3. Collected data

- To study life history aspects, the most needed data are species, length of body size (indicates fecundity) and age of fish (determined by analysing skeletons, otolith and scales' annuli).
- They evaluate population growth by calculating the frequency of fish length.
- They do not collect data about fish behaviour.
- They generally do not attempt to estimate the size of the overall Aegan Sea populations (number of fish) on the basis of their sampling data. The main reason is that they do not need to research questions. We assume than they only need measurements and trends collected from samples of

fish, without needing to calculate the overall population size. Besides, this estimation is difficult and uncertain, and can be done using ecological models.

4. Sampling methods

- They collect their data from dead fish.
- They collect samples of fish from experimental fishing, using small boats or trawlers, and sometimes from fish markets.
- Experimental fisheries are conducted using a random stratified sampling method[1]. We assume that strata are the different locations where samples are fished. Otherwise, it might be the different layers in the water column (i.e., fishing at several depths).

5. Species recognition

- Some species are difficult to identify, mainly because some exotic and thus unexpected species have recently immigrate from the Red Sea through the Suez Canal (note: the Red Sea ecosystem is similar to the Caribbean & Taiwanese ecosystems).
- They study pelagic fish, especially mesopelagic fish (i.e. living in depth between 200 & 700m). The main species are: anchovies, sardine, hake and mullidae fish (e.g., mullet).

6. Current data analysis tools

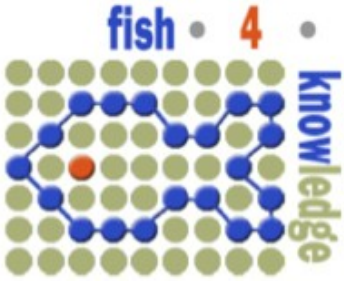
- They are using classical statistic tools such as the Food and Agriculture Organisation (FAO) tools[2].
- They do not encounter any particular issues using those tools.

7. Implications

- This interview gave insight in current biologists' practices, without using the Fish4Knowledge tool.
- These insights help to understand and define the user tasks, to envision how the Fish4Knowledge tool could exist in the biologists' context.
- We need to continue to investigate biologists' practices and explore their methods, techniques & tools implied in data collection, data analysis and data visualisation.

[1] http://oregonstate.edu/instruct/bot440/wilsomar/Content/MoreDesigns.htm#StRS

[2] http://www.fao.org/fishery/statistics/software/en



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix IV
Existing data collection and analysis methods

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document describes how biologists currently collect and analyse data, without being provided with a video analysis tool.

In March and April 2011, we conducted a user study to elicit user information needs related to the data analysis tasks we can support, and to the data we can collect from video. This work is discussed in the Deliverable 2.1 "User information needs".

We briefly describe here existing data collection and analysis practices. These insights are meant to support our understanding of:

- Sampling methods (section 1.1.)
- Techniques used to collect fish (section 1.2.)
- Data that can be collected (section 1.3.)
- Generation of relevant measurements from the collected data (sections 2.1. and 2.2.)
- Interpretation of measurements (section 2.3.)
- Compatibility of video data and manually collected data (section 2.4.)

1. Existing data collection methods

This chapter describes how biologists manually collect data. It is meant to help us understand how they would consider data that are automatically extracted from videos w.r.t. the way they manually collect data.

We consider two aspects of data collection practices: sampling methods (e.g., to select locations where samples are collected) and data collection techniques (e.g., fishing or diving in the selected locations).

1.1. Sampling methods

Biologists use sampling methods to select the locations where they collect or observe fish. Among the variety of existing sampling methods used, we explored basic concepts related to the random stratified sampling method that was mentioned by Prof. Stergiou. The definitions given below were collected from the website of the Oregon State University[1].

Simple random sampling: the most common sampling method, it consists of randomly selecting items while each item is equally likely to be selected. For instance, one can number each location, put the numbers in a box, and pick one. Usually, biologists use programs that generate random number.

Sampling without replacement: each item can only be selected once. For instance, once a number is picked up from the box, it is not put back in the box and won't get picked up next time.

Sampling with replacement: each item can be selected more than once. For instance, fishes that are observed while diving could be observed several times. Video data are collected with replacement, as a fish can be filmed several times.

Using quadrats and grid or coordinate systems: a quadrat is a square surface virtually drawn on the seabed in which fish samples are collected. Considering the whole area of study, quadrats must be placed in several randomly selected locations. There are 2 ways of using the simple random sampling method to select locations:

1. Coordinate system: use the simple random sampling method to select coordinates of points, and centre the quadrats on these points.
   In this case, quadrats may overlap (sampling with replacement). To get a sampling without replacement, it is accepted to replace an overlapping quadrat by another randomly selected one.
2. Grid system: draw a uniform grid by dividing the studied area into non-overlapping quadrat-sized rectangles, number the quadrats of the grid, and randomly select numbers of quadrats to use.

[1] http://oregonstate.edu/instruct/bot440/wilsomar/Content/MoreDesigns.htm#StRS

If the area of study is not rectangular, these method may select quadrats that fall outside of the studied area. In this case, these quadrats are usually discarded and replaced by other randomly selected ones.

Using line-intercept and coordinate system: instead of collecting samples inside a quadrat, biologists can draw a line and collect fish that crosses the line. A coordinate system and the simple random sampling method can be used to place the lines in 2 steps:

1. Randomly select coordinates of a point to centre the line on.
2. Randomly select an angle between 0 and 180 to define the direction of the line.

Random stratified sampling: consist of applying a simple random sampling to each subpart (called strata) of the universe of interest. Depending on the studies, biologists might consider as strata either different layers of the water column, different sub-areas on the seabed (e.g., boundaries between specific habitats), or different species that live in an area.
The number of samples collected in each stratum depends on the proportion of the strata. For instance, considering the sub-area *A* of $1m^2$, and the sub-area *B* of $2m^2$, the number of samples collect in *A* will be half the number of samples collected in *B*.
The data collected for each stratum can also be weighted to reflect the proportion of each stratum. For instance, to calculate the average density of fish *D* considering the sub-area *A* of $1m^2$ with a density D_A, and the sub-area *B* of $2m^2$ with a density D_B, we would use the formula $D = (D_A + 2*D_B) / 3$.

1.2. Techniques to collect data

To collect samples, biologists can access fish by either observing or catching fish through the following data collection techniques.

Fish market: as mentioned by Prof. Stergiou and Prof. Shao, biologists collect dead fishes from commercial fishery. We do not know what methods are used to select fishes and fisheries (e.g., which seller, which boat).

Experimental fishing: researchers fish exclusively for research purposes, using small boats or trawlers.

Diving observations: biologists observe fish in their natural environment. Contrary to the above techniques, they can observe fish behaviours and more of the underrepresented species (e.g., cryptic or hiding fishes).

1.3. Collected data

The types of data that can be collected do not depend on the sampling method used to select sampling locations, but on the chosen technique to collect fish (i.e. fishing for dead or alive fish, or diving to observe fish). For instance, fishing do not allow to collect data regarding fish behaviour, and only samples

of dead fish allow accurate measurement of:

- Fish age, by measuring body size, otolith size (inner bone located near fish eyes), or scales' annuli size.
- Fish fertility.
- Fish diet, by analysing the content of the intestines.
- Chemicals concentration in fish organisms (e.g., to study pollution).
- Species recognition, as sometimes only a microscopic analysis can differentiate similar species.

The following table resumes the types of data that can be collected w.r.t. the data collection techniques.

	Dead fish	Living fish		
	Caught by fishing	Caught by fishing	Observed by diving	Observed by cameras
Taxonomic classification	+++	++	++	+
Behavioural data	-	-	++	+
Fish body size	+++	+++	+	-
Fish age	+++ *(derived from fish dissection)*	++ *(derived from fish size)*	+	-
Fish fertility	+++ *(derived from fish dissection)*	++ *(derived from fish size)*	+	-
Fish diet	+++	-	++	+
Chemicals in fish organisms	+++	+	-	-
Data regarding cryptic or hiding fishes	-	-	++	-

Legend

- *- : this type of data can not be collected*
- *+ : this type of data can partially be collected*
- *++ : this type of data can be collected*
- *+++ : this type of data can be very precisely collected*

Table 1 - Types of data that can be collected w.r.t. data collection technique

2. Existing data analysis methods

This chapter describes how biologists currently analyse collected data and is meant help us understand how they deal with metrics calculations and statistics, and with data collected from different sampling methods and techniques.

2.1. Deriving biological measurements, patterns and statistics

Depending on their topics of interest, biologists can use a variety of measurements such as those mentioned in the Deliverable 2.1. "User information needs", in section 4.1. Sets of measurements (e.g., patterns of data) can give them insights into particular aspects, for instance to detect an environmental event (e.g., the low abundance of these species means that there is a pollution).

From a publication of Prof. Shao[1], we derived that biologists do not take into account every collected data. They do not use data that were collected from a too few fishes. For instance, if a species was observed less than 5 times, then it is not included in the count of species.

We identified basic metrics, but we still have to explore in-depth:

- Calculation methods to derive significant measurements, including mitigation of collected data (e.g., variability, threshold to take into account observations that were collected a significant number of times).
- Sets of measurements and significant patterns of data.
- Methods to study trends and evolutions overtime (e.g., ANOVA...).

2.2. Data analysis software

Prof. Stergiou mentioned that his team is using the FAO softwares[2]. and "*standard data analysis tools*", without encountering particular issues while using these tools.

We assume that biologists work with spreadsheets and frameworks such as Matlab, and we still have to study in-depth the data analysis softwares they use.

2.3. Analysing measurements calculated from samples

Biologists generally do not attempt to estimate data for the overall populations on the basis of their sampling data (e.g., overall number of every clown fish living in a specific area). They usually do not need that estimation for their research. However, it can be done using ecological models but it is difficult and

[1] *Effects of habitat modification on coastal fish assemblages* – Journal of Fish Biology - 2010

[2] http://www.fao.org/fishery/statistics/software/en)

uncertain.

We assume that biologists only need to analyse measurements calculated from samples of fish and their trends over time, without needing to estimate those values for the overall population.

2.4. Analysing data from different sources

Merging data collected by using different sampling methods and data collection techniques are likely to create biases, particularly regarding:

Underrepresented species: each data collection techniques are more or less accurate to detect specific species (e.g., cryptic, hiding or nocturnal species).

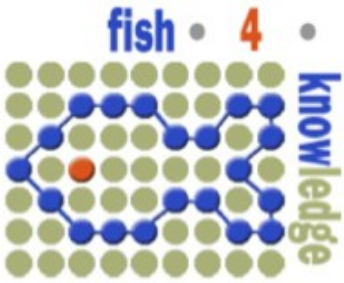
Sampling with or without replacement: some sampling process allow to take one fish into account several times. Diving and video analysis techniques allow to observe the same fish into the same sample (sampling *with replacement*). This is not the case with dead fish sampling (*without replacement*).

It is unsure that biologists could use both video analysis and manually collected data in the same analysis. We assume that they conduct data analyses by using data collected with the same sampling method and data collection technique. However, we assume that the results of such analyses can be compared, and that biologists can particularly compare trends over time rather than absolute values of measurements.

3. Conclusion

We derived that video analysis samples fish with replacement (i.e. one fish can be taken into account several times), and that it is similar to diving techniques except that cryptic species are still underrepresented. Locations of cameras would be selected using a stratified random sampling method.

Regarding data that can not be accurately collected using video analysis (e.g., fish size, nocturnal and cryptic fish), biologists may not merge video data and manually collected data, but they may compare results from video data analyses with results from analyses conducted with other data collection methods.



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix V
Unsupported tasks

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document describes biology research topics that are not in the scope of our project, but that could be studied through video analysis if provided with specific hardware or software resources.

In March and April 2011, we conducted a user study to elicit user information needs from which we derived biologists data analysis tasks and topics of interest. We elicited 3 tasks we can support: 2 data analysis tasks (population dynamics and impact of environmental events), and 1 video search task. We then derived the related information needs. This work is discussed in the Deliverable 2.1 "User information needs".

We describe here other biology topics that we can not currently support. These insights were also collected during the user study. They help to understand biologists' interests and could also be used for future work or other projects.

In this document, we will discuss data analysis tasks, information needs and feasibility issues of the following topics:

- Trophic systems (section 1)
- Fish reproduction (section 2)

- Migration (section 3)

Section 4 will finally resumes interesting insights collected about biologists needs for additional resources.

1. Trophic systems

This topic concerns the study of food chains and feeding interactions between fishes and other organisms. Biologists are interested in the diet of fishes and their feeding behaviours and strategies. They also study how a habitat's topography (e.g., type of holes for fish to hide, rocks for coral to grow) impacts feeding behaviours and thus species composition.

1.1. Trophic behaviours

Video analysis allows to detect feeding and predator-prey behaviours of various species. But we have to be aware that some behaviours may not be recognizable, especially those involving camouflage or hiding strategies, gestures of fins and tails, or toxic defence features (e.g., venomous fins).

Interesting trophic behaviours are:

- **Feeding**: a fish is observed while eating.
- **Predating**: a fish is observed while preying on another organism.
- **Preying**: a fish is observed while being preyed on by another organism.
- **Schooling**: a group of fish is observed while moving in coordinated movements.
  For some species that usually do not live in group or live a shoal (i.e. a group with no coordinated movements), schooling means that they are being preyed on.

We can derive the frequency of trophic behaviours occurrence depending on a timeframe and/or a location, and thus estimate **feeding periods** and **feeding areas** that are significant for a species.

1.2. Predator-prey relationships

From the observed trophic behaviours and the recognised species involved, we can derive predator-prey interactions between species, and supply the following information:

- **List of preys** observed for a given species.
- **List of predators** observed for a given species.
- **Trophic levels** to classify the various species depending on their diet (i.e. "who eats whom"). It could be the overall trophic levels (e.g., including algae, plankton), or simplified trophic levels (e.g., including

only fish).

To study fish diet, biologists usually analyse the content of fish intestine. This provide more accurate data than the detection of trophic behaviour through video analysis.

1.3. Discussion on feasibility

We excluded this topic because we are not certain that behaviour recognition would be able to recognise the complete set of trophic behaviours for every species involved (e.g., behaviours of cryptic species may not be recognisable).

For instance, the analysis of predator-prey relationship would be incomplete and we do not know if a partial list of predators would be useful for biologists.

We also assume that biologists would need to correlate those data with their existing knowledge about fish diet and habitats (e.g., role of coral species) and descriptions of trophic behaviours that video analysis can not recognise (e.g., camouflage, hiding strategies...). Currently our tool does not include data describing existing domain knowledge.

2. Fish reproduction

Biologists are interested in studying reproduction periods, habits and behaviours (e.g., mating, spawning, nursing), birth rate and fertility potential (i.e. number of eggs that may be produced). Recruitment is also an interesting part of the reproduction process. Also called settle down time, it means the period when young fishes are added to the overall population.

As mentioned by Dr. Day, biologists are also interested in identifying spawning aggregation sites (**SPAGS**) where fish regularly gather to reproduce at the same period.

2.1. Reproduction behaviours

Video analysis can detect reproduction behaviours such as mating, spawning, and nursing. But we must be aware that some behaviours may not be recognizable, especially those involving complex gestures of fins and tails. From behaviour detection, we can derive **reproduction periods and locations**.

Schooling behaviour may also be of interest. For some species, it implies that fishes are reproducing, mating or spawning, or evolving in a group of young fish (recruitment), especially if they usually are solitary fish.

2.2. Population growth

Video analysis can also provide measurements to evaluate the evolution of population size (i.e., growth rate). If there is a population growth in specific

locations, along with an overall population growth all locations together, then we can derive that population growth is due to reproduction. In the case where there is not a global population growth all locations together, but a rather constant overall population, then migration phenomena would be a more consistent explanation.

We can correlate growth rate and reproduction behaviour occurrences to derive that a population size is increasing due to reproduction phenomenon. This might allow to evaluate **birth rate**, and **recruitment periods and locations**.

But these evaluations would cary uncertainty as it would suppose that populations contain young fish. The evaluation of fish body size, which we can not support, would allow to evaluate the presence of young fish, and the study of recruitment process would be more reliable.

2.3. Discussion on feasibility

The study of fish reproduction is subjected to the following feasibility issues:

- Some reproduction behaviours might not be recognisable.
- The study of reproduction locations and SPAGS might need to cover a wide area with more cameras than we have, or with cameras that can be periodically moved to collect data in various locations over the studied area.
- The study of recruitment caries a high level of uncertainty due to the impossibility to detect fish body size, and thus to evaluate the proportion of young fish in a population.
- Fish fertility can not be evaluated as it also needs the evaluation of fish body size.

To evaluate fish body size, several technical solutions can be envisioned (e.g., stereoscopic cameras, infrared detectors), but none of them are in the scope of our project.

3. Fish migration

Biologists are interested in estimating migration purposes (reproduction, feeding or environmental reasons), periods and paths, as well as immigration and emigration rates.

3.1. Periods and path of migration

a) Ranges of period and path

Migrations can occur on a daily basis, or on longer periods on a seasonal or yearly basis. Daily migrations between shallow and deeper waters, called diel

vertical migration, is a common behaviour for many species and is often split between night and day for diurnal and nocturnal species.

Migration length can vary between a few meters to many kilometres, and migration paths can be horizontal (i.e. between areas) or vertical (i.e. between shallow and deeper waters).

b) Detection of migration period and path

Our tool can count the number of fish in various timeframe and locations. We could correlate variations of populations in specific locations with variations in the overall populations size all locations together. If the overall population size is rather constant, then we could assume that variations in population size in specific locations are due to a migration. Thus, we could derive **migration periods and paths** along with **emigration and immigration rates**.

These evaluations need a consistent sampling method to collect videos in a sufficient number of locations over the area of study. We would need either:

- Movable cameras operating in fixed locations for a fixed period.
- Static cameras in sufficient number and placed at representative locations.
- Correlation between video data and manual observations.

For some species, another way to detect migration could consist of detecting their schooling behaviours. For specific species it implies that fishes are migrating, as they usually are solitary or in a shoal (i.e. in a group of fish with non coordinated movement). But these schooling behaviours may not be recognisable as they may not occur near the reefs, i.e. close enough to cameras.

3.2. Purposes of migration

Many species of fish migrate, mostly for feeding and reproduction purposes. Some feed and reproduce in different locations because they cannot differentiate their own offspring from their food. Some migrations are (still) unexplained. Environmental factors such as currents, water temperature, pollution or oxygenation can influence migrations.

Biologists could derive the purpose of migrations by correlating detected migrations with detected reproduction and trophic behaviours, or with environmental data.

3.3. Discussion on feasibility

We excluded this topic because the areas of study are not currently covered by a sufficient number of cameras, and thus the detection of migrations would be highly uncertain.

For instance, considering one population in one location, a decrease in

population size could be explain by either mortality or migration, and an increase could be explained by either reproduction or migration. Without consistently evaluating the overall population size with a sufficient camera coverage, we can not doubtlessly differentiate migration, mortality and reproduction phenomenon.

Further, nocturnal migrations are impossible to detect without night vision cameras.

We also assume that biologists would wish to correlate video analysis data with their existing knowledge of migration phenomenon (e.g., it is known that species X migrate to the north in June for feeding purposes). Currently our tool does not include data describing existing domain knowledge.

4. Conclusion

We identified a number of requirements for biologists to be able to study fish migration, reproduction and trophic system. We identified the following needed additional resources:

- **Extensive camera coverage**: provide a sufficient number of cameras to cover the area of study.
- **Night vision**: provide cameras that can record videos during night time.
- **Fish body size**: provide the size of fishes and derive their age and their fertility.
- **Extensive species and behaviour recognition**: provide the accurate detection of a consistent set of behaviours for a targeted set of species (e.g., the accurate recognition all reproduction behaviours of species X). By the end of our project, we aim at being able to detect a wide range of species and behaviours. Currently we are not sure if we will be able to recognise the complete sets of behaviours needed to study reproduction or trophic system.
- **Domain knowledge**: provide a database containing prior biology knowledge that can improve automated video analysis, or that can help biologists to interpret video analysis data.

Some of these additional resources could also improve the study of the data analyses included in the scope of our project (population dynamics and impact of environmental events). The table below resumes the impact of unsupported resources on data analysis tasks.

	Population dynamics	Environmental events	Trophic systems	Reproduction	Migration
Extensive camera coverage	+	+	+	++	**+++**
Night vision	++	+	+	+	++
Fish body size		+		**+++**	++
Extensive species and behaviour recognition		+	**+++**	++	+
Domain knowledge	+	+	++	+	+

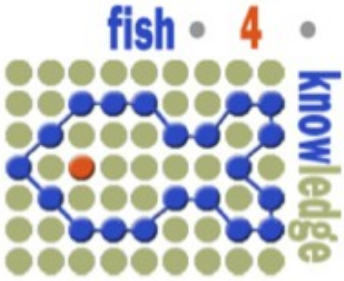
Legend

+ : the resource would generally improve the data analysis task

++ : the resource would dramatically improve the data analysis task

+++ : the resource is required for the data analysis task

Table 1 - Impact of unsupported resources on data analysis tasks



Fish4Knowledge Deliverable D2.1

User Information Needs

Appendix VI
Basic user tasks and functionalities

Principal Authors: E. Beauxis-Aussalet (CWI), L. Hardman (CWI), J. van Ossenbruggen (CWI)
Reviewers: F.P. Lin (NCHC)
Dissemination: PU

Abstract: This document specifies basic data manipulation functionalities provided for users to query meaningful information that implies data selection or computation.

In March and April 2011, we conducted a user study to elicit user information needs from which we derived biologists data analysis tasks and topics of interest. We elicited 3 tasks we can support: 2 data analysis tasks (population dynamics and impact of environmental events), and 1 video search task. We then derived the related information needs. This work is discussed in the Deliverable 2.1 "User information needs".

In this document, we specify basic functionalities provided for users to access information they need. These functionalities provide basic data manipulation that allow users to query for meaningful information that implies data selection or computation. They facilitate interactions between the database (low-level queries) and the user interface. They support flexible data analysis by allowing users to gather specific information of their interest.

After specifying these functionalities (sections 1 to 4), we gives examples of user queries, based on the most important queries (see Deliverable 2.1, Table 1, p.10).

1. Introduction

The functionalities supporting each user task are described as functions with inputs (i.e. parameters), and outputs. We give below a few explanations about special types of inputs.

Inputs indicated in *italic* are not mandatory, inputs following the character * can be set up with a cardinality 0-n.

Time units can be chosen among 10 minutes, hour, day, week, month or year (e.g., to calculate the number of fish detected each day).

Users may wish to look for fishes from a specific species as well as all fishes from a genus or a family. They may even wish to look for fish whose species, genus or family is unrecognised. Thus, users can define a **Taxon** as an input, that consist of 3 levels of taxonomic classification:

1. **Family**. The value is either:
   - A specific family,
   - All families (i.e. all fish),
   - Unrecognised families (i.e. all unrecognised fish)
2. **Genus**. If a specific family is selected, the value is either:
   - A specific genus from the selected family,
   - All genus (i.e., all fish from the family),
   - Unrecognised genus (i.e., fish from the family that genus is unrecognised)
3. **Species**. If a specific genus is selected, the value is either:
   - A species from the selected genus
   - All species
   - Unrecognised species

Users might want to take into account only fishes that were detected a sufficient number of time, as very sparse species might not be representative of the studied ecosystem. They can define the minimum number of occurrence of a fish taxon to take into account, as an input called *MinOccurence*.

2. Video search

Functionality	Input	Default input	Output	Priority
SearchVideo()	Free text query or *Timeframe,* *Location,* *Event,* ** Taxon,* ** Behaviour*	Possibility to input the timeframe and location of an event. Several taxon and behaviours may be combined (AND/OR).	List of video excerpts, along with the recognised fishes, species, genus, families and behaviours.	Low

TagVideo()	Video, * Fish, * Taxon, * Behaviour	*Biologists manually indicate fishes, species, genus, families or behaviours that have not been recognised by video analysis.*	Low
RemoveVideoTag()	Video, * Fish, * Taxon, * Behaviour	*Biologists manually remove fishes, species, genus, families or behaviours that have not been correctly recognised by video analysis (false positive).*	Low

3. Population dynamics

This table below contains the basic metrics to study populations dynamics:

Functionality	**Input**	**Default input**	**Output**	**Priority**
OverallAbundance()	Timeframe, Location, *TimeUnit,* *MinOccurrence*	TimeUnit = Timeframe MinOccurence = 1	Number of fish counted in each time unit. Species are taken into account if a minimum number of occurrences is reached.	High
Abundance()	Timeframe, Location, Taxon, *TimeUnit*	TimeUnit = Timeframe	Number of fish from the Taxon counted in each time unit.	High
RelativeAbundance()	Timeframe, Location, Taxon, *TimeUnit*	TimeUnit = Timeframe	Percentage of fish from the Taxon calculated for each time unit.	High
AbundanceLevel()	Timeframe, Location, Taxon, *TimeUnit*	TimeUnit = Timeframe	Range of value in which the relative abundance of a Taxon belong. Ranges are Abundant, Common, Occasional and Rare. Ranges are evaluated for each time unit.	High
DistributionInAbundanceLevels()	Timeframe, Location, *TimeUnit*	TimeUnit = Timeframe	Count and percentage of species belonging to each abundance levels (Abundant, Common, Occasional and Rare).	High
SpeciesRichness()	Timeframe, Location, *TimeUnit,* *MinOccurrence*	TimeUnit = Timeframe MinOccurence = 1	Number of species counted in each time unit. Species are taken into account if a minimum number of occurrences	High

			is reached.	
SpeciesComposition()	Timeframe, Location, *TimeUnit, MinOccurrence*	TimeUnit = Timeframe MinOccurence = 1	For each time unit: OverallAbundance(), SpeciesRichness(), and RelativeAbundance() for each retrieved species. Species are taken into account if a minimum number of occurrences is reached.	High

This table below contains metrics used to study evolutions of populations overtime:

Functionality	**Input**	**Output**	**Priority**
OverallGrowthRate()	Timeframe, Location, TimeUnit, *MinOccurrence*	Positive or negative increase percentage of number of fish counted in each time unit. Species are taken into account if a minimum number of occurrences is reached.	High
GrowthRate()	Timeframe, Location, Taxon, TimeUnit	Positive or negative increase percentage of number of fish from the Taxon counted in each time unit.	High
SpeciesRichnessEvolution()	Timeframe, Location, TimeUnit, *MinOccurrence*	Positive or negative increase percentage of number of species counted in each time unit. Species are taken into account if a minimum number of occurrences is reached.	High
SpeciesCompositionChange()	Timeframe, Location, TimeUnit, *MinOccurrence*	For each time unit: OverallGrowthRate(), SpeciesRichnessEvolution(), and GrowthRate() for each retrieved species. Species are taken into account if a minimum number of occurrences is reached.	High

This table contains metrics used to study behaviour occurrences:

Functionality	**Input**	**Default input**	**Output**	**Priority**
BehaviourOccurrence()	Timeframe,	TimeUnit =	Number of behaviour	Low

	Location, Behaviour, ** Taxon,* *TimeUnit*	Timeframe	occurrences counted for each time unit. If multiple Taxon are indicated, it means that fish from different species are implied in the behaviour. The definition of roles in behaviour might be refined later.	
BehaviourOccurrenceEvolution()	Timeframe, Location, Behaviour, ** Taxon,* TimeUnit		Positive or negative increase percentage of number of behaviour occurrences counted in each time unit.	Low

Note: behaviours could concerns two species, each species having a specific role (e.g., preying, territorial behaviour). When the specification of behaviour detection will be finished, these functionalities may be refined to include specific roles of species (e.g., count preying behaviours where species X is the predator and species Y is being preyed on).

4. Impact of events

Functionality	**Input**	**Default input**	**Output**	**Priority**
EventInfo()	Free text or Timeframe, Location	Free text is meant to match names or any characteristic of events	List of events along with characteristics of events (e.g., type, timeframe, location, description).	High
EnterEvent()	Timeframe, Location, Name, * Type, Description		Manual definition of an event.	High
WeatherInfo()	Timeframe, Location, WeatherDataType, *TimeUnit* or Event, WeatherDataType, *TimeUnit*	TimeUnit = Timeframe if an Event is entered as the input, then Timeframe = Event'sTimeframe, Location = Event's Location	Weather data from external databases and for each time unit. External databases are chosen depending on the needed WeatherDataType.	Low
EnvmtImpact()	Timeframe, Location,	Metrics are any functionality defined	Metrics calculated for the selected time	High

	WeatherDataType, * Metrics, *TimeUnit* or Event, TimeframeBefore, TimeframeAfter, TimeUnit, * Metrics	in section 3. Inputted Timeframe, Location (eventually this of the event) and TimeUnit will be used to calculate the Metrics. Users must also indicate other inputs needed for the selected metrics (e.g., Taxon, Behaviour, MinOccurrence).	frames (before and after event) and time unit.	

5. Example of user queries

We collected the most important queries our system would be used for (see Deliverable 2.1, Table 1, p.10). Here we reformulate these queries expressed in natural language into queries expressed using the system functionalities as specified above.

For each query, the table below gives the original user text, a comment on its interpretation, and its reformulation using system functionalities.

Query ID	**Original text** ***Comment*** **Reformulation**
Q1	`How many species appear and their abundance and body size in day and night including sunrise and sunset period.` *Night period and body size can not be studied.* SpeciesComposition(Timeframe=T, Location=L, TimeUnit=Hour)
Q2	`How many species appear and their abundance and body size in certain period of time (day, week, month, season or year). Species composition change within one period.` *Body size can not be studied. The example below is for a summer season.* SpeciesComposition(Timeframe=June2011-Sept2011, Location=all) SpeciesComposition(Timeframe=June2011-Sept2011, Location=all, TimeUnit=week) SpeciesCompositionChange(Timeframe=June2011-Sept2011, Location=all, TimeUnit=week)
Q3	`Give the rank of above species, i.e., list them according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.` *The list of species ordered by abundance is already given by the Species Composition queried above.* DistributionInAbundanceLevel(Timeframe=June2011-Sept2011, Location=all)

Q4	Fish colour pattern change and fish behaviour in the night for diurnal fish and in daytime for nocturnal fishes.
	A colour pattern change is due to the occurrence of a specific behaviour. We consider that users want to count the occurrences of a specific behaviour. *Night period can not be studied. The user interface could facilitate the selection of all nocturnal species, if a boolean property "nocturnal=true||false" is associated to species IDs.* *The query below can be repeated for each nocturnal species.*
	BehaviourOccurrence(Timeframe=X, Location=Y, Behaviour=Z, Taxon=nocturnal species)
Q5	Fish activity within one day (24 hours).
	We consider that fish activity is analysed using a tailored set of measurements, depending of each users interest. Below is a default set of measurements.
	SpeciesComposition(Timeframe=today, Location=all, TimeUnit=Hour) SpeciesCompositionChange(Timeframe=yesterday-today, Location=all, TimeUnit=day) BehaviourOccurrence(Timeframe=today, Location=all, TimeUnit=Hour, Behaviour=feeding) BehaviourOccurrenceEvolution(Timeframe=yesterday-today, Location=all, TimeUnit=day, Behaviour=feeding) BehaviourOccurrence(Timeframe=today, Location=all, TimeUnit=Hour, Behaviour=preying) BehaviourOccurrenceEvolution(Timeframe=yesterday-today, Location=all, TimeUnit=day, Behaviour=preying)
Q6	Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behaviour of various species.
	The query below can be performed for each behaviour and species of interest.
	BehaviourOccurrence(Timeframe=X, Location=Y, Taxon=T, Behaviour=B)
Q7	Growth rate of certain species for a certain colony or group of observed fishes.
	We consider that "colony or group of observed fish" concerns a specific location which fishes are studied. For instance, it can be Lanyu Island. The user interface could facilitate the selection of all camera of a specific area of interest, if areas of interest are defined and associated the targeted camera. *The query below can be performed for each species of interest.*
	GrowthRate(Timeframe=this month, TimeUnit=day, Location=Lanyu, Species=X)
Q8	Population size change for certain species within a single period of time.
	The example is given for a period of time set to March 2011.
	GrowthRate(Timeframe=March2011, Location=X, Species=Y) Abundance(Timeframe=March2011, TimeUnit=day, Location=X, Species=Y)
Q9	The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.
	Users will simultaneously visualise the two queries above (Q8) with the query below.

	WeatherInfo(Timeframe=March2011, TimeUnit=day, Location=X, WeatherDataType= all)
Q10	Immigration or emigration rate of one group of fish inside one monitoring station or one coral head.
	Immigration and emigration rates can not be studied. But users can study the evolution of counts of fish, for instance with the query below.
	OverallGrowthRate(Timeframe=2011, TimeUnit=month, Location=X)
Q11	Solitary, pairing or schooling behaviour of fishes.
	Same as Q6
	Same as Q6
Q12	Settle down time or recruitment season, body size and abundance for various fish.
	Reproduction, including settle down time or recruitment season, can not be precisely studied. Neither can body size be studied. Only abundance can be studied.
	Abundance(Timeframe=X, Location=Y, Species=Z)
Q13	In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic character to distinguish some similar or sibling species.
	This concerns specific trust aspects. Certainty scores of species detection are calculated every time a fish species is detected. These scores will be accessible to users. The specification of these scores is not finished yet. Thus, the way users would analyse certainty scores will be investigated in detail later on. *Similar remark stand for the "diagnostic character to distinguish species", as their specification is not completely defined yet.*
Q14	Association among different fish species or fish-invertebrates.
	"Association among species" is a synonym of "species composition". Biologists are not interested in analysing which fish co-occur simultaneously. They are rather interested in analysing species that live in the same area. Thus they would use the first query below. *In case they want to watch videos where two species co-occurred, they would use the second or third queries below.* *And if they want to study interactions in terms of behaviour, they would use queries such as the fourth one.*
	1. SpeciesComposition(Timeframe=X, Location=Y) 2. SearchVideo(Taxon={T1,T2}) 3. SearchVideo(Taxon={T1,T2}, Behaviour=B) 4. BehaviourOccurrence(Timeframe=X, Location=Y, Behaviour=B, Taxon={T1,T2})
Q15	Short term, mid-term or long term fish assemblage fluctuation at one monitoring station or comparison between experimental and control (MPA) station.
	"Fish assemblage fluctuation" is a synonym of "species composition". Comparison can be done only between locations that are equipped with cameras.
	SpeciesComposition(Timeframe=X, Location=L1)

	SpeciesComposition(Timeframe=X, Location=L2)
Q16	Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.
	Comparison with other data collection and analysis techniques are not in the scope of the user interface.
Q17	The difference of using different camera lens and different angle width.
	A proper sampling and data analysis protocol must be set up to proceed to such comparison.
Q18	Is it possible to do the same monitoring in the evening time.
	This concerns the certainty (i.e., error rate) of video analysis done in evening time. As for Q13, certainty score specification is not finished yet, so this query will be investigated later.
Q19	How to clean the lens and solve the biofouling problem.
	This is not in the scope of the user interface. We do not know yet if the automatic evaluation of lens cleanliness will be implemented.
Q20	Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.
	This is not in the scope of the user interface.
Q21	What is the average body size for *species X?* How many percent of fish are *small*, *normal* or *big*?
	Currently, body can not be studied.
Q22	What is the number of fish in *area X* for indicative species related to pollution?
	The user interface could facilitate the selection of specific indicative species, if a dedicated property is associated to species IDs (e.g., indicatePollution=true||false).
	Abundance(Timeframe=X, Location=Y, Taxon=list of indicative species)
Q23	What is the distribution and number of fish for indicative species of *factor X*?
	The user interface could facilitate the selection of specific indicative species, if a dedicated property is associated to species (e.g., indicateFactorX=true||false). The "distribution of fish" concerns the location where fish populations are settled. Thus the query below can be performed for each location of the area of interest.
	Abundance(Timeframe=T, Location=L1, Taxon=list of indicative species)
Q24	What is the analysis of *factor X* impact, using *pattern of indicative data Y*?

	A "pattern of indicative data" is a set of measurements dedicated to the analysis of specific environmental conditions. The user interface could facilitate the selection of these sets of measurements, if a dedicated property is associated to measurements (e.g., indicateFactorX=true||false). The same remark stands for indicative species or behaviours that might be used to calculated metrics.
	EnvmtImpact(Timeframe=T, Location=L1, Metrics=list of indicative metrics, *Taxon=list of indicative species, Behaviours=list of indicative behaviours*)
Q25	`What are the areas and periods of time of` *`species X`* `migrations?`
	Migrations can not be precisely studied using our tool. But users can study the evolution of species abundance in various locations and periods of time using the functionalities below.
	Abundance(Timeframe=X, Location=Y, Taxon=T) GrowthRate(Timeframe=X, Location=Y, Taxon=T)
Q26	`What are the areas and periods of time of` *`species X`* `SPAGS`[1]`?`
	Reproduction can not be precisely studied using our tool. But users can study the the occurrence of spawning behaviours in various locations and periods of time by using the functionalities below.
	BehaviourOccurrence(Timeframe=X, Location=Y, Taxon=T, Behaviour=spawning)
Q27	`What are the SPAGS`[1] `periods in` *`area Y`*`?`
	Reproduction can not be precisely studied using our tool. But users can study the the occurrence of spawning behaviours in the locations of interest, and over various periods of time.
	BehaviourOccurrence(Timeframe=X, Location=L, Behaviour=spawning)

Table 1 - The most important queries envisioned by potential users, and their formulation using system functionalities

1 SPAGS: SPawning Agregation Sites, where fish gather to reproduce.

