ICES WKSEATEC REPORT 2017

ICES CM 2017/SSGIEOM: 29

REF ACOM AND SCICOM

Report of the Workshop on Technical Development to Support Fisheries Data Collection (WKSEATEC)

12 - 14 September 2017

ICES HQ, Denmark

ICES CIEM
International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2018. Report of the Workshop on Technical Development to Support Fisheries Data Collection (WKSEATEC). ICES WKSEATEC Report 2017, 12 - 14 September 2017. ICES HQ, Denmark. 81 pp. https://doi.org/10.17895/ices.pub.8685

Contents

Executive summary

This workshop, chaired by Dave Stokes (Ireland) and Marcellus Rödiger (Germany), was held at ICES headquarters, Copenhagen, from 12th – 14th September 2017. There were 16 attendees and 2 remote participants representing 12 countries from the Mediterranean, Baltic and Atlantic areas including the US.

WKSEATEC was set up by the ICES Steering Group on Integrated Ecosystem Observation and Monitoring (SSGIEOM) to make recommendations on technical solutions for the collection and quality assurance of fisheries data at sea and in ports. The objectives can be summarised by two complimentary themes; i) to maximise the effectiveness of quality assurance checks in a data collection program greatest focus and support needs to be on the very start of the process, during sampling itself; ii) this can only reasonably be achieved once the data are in a digital format which requires improved uptake in electronic data capture methods.

To address the Terms of Reference the workshop structured the meeting and report into 4 topic sections:

1. Overview – benefits and impediments to paperless sampling: The report highlights the difference between reactively screening data products in Quality Control (QC) to proactive Quality Assurance (QA) where information is fed back during the process of sampling itself. The ready availability of affordable powerful computing and the effectiveness of open source visual and statistical data checks in open source code such as R is widely known. However, there has been less than 50% uptake on IBTS surveys for example, so barriers to wider implementation of paperless sampling during the measuring process are discussed.
2. Review of current approaches: Summaries of the main presentations during the workshop covering the technology that has been either purchased or developed in-house are given in this section. How the technology has been applied, customized and received have been key themes throughout the workshop. Presentations from the group illustrate that whether purchasing off-the-shelf or developing technology in-house significant time and resources can be expended, but effective solutions are very achievable. Smaller scale open source projects have also been successful and are currently being field tested.
3. Data quality control and management: Here approaches to error trapping and feedback to the measuring board user is examined using three case studies presented during the workshop. The important measurement errors likely to occur as well as the efficiency of graphical data display to highlight and aid correcting potential errors is also illustrated. Once collected the ease with which data formats can be understood and exchanged is of great importance to how widely and efficiently they can be utilised to answer questions and guide management advice. Therefore guidance on a range on internationally available and maintained standard reference lists is also given.
4. Collaborative potential: The general applicability of the technology solutions and issues discussed during the workshop is outlined in the final section. Experience from compiling the DATRAS and RDB data exchange formats highlight some general principles. Significant resources have been expended and experienced gained by individual institutes addressing this issue and guidance is given on how best to collaborate and support increased engagement with paperless sampling.

WKSEATEC discussed a broad spectrum of content from user perception and uptake to technical programming details. Two overarching conclusions were reached however: 1) no single factor seemed to impede a move to paperless sampling. A few reasonable concerns combined with uncertainty around what exactly should be expected from investment in costly electronic data capture system can lead to inertia; 2) significant progress in electronic data capture has been made in the Atlantic/Baltic area while progress in the Mediterranean has centred on implementation of a standardized data checking routine across surveys. To maximise the benefits of these and other developments the group proposed an agile approach to supporting development and technology exchange. The first is to broaden the regular static data exchange format approach into a more generic Fisheries Data Language. The second, more ambitious concept is to combine this with an Application Program Interface (API). API's are proven in many other fields, but seen by the workshop as potential 'game changers' in supporting the integration of technology and open source "data tool boxes" for fisheries data collection.

1 Opening of the meeting

The Workshop on Technical Development to Support Fisheries Data Collection (WKSEATEC) met in Copenhagen, 13th–14th September 2017. The chairs were Dave Stokes of the Marine Institute, Ireland and Marcellus Rödiger of the Thünen-Institut, Germany. There were 16 attendees and 2 remote participants covering 12 countries from the Mediterranean, Baltic and Atlantic areas including the US. Contact was received during the meeting from a further country requesting current and future participation. The list of participants is in Annex 1.

2 Adoption of the agenda

The agenda was adopted by the group and given in Annex 2.

The **Workshop on Technical Development to Support Fisheries Data Collection (WKSEATEC)**, will make recommendations on technical solutions for the collection and quality assurance of fisheries data at sea and in ports. The workshop will be co-chaired by Dave Stokes*, Ireland, and Marcellus Rödiger*, Germany, will meet on 12-14 September 2017, at ICES HQ, specifically to:

a) Review data QC utilised by data managers, advisory groups and recommended by relevant projects such as fishPi (MARE2014-19), SDEFQuality (MARE/2014/19Med&BS), FishTrawl (Casciaro et al., 2015) and Rome R (Bitetto et al., 2017). Where relevant, recommend how best these could be implemented efficiently and effectively during sea/port sampling with the use of electronic data capture.

b) Review current and potential electronic data capture solutions applicable to fisheries data collection at sea and in ports. Recommend a shortlist of approaches to reflect the resources and technical support that may be available.

c) Review the applicability of open source and General Public License (GNU/GPL) software to promote the affordability, flexibility and uptake of technical solutions to fisheries data collection. Recommend a structured approach to collaborative maintenance, development and support of an open source system.

3 Overview – benefits and impediments to paperless sampling (DS, MR)

3.1 Background

The benefits of electronic or paperless sampling is often debated among survey scientists, however the uptake generally is limited with the reasons not being entirely clear or consistent. At the ICES International Bottom Trawl Working Group (IBTS) meeting, in Sète, France, 2016, a collaborative session with members of the ICES Mediterranean International Trawl Survey (MEDITS) group took place. This included a presentation from IFREMER on a development project for a new electronic measuring board. It became clear that the majority of countries represented were involved in some form of electronic data capture projects, all for several years with various levels of resources, but all independent and without full implementation at the time.

Discussion about the reasons for limited coordination and uptake of technology in data capture and across member states (MS) centred on a shortlist of themes, around which there was no universal agreement or conclusion. IBTS therefore recommended a workshop, WKSEATEC, be set up with TORs to address these themes and make recommendations for coordination and progress in this area.

Following submission of the recommendations to ICES one of the open source measuring board projects mentioned at the meeting, OpenSMB from the Thünen Institute, presented their work at the Baltic Re-gional Coordination Meeting (RCM) in Rostock, August 2016. Recommendations for a follow up work-shop were also made there. Through ICES contact was proposed between IBTS and the Thünen Institute and a follow up meeting arranged in Q1 2017. This meeting helped exchange ideas, coordinate and ensure interests were aligned then for a joint workshop, WKSEATEC, for all interested parties in late 2017.

3.2 Advantages for data quality management

Substantial resources are expended annually on fisheries data collection to support stock assessment and inform fisheries management decisions. Despite advances in affordable computing and powerful open source statistical software such as R or Python, this technology is not routinely implemented during the critical data collection process itself. Once the sampling window closes the quality management role becomes increasingly one of screening rather than correcting the data.

Furthermore, where data are aggregated under coordinated data collection programs such as the CFP, having different approaches to data checking at various stages in the data collection cycle between participants adds a further layer of complexity to the issue.

As highlighted above, despite a lack of consensus as to why paperless sampling is not more routinely implemented, there was general agreement on two important points:

1. The majority of data checks applied at all stages of the data collection pathway, are simple, effective and quick to apply once the data are in digital format
2. The opportunity to correct or confirm ‘unusual’ records rather than simply remove or accept them reduces exponentially with time from the point of sampling

This second point highlights the critical phase in data quality management of moving from Quality Assurance (QA) to Quality Control (QC) (Fig 3.2.1.). In general, once sampling is complete data are consolidated into a local sampling database and fish and sampler go their separate ways. Thereafter the data progresses to an Institute database likely followed by a national and/or international database such as DATRAS or InterCatch. At each step checks are applied, but the window to engage again with the activity of the sampling 'process' and correct data (QA) is limited. Measuring activity over, quality management becomes largely a screening process of the resulting data 'product' (QC). It cannot be overstated that the QA process of engaging directly with the sampling event offers the most constructive opportunity to trap measurement error and maximise data accuracy.

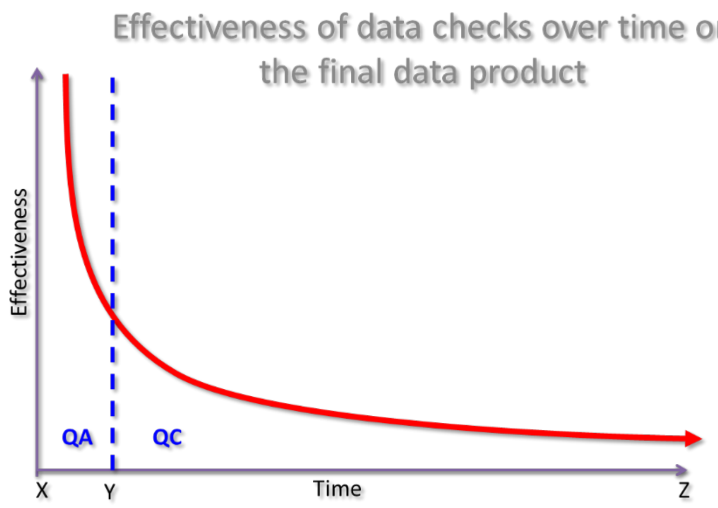


Fig 3.2.1. A generalised schematic showing the data quality management process over time, from sampling (X) to final data product (Z). The effectiveness of data checks to correct data and actively target the Quality Assurance (QA) process dissipates quickly once the sampling process is complete (Y). Thereafter Quality Control (QC) of the data takes over to screen out anything outside agreed standards and control the quality of the final data product.

As an example, a faulty weighing scales that drifts in accuracy (systematic error) or a sampler that forgets to tare a scales for the sampling container (random error) would lead to small, but significant Measurement Error. Well-designed exploratory data plots can invariably pick these errors up (see section 5.2.) in near real-time. Timely intervention allows problems be addressed proactively and corrected accurately rather than estimated or simply removed as erroneous (Fig 3.2.2). More importantly, it actively informs QA for successive sampling so calibration may be increased for example if the scales cannot be replaced.

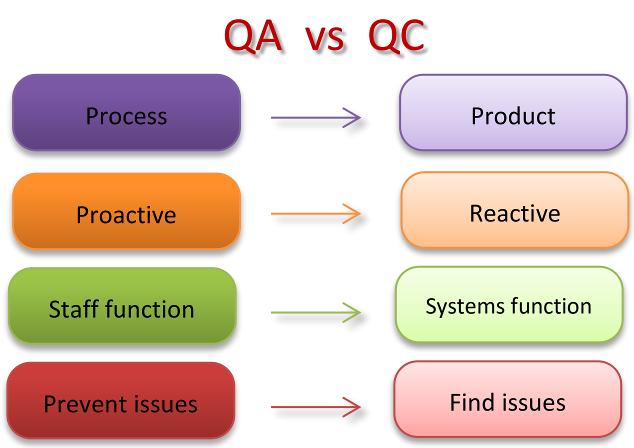


Fig 3.2.2 A data quality management overview of the contrasting functions of Quality Assurance (QA) vs. Quality Control (QC).

The objective of the SEATEC workshop therefore is to promote and support the application of technology and data processing skills currently prevalent throughout the QC process, to the critical point QA process of data collection.

3.3 Impediments to paperless sampling

While the advantages for data accuracy of paperless sampling may be intuitive, it is not generally applied. A lot of time has been allocated by IBTS for example to check the international survey data held in DATRAS and propose standard checks now within the upload process for all IBTS surveys. Despite the errors and methods to trap them being familiar, less than half of IBTS surveys routinely implement paperless sampling.

The reasons for this do not appear attributable to any single specific factor. However, discussion around this topic generally falls under three headings. The influence of each in promoting or impeding development in this area is generally specific to each Institute.

3.3.1 Financial/time cost

Initial investment in terms of budget for hardware and developer time can obviously be significant. In addition people’s time for testing, training and upskilling is likely to be required. Furthermore, existing data models and checking routines may need revision.

Highlighted by several presentations during the workshop however, significant budgets are not always necessary. Simple projects to capture and validate simple length, weight and basic meta-data can be built for a few hundred to few a thousand euros. Initially this approach tends to be driven by enthusiastic multi-skilled individuals often on their own time to some extent with little or no budget. It often results in long development periods and/or a ‘rustic’ product that may require some upskilling to use and maintain.

Given low budget approaches are feasible; cost alone doesn’t appear an exclusive factor therefore. However, it is at the higher end of the budget scale we start to see a conundrum. Few ‘’off-the-shelf” systems exist and the levels of functionality and support varies significantly. There is little in the way of a “standard” requirement or level of functionality as a result. Despite this, boards with even basic functionality often start around €5k-€8K each. For Institutes located in the EU not wanting to develop technology themselves they are then drawn into the public procurement process. Given the budget to equip a groundfish survey with one of the higher end systems currently in use is likely to be well in excess of €100k the tender evaluation process will benefit from clear specification.

Without being familiar with various technologies, functionality and what to expect from a ‘good’ system it can be difficult to specify the tender. Likewise evaluation of value for money is not a simple matter where proposals might range greatly in complexity, cost, levels of support and software licensing agreements, while track record is likely to be limited in such a niche market.

3.3.2 Technical expertise

Moving from paper and pencil to full electronic data acquisition may require some shift in skillset for some users. However, given the normality of technology in daily life nowadays this unlikely to be the event horizon it may have been 10 or 20 years ago. On the contrary, many current paper and pencil users in the field are likely to carry splash-proof personal smartphone technology in their back pocket. This is powerful, affordable and data secure and can reliably stream media and validate communications while a clipboard and pencil is utilised for costly scientific research data logging.

From an end-user perspective, the skills required for an electronic acquisition system should be comparable to any new business or personal technology. However, there is likely to be at least some upskilling for behind the scenes administration tasks such as configuration, hardware/software maintenance and data editing. What amount will depend on how developed the graphical user interface (GUI) is and the technology implemented. In turn this will depend on resources available to the project, skills level of the user team and support provided by the supplier or development team etc..

Finally, continuity and flexibility are further important technical considerations. Can the system be supported and developed over the longer term as new requirements and technologies arrive? Whether this support is provided by a supplier through a service level agreement (SLA) or by other project development team within the Institute will have implications for cost, flexibility and longevity. Indeed a considerable opportunity for ideas, IT skills and development drive could and should come from within the data user community. A great deal of QC work in particular, including the open source code, being done by many experts and users of fisheries data has been collated by various projects such as FishPi , Rome and FishTrawl are expanded on later in this report.

Technical expertise is still an important resource consideration in any scoping a project or tender. Participant discussion at WKSEATEC however suggested it is significantly less of a barrier now in terms of user acceptance than even a decade ago. This obviously is an important and timely shift in terms of stakeholder 'buy in' for teams wanting to get momentum behind the topic locally.

3.3.3 User acceptance

Any technology solution for data gathering is likely to have some costs in terms of one or more of the following:

1. Set up time for a sampling event
2. Access to and responsibility for expensive pooled equipment
3. Increased time for data entry
4. Upskilling
5. Requirements for greater/protected sampling space with power
6. Physical carrying of the equipment

This is not an exhaustive list, but at least indicative of the concerns often raised by users faced with moving to paperless sampling. A significant body of research around subjective measures to predict user uptake of a technology exists. Much of this resulted in the mid 90's in a Technology Acceptance Model (TAM), still used today (Davis, 1989, Venkatesh, 2000). In essence the model proposes that a person's likelihood to embrace and continue to use a technology is determined by 2 beliefs from the outset:

1. Perceived usefulness
2. Perceived ease of use

Their research findings suggest that, in reality, perceived usefulness can be up to 50% more influential than perceived ease of use in determining whether the technology will be successful. In other words, believing the technology has benefits for the individual themselves is likely to significantly outweigh concern around any ‘reasonable’ up-skilling that might be required to use it.

The underlying point here seems obvious, “perceived usefulness” is critical to the popularity of any technology project and therefore its successful implementation. Most people would agree that electronic data capture and checking from the point of measurement is likely to efficiently and consistently produce a high quality data product and that is very “useful”. The more controls that are put on the input data during sampling of course the less work that needs to be done by the data user at the end of the process. That of course means, potentially, the more work that needs to be done by the sampler. Their day may become longer or more technically demanding with motivation or energy to reapply these summary checks or find the exact problem fish diminished as a consequence.

User acceptance then needs to identify tangible efficiencies; broader understanding and greater ownership of the whole data collection process for everyone involved. In that context, technology development in this area should naturally offer a direct and mutually beneficial interface between measurer and manager where ideas and expertise can be traded and translated into informative plots and reports as well as user driven procedures to underwrite data quality.

4 Review of current approaches

4.1 Introduction

Electronic data capture in fisheries is established well over thirty years with RFID Tags (Prentice, *et al* 1990[1]) and Bar Codes (Sigler, 1994[2]) have being implemented since the mid-1980's. This section provides an overview of the presentations given by the WKSEATEC attendees for the technologies currently in use or development. It is not intended as a comprehensive review of all systems available, but gives a good general account of the issue list and technology being applied and developed currently within EU countries engaged in fisheries data collection.

4.2 Smartfish - ILVO

4.2.1 Overview

Requirements

The initial question for the use of an electronic measuring board came in the context of the implementation of the National Data Gathering Program (NDGP).

Yearly, Belgium takes samples for the NDGP during 40 commercial sampling trips and 2 surveys (BTS & DYFS). In total 500.000 fish length measurements and 15.000 fish biological parameters are gathered.

Since 2012 two Scantrol measuring boards were in use but most of the length data were collected via manual measuring.

Because manual length registration is labour intensive and error-sensitive ILVO looked for solutions to automate and improve the registration of length data. The main goals for an electronic measuring board where:

1. Improve efficiency
2. Improve data quality
3. User friendly
4. Reliable
5. Length and metadata
6. Real-time data integration
7. Cheaper than existing market solutions

No of Institute/sampling program users

ILVO / Fisheries and Aquatic production

Fishery Biology Unit: National Data Gathering Program

- Commercial sampling (40 trips): 4 observers
- Onshore sampling (40 trips): 2 observer
- Surveys (BTS and DYFS): 4 scientists + 4 observers

---

[1] E. F. PRENTICE, T. A. FLAGG, C. S. MCCUTCHEON, D. F. BRASTOW, AND D. C. CROSS, 1994. Equipment, Methods, and an Automated Data-Entry Station for PIT Tagging. American Fisheries Society Symposium 7:335-340, 1990

[2] M. Sigler, 1990. An electronic measuring board with bar codes. *Transactions of the American Fisheries Society* 123: 115-117.

Fishery Technic Unit

◎ Surveys: 4 scientists + 2 observers

ILVO / Aquatic environment and quality

Surveys: 4 scientists + 2 observers

Why commercial/customised solution?

In 2014 ILVO decided to build a custom measuring board. This choice was mainly determined by the fact that the existing commercial solutions where to expensive and not flexible enough.

4.2.2 Technology

Lessons learnt

ILVO had built the SmartFish measuring board in cooperation with a local company. The implementation from the design of the first prototype to the production version took more or less 3 years.

It was labour-intensive process to build a custom measuring board, but in the end all concerned are satisfied with the final result.

General

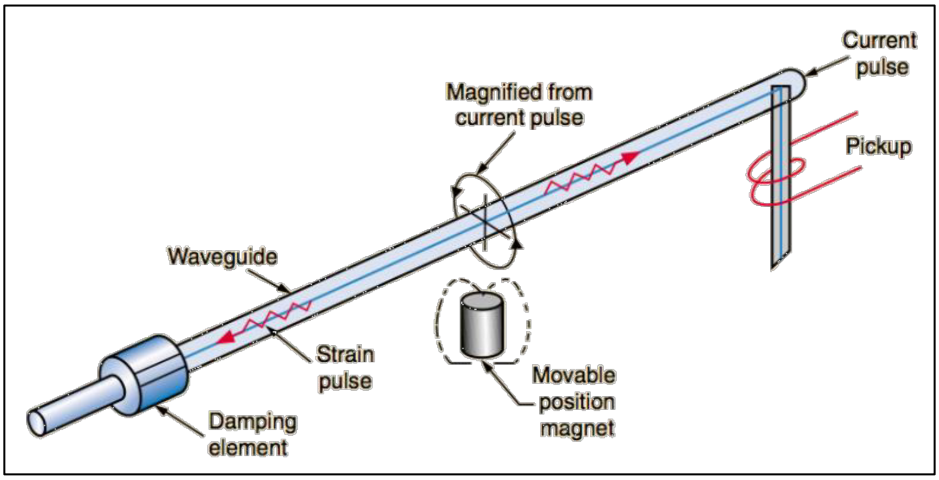
The SmartFish measuring board is an input device only:

- Similar to a keyboard
- No integrated computer
- No internal data storage

Technology

- Linear magnetic transducer
- Micropulse transducer in a profiled housing
- Measuring range: 105 cm

    - 100cm for measurement
    - last 5 cm for specific functions
- Resolution: 1 mm



Hardware specifications

- Dimensions
    - Length: 125 cm ; width: 21 cm ; height: 3,5 cm
- Weight: 11kg
- Material: PVC-CAW
- Reliability / durability
- Waterproof - IP67
- Shockproof
- Corrosion proof

Power component

- External battery
- Battery specifications
    - Power bank Li-ion battery
    - IP66
    - Battery status indication
    - Output: 5V
    - Overcharge / discharge protection
    - Autonomy: 6-10 hours
- Integrated in waterproof case

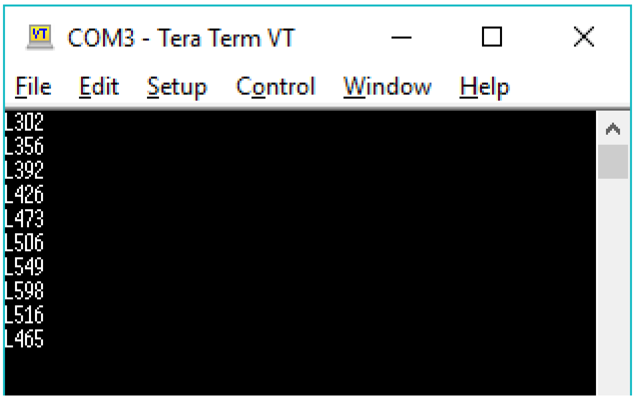
Other components

- Input/output connector
    - Single connector (Fischer push-pull)
    - Replaceable
    - Input: power (5V)
    - Output: data (serial – Bluetooth alternative)

- Led bar flasher
    - Power indication
    - Interface status indication
    - Data transmission indication

Data interface

- Data interface
    - Bluetooth module
        - Roving networks RN42 Bluetooth adapter
        - Low power consumption
    - Alternative: serial cable
        - Converter RS232 -> USB
- Data
    - Output = only length data



Length data conversion

◎ Handled in Smartfish software.

o The Smartfish software is custom registration software developed at ILVO. The output of the measuring board is always length data. The conversion from length data into alpha numeric or function key input is handled in the software.

◎ Input mode

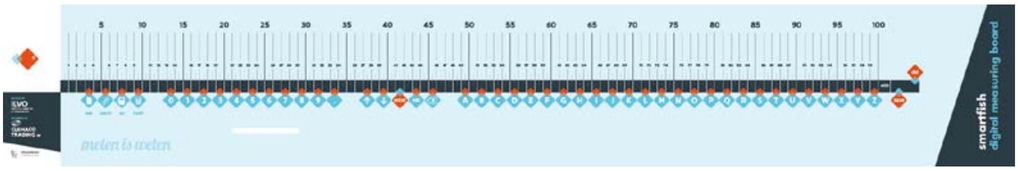
o Length

◎ Output value (conversion in Smartfish software)

o Length data

o Alpha numeric

o Function keys



Issues remaining?

The SmartFish measuring board is in production since 2016/09. All teething problems have been resolved.

The past 18 months the SmartFish measuring board has been intensive used, all involved are happy with the developed solution.

4.2.3 User expectations

Strategy to board users and data end-users?

The main expectation for the end-users was a dummy proof reliable solution.

Manage/balance/cost expectations?

One of the requirements was a cost which was lower than the commercial solutions. The building process in cooperation with the external company was a labour-intensive process. The total cost was higher than initially estimated.

Because it’s a small production, a single board cost around 10.000€.

Feedback

Because ILVO did not have the engineering skills in house, it was necessary to work together with an external company. The choice of the external company is decisive for the success of the project.

4.2.4 Implementation

Implementation difficulties and duration

To make the measuring board waterproof was a challenge. More attention during the design of the measuring board in function of the waterproofness had been very useful.

The implementation from the design of the first prototype to a production ready measuring board took more or less 3 years:

- 2014-2015: build prototype 1
- 2016/01-06: build prototype 2
- 2016/09: delivery 6 Smartfish measuring boards

4.2.5 Time needed for staff training

The time needed for staff training was rather limited: one or two working days.

All observers were involved in the developing process. Their suggestions were always taken into account. At the end of the process, there were no surprises.

4.3 Open SMB – Thünen Institute

4.3.1 Background

The "open scientific measurement board" (openSMB) uses a Beaglebone Black as embedded computer, running debian linux as operating system (Fig 1). Lengths are registered via megnetorestrictive linear position sensor. These sensors are widely available, rugged and accurate and have been shown to be a relatively affordable technology with which to develop electronic measuring systems (Towler and Williams, 2010[3]).

4.3.2 Requirements

Use cases

1. Observer at sea
2. Lab at sea
3. Lab in institute

Resistance / key parameters

1. IP67 / sea water resistant
2. Mains operation (230V) /recharchable battery (min. 10h)
3. resolution 1mm

Data handling

1. Compatible to TI-DB
2. Standard dataprotocolls (submission)

[3] R. Towler and K. Williams, 2010. An inexpensive millimeter-accuracy electronic length measuring board. Fisheries Research 106 (2010) 107–111

Future safe

1. Open source (circuit digrams, software, hardware)
2. Platform independent
3. Modular (use of standard-Hardware)
4. Interfaces for future requirements
5. Scalable

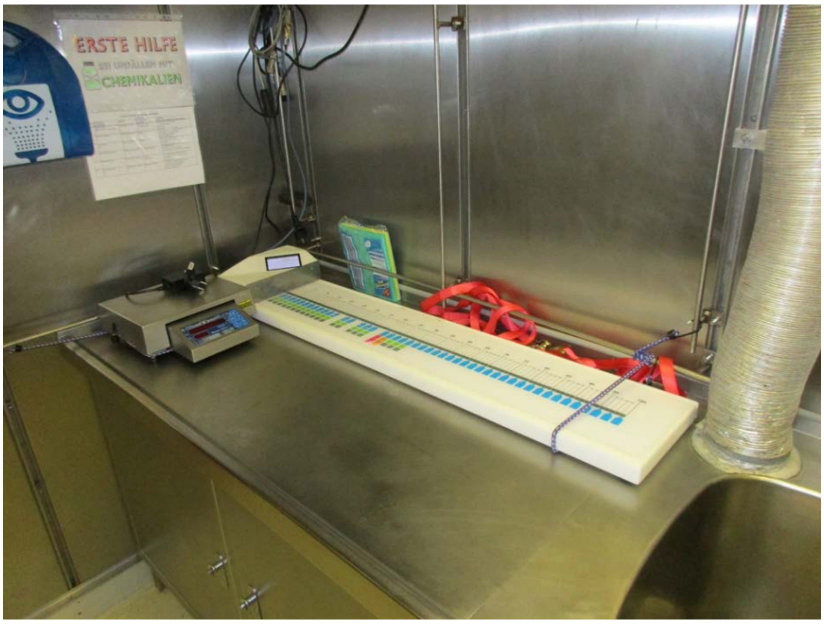


Fig 1. openSMB measuring board using linear magnetorestriction sensor and linked to electronic marine scales.

Software core components of the openSMB are a hardware service program and a PostgreSQL database server. The service program is reading user defined measurement instructions from the database. It interprets sensor signals in dependency of the measurement instructions and writes the resulting data values into the data storage of the database.

In detail the measurement instructions represent a user defined workflow / process. They are a sequence of actions which start a predefined or user defined set of procedures. This can be an output on the user interface (e.g. "press on scale to generate a length measurement"), the way how a generated data point is processed (e.g. "store it in the database, send it through an API to a server", "put it out on the serial port", "check for conditions", etc.) or any kind of action that can be scripted on the operating system or in the database management system.

The data itself is stored in a simple table with a recursive structure, which is based on the relations between data points.

4.3.3 Example 1 - Standalone

For illustration, we will use a simple measurement instruction where the trip is the root element. A haul is part of the trip. Then trip is the parent and haul the child. The haul itself is parent of each species found, while sorting the catch into buckets. The species (or to be more precise: the bucket containing the members of a species) is a child element of the haul and parent of each individual. All measured data points (length, weight …) have no further child, but they are a children of each individual.

Prototype testing has been started in 2016 during the IBTS. The data requirements are defined in IBTS manual. A process based on the IBTS manual has been defined and implemented as a measurement instruction. The PostgreSQL database server contains tables with measurement instructions and data. An entity-relationship-diagram is available.

The graphical user interface (GUI) is a webpage on the device. It connects to the hardware service program, measurement instructions and data through an "application programming interface"(API) written in python. Software other than the GUI then has the possibility to connect to the API. So the openSMB can talk to automated processes, other API's and external GUI's.

The API can be used to configure basic settings, the operating mode and trigger actions. It sends and receives data objects in "JavaScript Object Notation"(JSON) using HTTP(S).

Using the stand alone mode, the openSMB stores all data in the internal database. The data can be exported to a file or an external database later. Using the unidirectional mode, a complete dataset (e.g. of an individual) is pushed directly to a server / another API. The bidirectional mode allows a server reaction to the submitted data (e.g. trigger an alarm, if the submitted data seems to be inaccurate).

Interconnected operations allow data flow between a server and multiple openSMB's. This can be interesting for workflows with globally defined conditions. Another interconnected scenario is the direct interaction between two measurement boards. This can be useful to clone measurement instructions or copy data between devices.

4.3.4 Example 2 - Networked

An example for using the openSMB API with an external API is the central database system of the Thuenen Institute. The API of the "Database for Monitoring Aquatic Resources" (DMAR) is installed on the servers in the institute and on mobile data collection systems. An openSMB can connect to servers from the laboratory or in the field to the mobile devices.

Mobile data collection systems are made to collect data from different sources like ship sensors, manual input, etc... The installed software is the same like the server software in the institute's network. Data are stored temporarily and forwarded into the main servers.

APIs of the openSMB and DMAR are very similar. They are based on web server techniques and use http(s) as transport layer. The data itself and instructions are capsuled in communication objects (comObject). They are basically text strings containing key value pairs.

comObject strings consists of JavaScript Object Notation (JSON) defined in the "JSON Data Interchange Standard" ECMA-404. A key value pair is the basic element. A value itself can be a simple value, a list of values or a complete JSON object.

The basic structure of a comObject contains the (sub-) objects "data", "meta" and "errors". Data includes the data of an object like a fish. Meta contains additional/optional metadata of the data object and commands. The error object transports information about unexpected behaviours.

In an example situation the user clicks a link in the GUI. The GUI generates a comObject with a meta.command string asking for details of a specific procedure and sends the comObject to the measurement boards API.

After receiving a command through the API, the openSMB prepares an answer/comObject and returns it to the requesting instance. In the example, the GUI displays a detailed view on the requested object.

For the openSMB it makes no difference if it is instructed by an internal GUI, an external GUI or other automated processes, because everything is handled by the same API.

For different user groups the openSMB offers different possibilities of access in dependency of the user's level of experience.

4.4 ICROS - IEO

4.4.1 Background

Our survey data acquisition model has not changed in decades except for the capture of intrinsically digital data such as GPS, echosounders, etc. The remainder of the data collected (biological sampling, length frequencies distributions, etc.) is still mostly being collected analogically (pen and paper).

The analog capture model is:

- Inefficient, because it requires at least two people (sampler and recorder) for the majority of the data and, at least, another individual for data entry (typing) and QC of the data. That is, every single datum passes through, at least, 3 people.
- Prone to data transmission errors, due to the high noise level of the work environment.

On the other hand, the growing multidisciplinary nature of the fisheries research surveys makes it difficult to have numerous fish sampling teams. It makes sense to think about developing systems which improve the process of data collection:

- Making data collection and recording more efficient: Every single datum has to pass through the minimum amount of people possible.
- Reducing errors caused by bad communication/transmission of data.
- Reducing the incidence of transcription/typing errors.

Potential users and scope of the system

Though the system was conceived for survey and lab sampling, the considerable amount of data and measuring operations which are carried out by observers on board and at the fish markets highlights the need for the system to be extended to those areas. This would imply implementing another approaches in relation to software development but also regarding to hardware, providing the infrastructure available at fish markets and commercial vessels doesn't meet the system requirements.

For an estimation of potential users there are two main sampling scenarios:

- Surveys at sea on research vessels, or land laboratories, where the whole system can be installed and used by the teams operating the vessel or lab. In the case of Spanish Institute of Oceanography's surveys and labs, it means up to 4 on-vessel setups and 9 labs, with an average of 4 measuring boards and sampling kiosks in each setup, so about 13 complete systems and around 50 measuring boards would be enough despite for the laboratories a centralized setup could be considered, thus reducing the number of servers, etc. needed for the system to run.
- Fish market/on board observers: In this case deployment of complete systems is not possible but using the measuring boards with a suitable software in

hand-held or portable devices is desirable and one person, one device approach is more realistic. In 2017, Spanish Institute of Oceanography's fisheries sampling program (covering CECAF, ICES, Mediterranean and other areas) involved 75 samplers, and similar numbers in 2016 and 2015, so between 70 and 90 devices/portable systems could be considered for this scenario.

There are other institutions in Spain with responsibilities in fisheries management, not only at sea but in inland waters which could be interested in using these devices; however the potential number of users is unknown.

Why Open Source (OSS) and Open Hardware?

- Most OSS licenses allow the user to operate the system with no limitations apart from distributing the source code with the application.
- System evolution is linked to users (scientists) needs: No particular interests should block the evolution of the system (commercial or others).
- On the cost side: All the software is free, hardware components are cheap and off-the-shelf helping keep within a low budget. The system is conceived for systems with reduced performance, so old systems can be reused instead of investing in expensive up-to-date servers, computers, etc. For instance, the system at Cadiz laboratory and in research vessels Miguel Oliver and Ramón Margalef for IEO's surveys ECOCADIZ and ECOCADIZ-RECLUTAS runs on a 2008 HP WX8400 Workstation with 4 sampling kiosks attached without issues.
- Ease of building a community of users for support and enhancement of the system.

4.4.2 System basics

The key aspect of the system is the change in the model for data storing and management. Given the traditional model of storing the data in file-based databases (MS Access, DBASE...), in this case a server-based RDMS (relational database management system) is used, which makes the following possible:

a) Data typing/management can be performed by several people simultaneously.
b) Real time data availability in multiple places, for editing, visualization.
c) All survey data can be managed and stored in a single place.
d) Serious RDMS have features like user management, replication, synchronization, and many others which makes it possible to build both simple systems as well as sophisticated ones.
e) RDMS like PostgreSQL or MySQL are scalable so are able to run both on low-end computers (for stand-alone sampling at lab, for instance) or as a full-scale server (for institution/lab scenarios).

The IcrOS system is, wherever possible, based on open source and open hardware technologies.

Software:

i) PostgreSQL as RDMS, with PostGIS extension for management of spatial data.
ii) Shiny-server and QGIS for visualization.
iii) Python and R as preferred programming languages. Both are mature and thoroughly tested programming languages, with active development communities and are well accepted by the scientific community, having plenty of libraries and

extension modules to satisfy development requirements. Also, Python is an interpreted programming language makes undesirable compilation steps unnecessary.

iv) LTSP4/PiNetDev for sampling kiosks infrastructure.
v) dnsmasq as DHCP server.
vi) O.S. Linux, on the main server, all the sampling applications are multiplatform and Windows could be used as the server if a solution for terminal server is provided.
vii) Data acquisition and management software developed so far:

- osb_pelakamp, for catch sampling, length frequency distributions and biological sampling.
- lancero for haul metadata (position, time...) recording.
- icrosgui for LFD sampling with digital fish board icrOS
- osbreportgenerator for reports (biological sampling, LFD data...)
- Several shiny apps for real time visualization and fast reporting.

The Hardware:

Annex 1: Main server: Currently the system runs on a refurbished HP WX8400 Workstation for surveys at sea and at the lab. However, the whole system will also run on a Toshiba Tecra-M9 laptop.

Annex 2: Sampling kiosks: Raspberry Pi B+ (the system could work with old Raspberry Pi B as well, and possibly other boards of this kind like OrangePi, Beaglebone, etc.)

Annex 3: Networking: Uses a Gigabit network (switches, cards) is mandatory due to the requirements of LTSP. Use of Power Over Ethernet (PoE) is desirable, but not mandatory.

Annex 4: Connection of RS232 devices (GPS, echosounders, some marine scales) to the system: RS232-Ethernet Expert EX9132C2 converters.

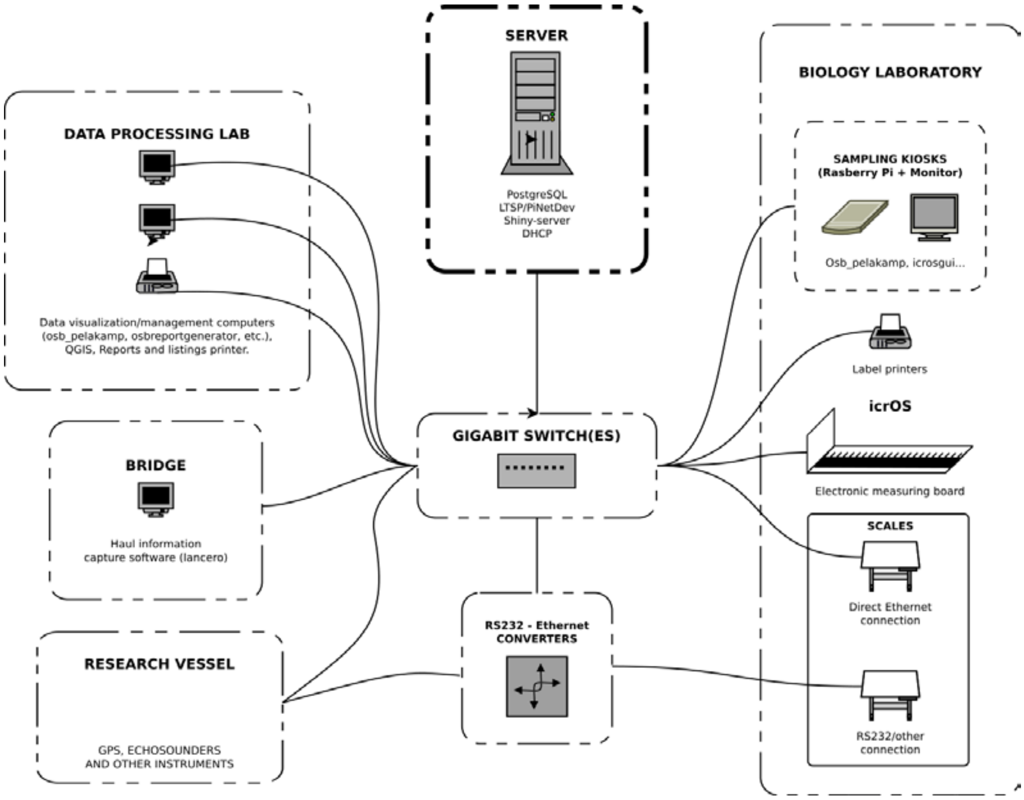
Annex 5: NAS for data backup and storing.

Annex 6: Digital fish board icrOS.

---

4 Linux Terminal Server Project

Annex 7: Other network hardware: Splitters PoE, switches 10/100/Gigabit.



4.4.3 Sampling kiosks: LTSP, PiNetDev and Raspberry Pi

These kind of systems should be used in processing rooms and wet labs. In both cases, the environmental conditions are tough: Humidity, hot/cold temperatures, condensation and dirt all around. This makes it risky to use laptops for data entry/capture, and the alternatives (ruggedized devices) are expensive. Moreover, using several machines for sampling requires individual maintenance of software (applications, OS, etc.), thus a low budget, low maintenance system needed to be found. The software solution comes from LTSP/PiNet and the hardware by using Raspberry Pi.

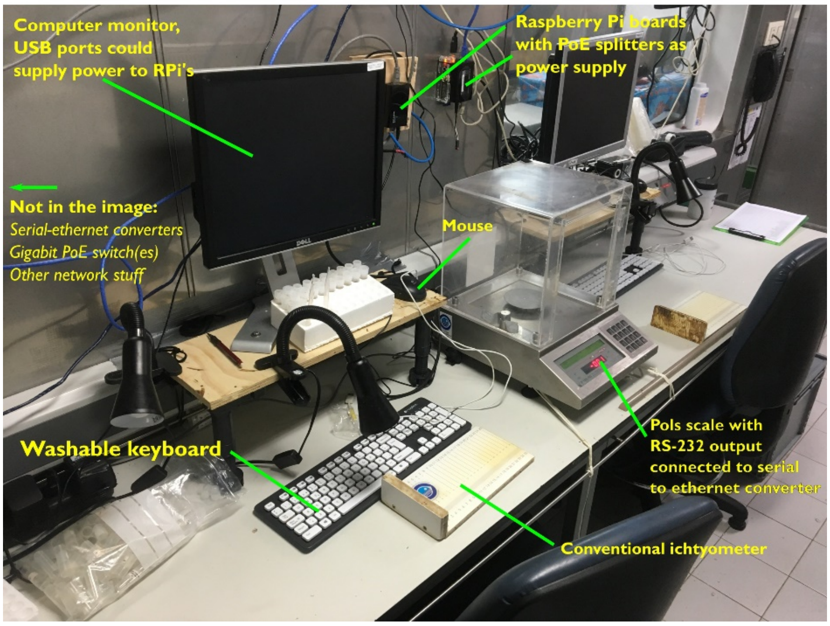
LTSP (Linux Terminal Server Project, http://www.ltsp.org) is a terminal server designed for easy deployment. The server stores an Operating System image (containing all the software necessary to run the system) which is served to the terminal clients at boot time through the network, it is only necessary to update and maintain a single copy of the software. Any change in OS or software application is distributed to the terminal clients by the server, ensuring all the client devices execute the same versions of the software.

PiNet (http://pinet.org.uk) is a project which significantly eases the deployment of LTSP setups.

Raspberry Pi is a low cost system-on-a-card computer (about 40€ for the latest Raspberry Pi version 3) with usb connectors, HDMI output and Ethernet connection. The configuration of these boards is minimal with LTSP in charge of serving the OS to the boards. This eases the maintenance of the system, because it is only necessary to maintain an OS which is distributed to all the boards. The boards run on 5V DC so can be easily powered by USB from monitors, PoE or even batteries.

LTSP clients (RaspberryPi's in this case) can operate in two modes: As thin clients, in which all processes of the clients occur in the main server with the kiosks acting as mere screen terminals, or as fat clients, in which the processes occur on the client itself. Fat clients cause less overhead on the main server, so other processes can be run (Shiny server, for instance) but also makes it possible to plug USB devices (as well as keyboards/mice) into the clients such as label printers, calipers/measuring boards or scales, however, the system concept favors using network connections to allow resource sharing.

Sampling system setup onboard R/V Ramón Margalef
ECOCADIZ-R-2016 survey, October, 2016



Networking: As stated above, the network should use gigabit components at least for the main server and switches. However, for testing and standalone installs (max 2 fat clients in the system), Fast Ethernet (100 Mbps) connections seem to be sufficient. The actual limit of connected fat clients (sampling kiosks) for such a network connection has not been tested however. Ethernet connections are preferred because programming is not as platform-dependent as is the case for serial/Bluetooth connections. Moreover, attached devices can be easily shared between kiosks. For instance, scale sharing makes multi-sampling or shared sampling operations possible, while GPS/echosounder NMEA streams make position and depth available to any computer attached to the system.

Apart from icrOS, the main input devices are the inexpensive Logitech K320 Washable Keyboards (~20 €) which can be washed with water and soap after use. Experience has shown that the latter devices are tough enough to survive in a fish lab environment for many surveys with little or no maintenance apart from cleaning of scales and body fluids.

4.4.4 System capabilities

The system is in beta stage and still under development, however in its current state it is used for surveys at sea and at the lab.

Software has been developed in Python 2.7 and PyQt4 libraries, a custom widgets library is available to ease the creation of new GUI's. These libraries are available under a GPL 3 license in https://github.com/jtornero/jtCustomQtWidgets. However, for data visualization and reporting migration to R/Rmarkdown/Shiny web apps is in progress to the ease report and web programming using the R ecosystem than Python/Reportlab.

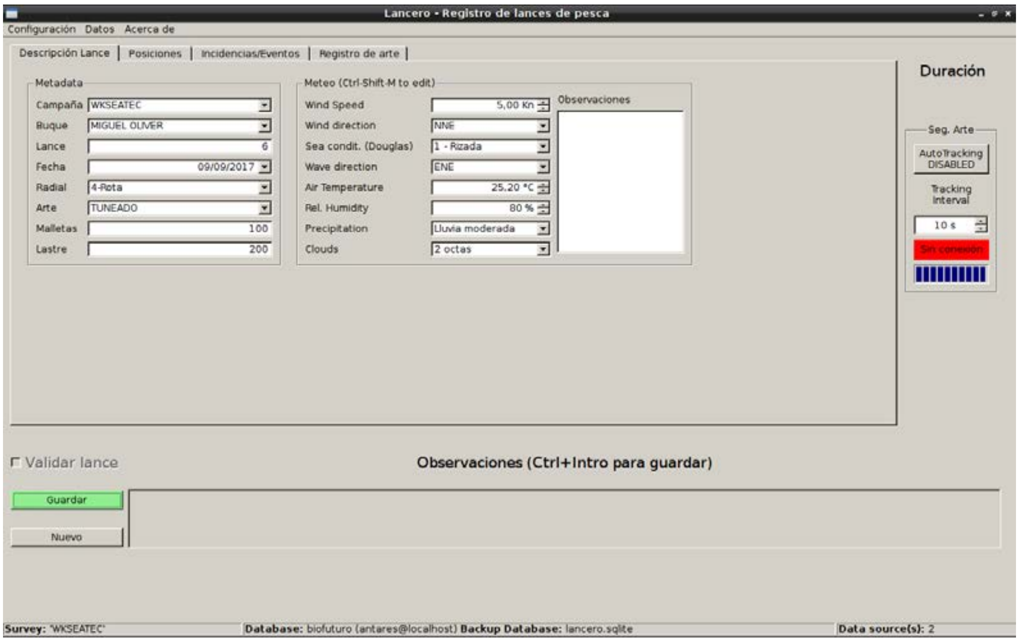
Haul data entry

It is still possible to type the data (haul, coordinates, depth...) from a form; however, effort has been expended to make an application, lancero, which makes it possible to capture the data directly from the vessel's instrumentation. The application has the following capabilities:

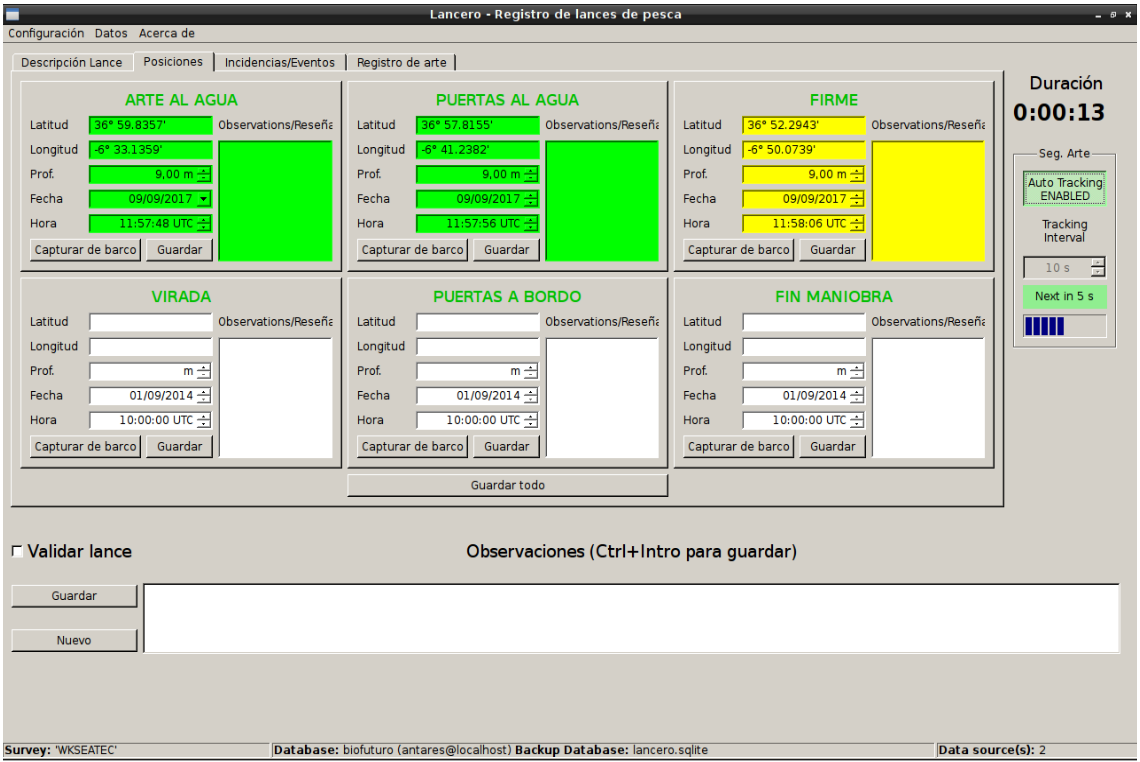
1. Data are made immediately available for visualization in spatially enabled applications (QGIS, etc.)
2. The configuration of data sources, DB, etc. is performed through simple plain text files:

```
[dbserver]
server=192.168.X.X
user=XXXXXXXXX
password=XXXXXXXXX
database=biofuturo
autologin=True

[backupdb]
filename=lancero.sqlite

[survey]
name='WKSEATEC'

[source1]
sourcename=LABO-UDP2
type=GPS
protocol=UDP
ip=localhost
port=24000
data=depth

[source2]

…
```

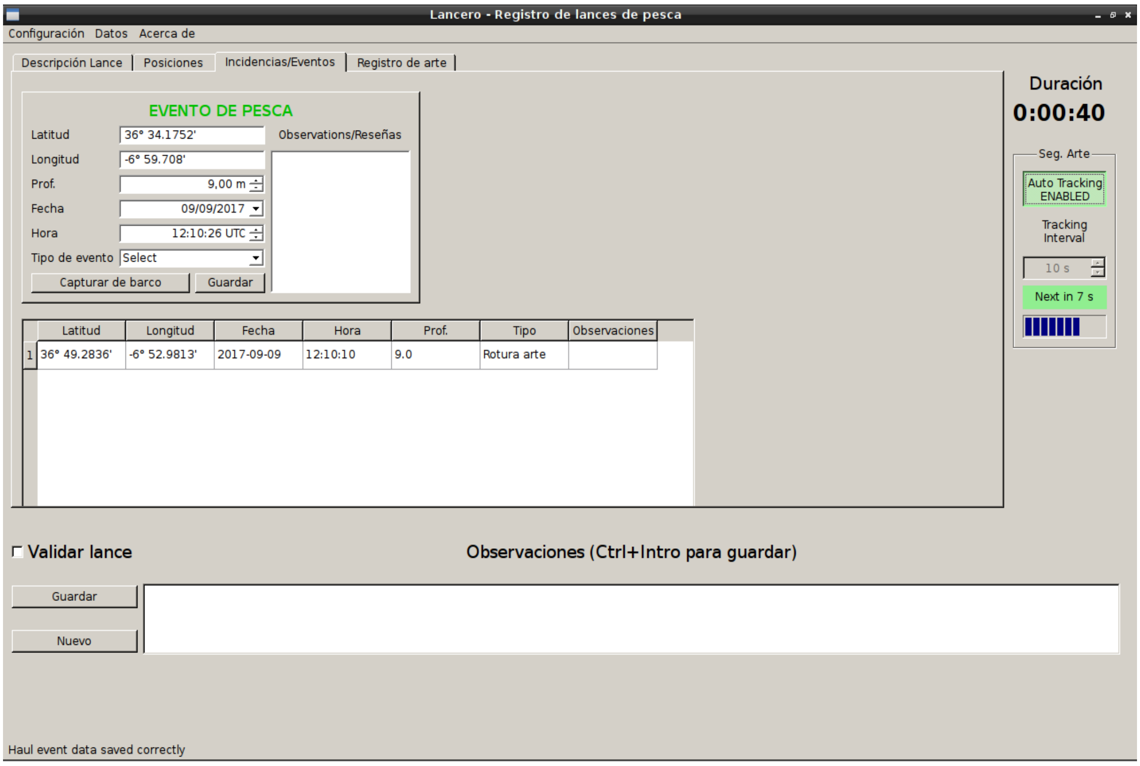
Manual input of haul metadata like: Survey, Vessel, Haul, Transect, gear configuration (wind ropes, door, ballast) and meteorological data.



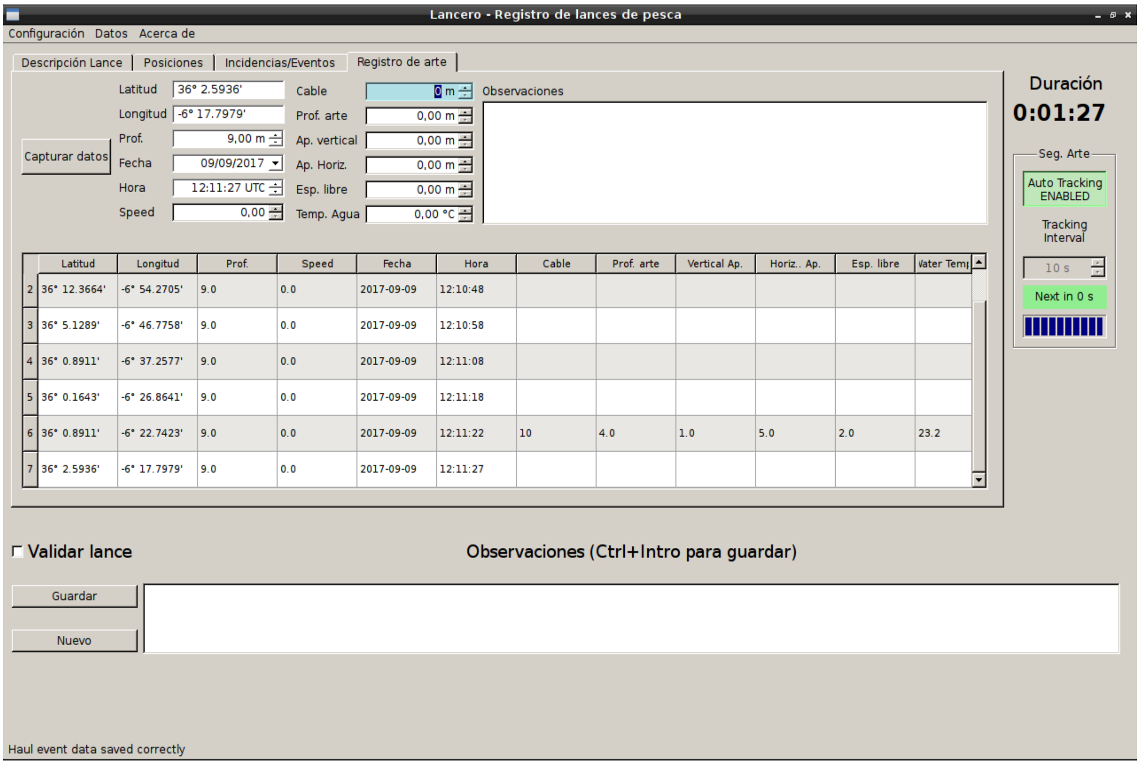
Coordinate depth and date/time captured from the vessel for the haul events (net shooting, otter boards, haul start…).



The same for haul incidences (presence of obstacles in the sea floor, problems with gear, whatever).



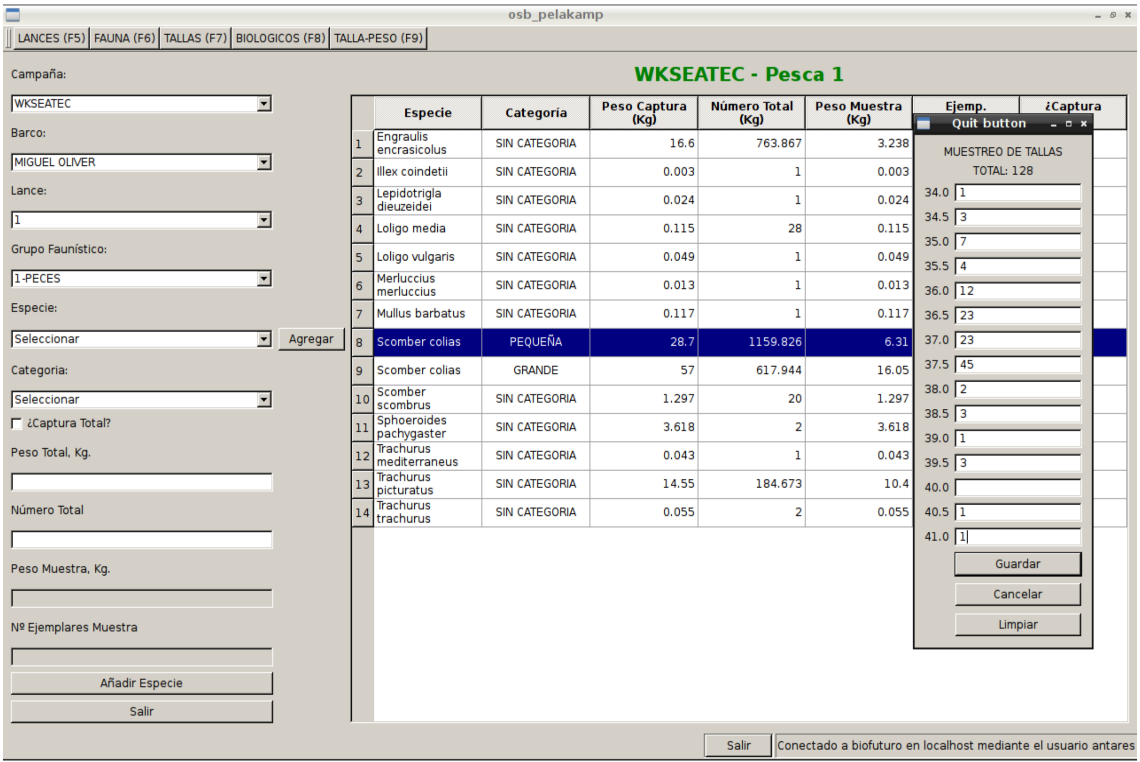
Regular interval haul tracking with coordinates, depth, time... Recording of data from gear sensors in progress, but difficult to scale due to the variety of systems present on vessels.



Catch and LFD data

One of the changes proposed is forgetting about codes-everywhere. Having drop-down menus connected to the DB with foreign key resolution makes it possible to select species, sexes, maturity stages by their nominal value rather than having to introduce (and later check) codes. It also makes it possible to get rid of "value not in the list" errors and, as a consequence, dispense with quality control on those issues (at the end, the QC is performed, but at the earliest stage possible, virtually before data input).

Both catch recording and LFD sampling can capture weight from scales, so it is possible not only to type the data in the traditional way (from forms) but input the data directly into the database capturing weighs during sampling. Again, this makes possible fast data availability.



As a sampling aid, weight fields can operate in sum mode, in which several weights can be added up and the field is finally updated to the sum of all the individual weights entered. It is useful in the case of weighing several boxes/baskets for a species.

Biological sampling

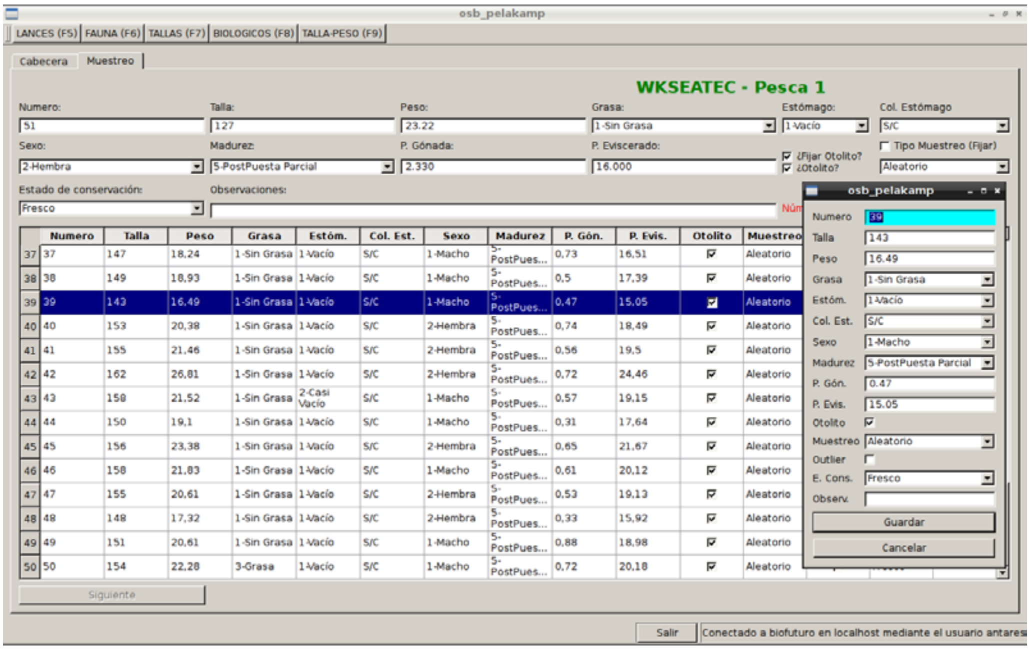
Taking into account that the software for biological sampling has been developed for the small pelagic surveys of IEO's PELCOSAT project, is not as flexible as desirable. However, efforts are being made to make it more flexible in terms of sampling protocols depending on species, etc.

Configuration of scales, etc., is done through plain text files as depicted before in the case of haul data capture software.

All weight data are captured from scales: Nowadays, capturing weights is possible from MAREL, POLS and Mettler-Toledo scales. Multiple scales can be configured and it is possible to do that on the fly (for instance, the scale for weighing the gonad is different from the scale for weighing whole fish).

Multi-samplers or shared sampling: Because scales can be shared, it is possible to sample different species with the same scale, or two or more people sampling the same species with the same (or not) scale.

Quality control: Some simple quality checks are available, like total weight<=gonad weight + gutted weight. Weight or TL>0 and weight>0. Also, with R/shiny apps simple graphical QC has been implemented, based on length-weigh scatterplot.



Other sampling aids: A control on number of mature females in the sample has been implemented for DEPM surveys. Also, automatic sample code generation for labelling, as well as the functionality for printing labels has been implemented. However, in the present version those features required source code modification, so implementation of additional features is not trivial, a different approach needs to be taken.

LFD sampling with icrOS

The measuring board icrOS is based on OpenSource/OpenHardware (Arduino) technologies and works under a simple operating principle: The operator reads the length of the fish (just like in a conventional measuring board) and presses a key for the length class, it is transmitted, along with the specimen's sex and whether it is an “add individual” or a “remove individual” operation (in the latter case for error correction purposes) to the sampling software, where is processed.

Despite Arduino boards have out-the-box serial USB communications enabled, communications by LAN are default in the current state of the Arduino sketch (script which runs into the arduino). However, as Arduino Sketch is OpenSource, the user have the opportunity of fitting the device to their particular needs and switching from one communications technology to the other is fairly simple.

The board operates using membrane keyboard technology, which offers a low budget, waterproof and reasonable durability solution. The technology is widely available in most countries and the price of the keyboard itself decreases dramatically depending on ordered quantity.

The board has a length of 60 cm. This selection takes into account the case of IEO's ECOCADIZ/BOCADEVA series, it covers almost 100% of sampled fish (only 46 over

185127 sampled fish for the whole series measured over 60 cm). However, it could be extended up to 95 cm without major changes due to the limitations of the Arduino Mega boards. Possible solutions for length extension could be obtained, however, through special measuring modes like origin shift, board inversion etc. or by mean of extension boards. However, these possibilities are still at the design stage.

Each key is for a length class, but also five additional keys are available: Delete mode key; female, male and undetermined sex key; function key. The latter makes possible to alter the behavior of the other keys of the board, for instance. Because developers have access to the key matrices in the source codes, all keys are programmable through sketch modification without issues.

Key activation pressure is 350 g, and it can be adjusted at manufacturing time varying the key dome specification, this activation pressure is high enough to avoid accidental keypresses so far.

In its current design the board is **powered through PoE** (Power over Ethernet) which is an advantage in wet conditions, having to rely on just one cable connection, though battery operation is being tested.

Communications are enabled by LAN connection. The system to which the board is connected takes care of the connection as well as reading the measurement data emitted by the board. Network configuration of the board is stored on a SD Card attached to the hardware, it is readable/writable through the network via a simple telnet connection. IP address can be either fixed or through DHCP. The port for communications is also configurable. A configuration file consists of a single JSON string as follows:

{"DHCP":"False","IP" :"192.168.2.100","PORT":23000}

which indicates whether the IP assignment is going to be performed manually or by DHCP and, in the former case, the IP and port of the device to communicate on.

Communication protocol is as follows:

- **Data output:** A string containing type of response (currently only "MEAS"), whether the board is in "add" or "delete" mode, sex and length class in mm, with semicolon as field separator:

  MEAS;D;1;125

  "Delete a male of LC 125 mm"

1. **Configuration and miscellaneous commands:**

   a. **c** command outputs the configuration (IP and port) of the board

   b. **ip** sets the IP of the device, i.e. ip192.168.2.100

- **Acknowledgement signals:** The board emits a loud beep with different tones depending on operation when receiving one of the acknowledgement signals. This makes it possible for the system to which the board is attached to certify that not only the measure has been received but successfully stored in the system. Currently the signals are:

  a for an individual adding operation.

  d for an individual removal operation.

The system is conceived to keep a low budget profile. All component and spares, except for the keyboard and the housing itself are off-the-shelf components which should cost less than 30 € and, in the case of keyboards and probably housings their price will drop dramatically with quantity. However, for the last reason, this budget should be considered carefully:

Considerations regarding design and further development: While developing the board some problems had to be faced:

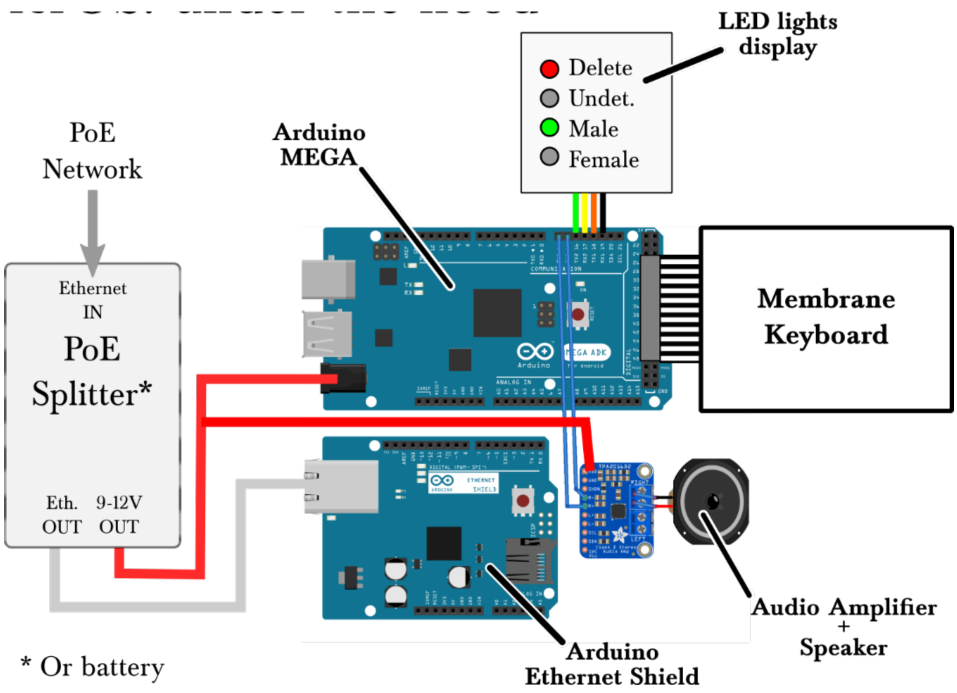
- The membrane keyboard technology seems to have some limits, but they could be overcome:
    - Because the keyboard production involves screen-printing of the circuitry and decals, **maximum length of the keyboard could be limited** by the keyboard's manufacturer's machines.
    - **The size of metal domes used for the keyboard keys impose restrictions about how close the keys can be placed.** That's why it is not possible to have a single row of keys, in icrOS, whole centimeter keys are on top of the board and half centimeter keys at the bottom. However, this is also manufacturer dependent and smaller domes could be employed to make single line key arrangement at least on each side instead of the actual staggered pattern, to improve the ergonomics of the device.
    - This technology imposes other restrictions, mainly due to the **need for enough room for PCB tracks.** However, this could be superseded adding more layers to the circuitry, at the expense of raising the cost of the keyboard.
    - **Special care should be taken with respect to EMI** (Electromagnetic Interference), taking into account the "perfect antenna" effect of having linear, long PCB tracks in the keyboard's circuitry and working in an EMI favorable environment due to ship motors, etc. At least with the first icrOS prototype it was an important issue (up to the point of making it unusable under certain conditions) and proper shielding of cables and keyboard was necessary. In the next production of keyboards, shielding will come from the factory and a rearrangement of electronic components (closer to

keyboards output, so no need for long cabling) could be adopted. In this sense, at design stage it is worth thinking about the keyboard's cable end placement to make it as close to the electronics as possible, as well as designing the keyboard track in a more EMI-proof manner.

- A change in measurement units (metric/inches) would require a different keyboard design to be usable, not only the top layer (where scale, texts, etc., are printed) but also the PCB where switches are place, though a software solution (measure translation) may be possible but has not been tested.

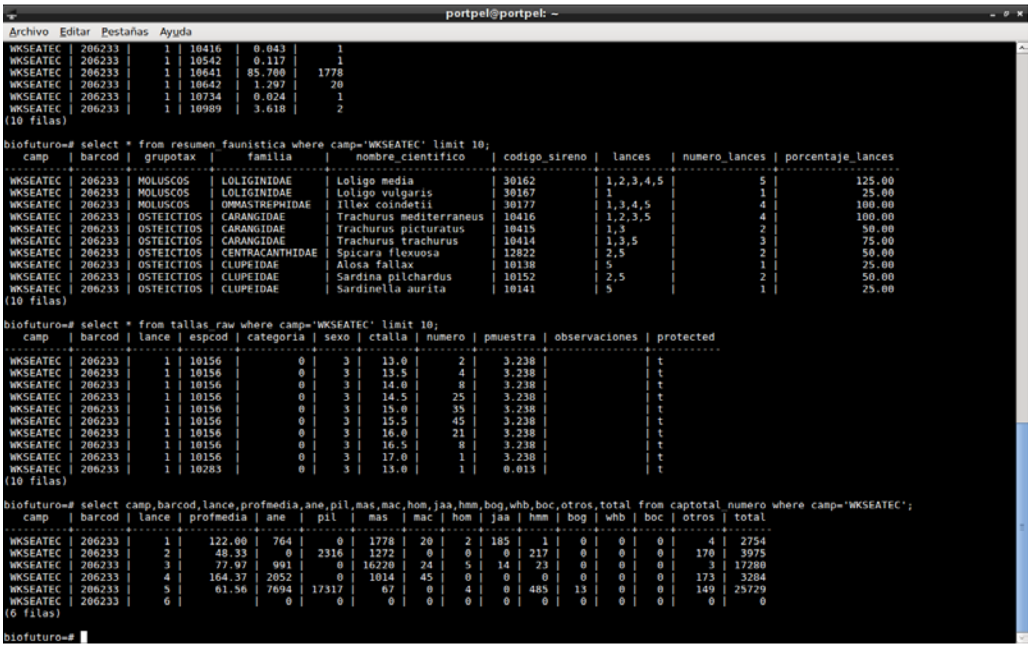
- The electronic casing is currently situated at the leftmost end of the board, but placing electronics under the board would make possible getting rid of *hard stops* at the end of the board in favor of movable ones. That would make the device more compact and easier to transport.
- **Further developments:**
  - The modularity of Arduino hardware makes it possible to enhance the device with limited further investment, i.e. adding GPS and GSM capabilities to a board could cost under 120€ plus the code modifications, so new functionality could be implemented with less effort:
    - **Battery operation** for non-PoE environments.
    - **Serial communication:** As stated before, Arduino have out-of-the-box USB serial communications, so it could be matter of just changing several lines of code. However, it would involve providing the box with another waterproof connector. In this case, powering of the board could be performed through the same USB connection.
    - **Bluetooth serial connection** could be implemented with little coding effort as well, but in this case powering the device with batteries is mandatory if complete wireless operation is desired. Otherwise, a power source must be used.
    - **Self-contained data:** The Ethernet shield, which is part of the device, has a SD card slot in it (where the configuration file is stored), it could be easy to implement routines for data backup and/or self-contained sampling. But in the latter case, in which icrOS became a complete sampling system, a screen and more complex methods of data input (for sampling metadata) should be provided, unless another device could act to support the sampling system.
    - **Audio/visual signaling:** For user feedback could be improved.
  - **With respect to software**, possible enhancements include:

- **Add board extension capabilities**, i.e. setting the origin to +50 cm for measuring longer species. Board inversion, as seen in an old commercial solution.
- **Add metadata input capabilities** for self-contained sampling, though it probably involves changes in the hardware as well.

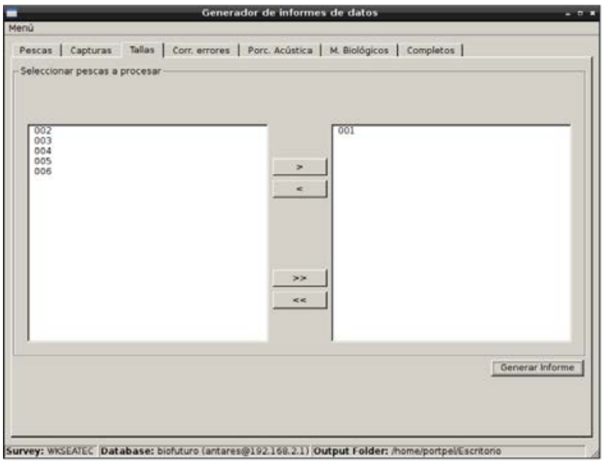


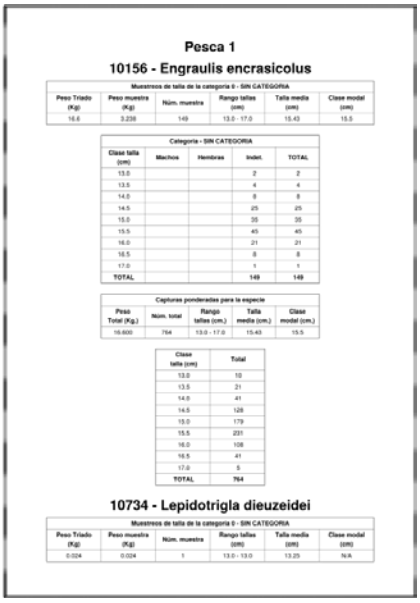
4.4.5 Visualization and reports

One of the main attractions of a system which relies on a RDMS server is the fast data availability and the ability of the database to perform processes on that data automatically. In the case of PostgreSQL this is performed through queries to tables, views and materialized views from database clients. Those clients could be terminal-like clients in which the user interacts directly with the database server or different purpose applications. Here the limits of what can be done are just the developer/user skills in programming such queries.

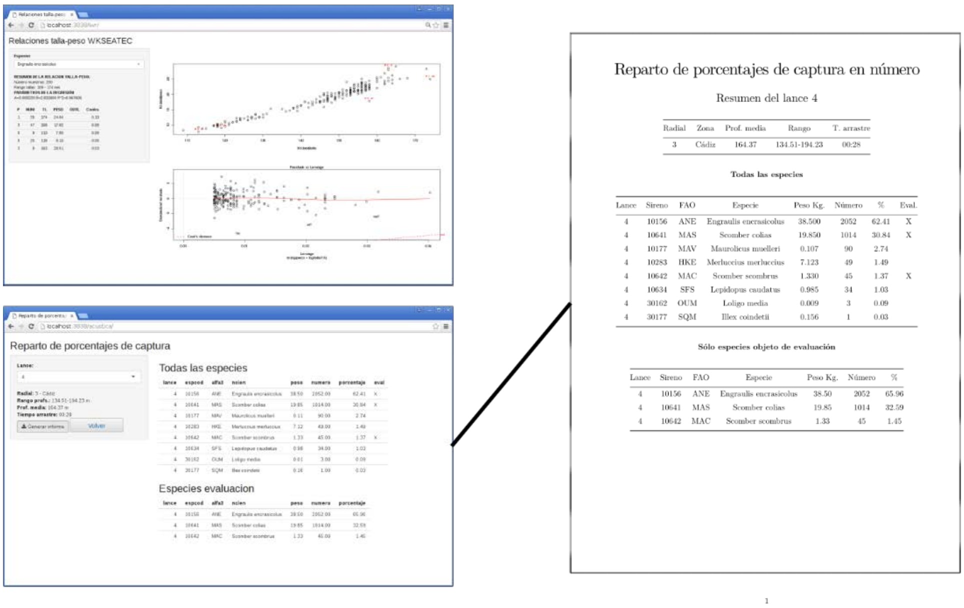


A Python/Reportlab application (osbreportgenerator) produces PDF reports like haul metadata, catch report, biological sampling, LFD report, and a QC control of catches/LFD data:





However, efforts are being made to migrate to R/Shiny web applications, which provides more friendly coding of the report applications and a better user experience through the use of simple forms, drop down menus, etc , in a more familiar, friendly environment like a web browser. Until now, apps for otolith collection control, LWR control, biological sampling to LFD conversion, catch report among others have been developed.



Reparto de porcentajes de captura en número

Resumen del lance 4

Radial	Zona	Prof. media	Rango	T. arrastre
3	Cádiz	164.37	134.51-194.23	00:28

Todas las especies

Lance	Sireno	FAO	Especie	Peso Kg.	Número	%	Eval.
4	10156	ANE	Engraulis encrasicolus	38.500	2052	62.41	X
4	10641	MAS	Scomber colias	19.850	1014	30.84	X
4	10177	MAV	Maurolicus muelleri	0.107	90	2.74	
4	10283	HKE	Merluccius merluccius	7.123	49	1.49	
4	10642	MAC	Scomber scombrus	1.330	45	1.37	X
4	10634	SFS	Lepidopus caudatus	0.985	34	1.03	
4	30162	OUM	Loligo media	0.009	3	0.09	
4	30177	SQM	Illex coindetii	0.156	1	0.03	

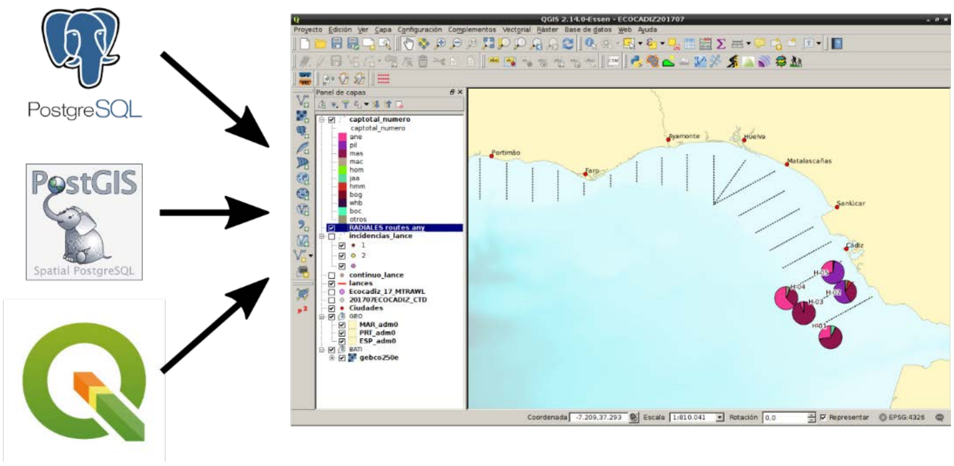
Sólo especies objeto de evaluación

Lance	Sireno	FAO	Especie	Peso Kg.	Número	%
4	10156	ANE	Engraulis encrasicolus	38.50	2052	65.96
4	10641	MAS	Scomber colias	19.85	1014	32.59
4	10642	MAC	Scomber scombrus	1.33	45	1.45

1

Biological sampling QC has been proved particularly useful through length-weight scatterplot and data identification as shown above.

Finally, enabling the PostGIS extension in the database makes it possible to add spatial analysis capabilities to the system. This enhancement allows spatial QC (for instance, checking if a haul is inside the survey area) to be applied to the data and also the possibility of mapping results in almost real time through spatially enabled applications (QGIS, SpatialR...)



4.4.6 Experiences implementing the sampling system osb_pelakamp

The development of osb_pelakamp system and associated icrOS measuring board began in 2011 and the system was conceived, at first, to assist in the biological sampling of European anchovy (*Engraulis encrasicolus*, Linnaeus, 1758) both at surveys at sea and laboratory. This is the timeline of the project so far:

YEAR	OSB_PELAKAMP DEVELOPMENT	icrOS development
2011	First use of the system, limited to biological sampling of anchovy in BOCADEVA0711 survey. Software ran in a laptop, no LTSP. No weight capturing from scales implemented.	
2012	Implemented weight capturing capabilities for MAREL and Mettler-Toledo scales.	First 25 cm PCB prototype, with tactile switches. Not operative but useful as development board for firmware development.
2013	Implemented haul data, LFD sampling and species/weight data typing GUIs.	
2014	Added some QC capabilities on biological sampling. Both anchovy and sardine are sampled with the system in surveys at sea.	Moving into Ethernet communication instead of serial.
2015	Implementation of routines and GUIs for multiple samplers biological sampling. All biological samplings of ECOCADIZ-R-2015 (8 species, about 5000 sampled specimens) survey are carried out in sampling kioks. First deployment of complete system (LTSP+Raspberry Pi as thin clients). Development of *osbreportgenerator* as reporting application.	
2016	Development and implementation of *lancero* software for haul data capture from GPS and echosounders. Improving data editing capabilities of the GUIs to get rid of SQL data fixes.	Moving to membrane switch technology. First 60 cm operating prototype and tests in surveys ECOCADIZ201607, PORCUPINE16 and ECOCADIZ-R-2016. Detected issue with EMI under certain conditions (vessel)
2017	Implementation of a R-Shiny server into the system for QC and reporting. Improved LFD and species weight data typing GUIs. Implementation of routines for sample labeling in biological sampling.	Improved prototype with EMI shielding. Implemented remote configuration of the board. Successful tests in ECOCADIZ201707 survey with new software for data capturing. Presentation at ICES WKSEATEC .

So far, the system has been used in 15 surveys of the IEO's ECOCADIZ and ECOCADIZ-RECLUTAS small pelagic fish survey series by about 15 different users. Despite no poll has been carried out for a precise knowledge of the users feelings, in general users are satisfied with the system. After all, it saves lot of times not only in typing but in hands availability for other tasks releasing pressure on fishing team.

The most welcomed improvements are those related with QC and some routine operations which are time consuming like, for instance, a recently implemented R-Shiny app which transform length measurements from biological sampling into length frequency distributions.

Testing in real survey conditions has made possible improving the system in many ways, not only with new code. For instance, it seemed that users who sample at the right of the scale (when two sampling kiosks share a scale) had more length-weight inconsistencies while sampling, even when weight data were captured from a scale. After some observation, it was pointed out that right side users tend to put the fish in the scale with their left hand and press the weight capture key with the right without waiting for the scale to be stable. Since then, when training users on system special emphasis is done for this mistake to be avoided and data quality has improved as a consequence.

Ensuring data backup as shown as one of the most efficient tools for overcoming the main reluctances to switch from paper to paperless sampling. Users need to be guaranteed (up to an extent) that their data (at the end it means their work, their time, their effort) are not going to be lost for a system failure. In the first times of osb_pelakamp every single sampled fish data were print in a dot-matrix printer. That was noisy and bulky. Nowadays users get visual confirmation of data storing, periodic whole database backups are done automatically (at least daily, by means of cronjobs) and reports with the data are printed preferably after each sampling event.

4.4.7 Future and challenges

In its present state, despite having been used in production (surveys and lab), the system needs refinement, debugging and many improvements. An almost comprehensive list of the tasks and challenges to be faced are:

- **Document the project,** to make the project attractive for developers and users and, eventually, ease the building of a development community. It involves not only code documentation but user guides, GUI help.
- **Release the code to the public,** so the project become a *real* OpenSource project. It implies not only the code itself but also database schemas, blueprints and electronic schematics of the hardware., etc.
- **Staff training:** Not only for using the system but for deployment and troubleshooting.
- **Software development:**
  - Improve usability of GUI's and code debugging
  - Implementing quality checks thoroughly, specially at early stages of the process.
  - The flexibility of the software, mostly with respect to biological sampling must be greatly improved, so any species could be sampled. In its current state, only certain variables can be sampled. Also sampling protocols should be implemented.
  - Provide translations of the software for, at least, English.
  - **Mantain software up-to-date:** Both PyQt4 and Python 2.7, despite being perfectly usable, are close to their end-of-life. Migration to PyQt5 and Python 3.x has to be tackled. This is probably one of the most demanding tasks ahead.
- **Improve design and ergonomics of the board:** Placing electronics under the board, using smaller domes for the tactile switches, better design of the layout of the measurement scale, improvement of acoustic signaling are some of the possible enhancements. **Hardware:** Serial, Bluetooth and wiFi communications are of easy implementation but battery operation would require a detailed power budget study but it seems that 6 hours autonomy and be reach with not great effort.

4.5 DCS board & Allegro Campagne application

The system being implemented by IFREMER is to utilise an off-the-shelf electronic measuring board system, DCS, from BigFin Scientific. On top of which it has developed a more comprehensive custom application for data management and quality checking.

4.5.1 DCS - BigFin

The Big Fin Scientific Data Collection System[5] (DCS), similar to several boards presented, uses a Linear Magnetic Transducer for registering measurements, housed in a rugged aluminium chassis. Data collected is displayed through its customizable application which runs on Android smartphones or rugged tablets. There is the option to push collected data up to a web based application and cloud account for storage, sharing and reporting functionality.

Although the fish-board is a commercial product and therefore not "open" in the sense of all schematics and code being published, the communication protocol is open and well documented. That is to say, the boards can be added as a "peripheral" to end-user software using simple serial communication commands over Bluetooth or USB. Essentially then the system has proprietary software, but open hardware "integration" or API.

4.5.2 Allegro – Tutti Project, IFREMER

IFREMER carries out approximately 20 fisheries surveys per year on vessels from 20m – 74m staffed by teams of between 2-15 scientists. Since 2013 the development of a new software application, Allegro Campagne, has been undertaken as part of an overall project, Tutti.

Allegro is an open source Java based software, however some underlying tables are specific to, and held by IFREMER. The application has been developed to import data from the BigFin DCS measuring boards, Sylvac Evo electronic callipers and Marel scales (Fig 4.5.2.1). The application then takes over the management, storage, quality control and general reporting of the data thereafter.

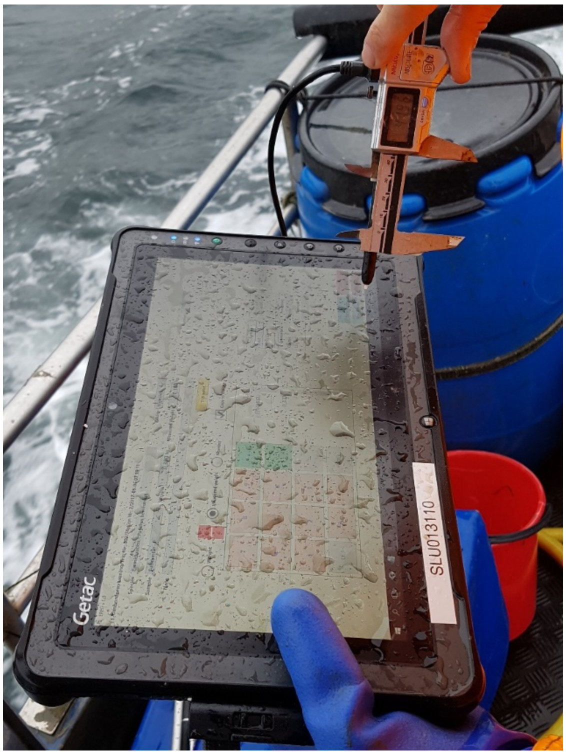
Fig 4.5.2.1 BigFin DCS board being used at sea with rugged Android tablet.

[5] https://www.bigfinscientific.com/documentation/

4.6 E-reg – SLU Aqua

4.6.1 Overview

E-reg is software developed in-house at SLU Aqua that runs on rugged tablets with Windows OS. Designed for data entry by touchscreen and data collection from external electronic tools (digital callipers, electronic scales, GPS, etc.). This approach allows for all types of data entry, not only biological. The application is currently used only on a national scale in coastal surveys and on-board sampling of commercial vessels targeting Nephrops. An in-house developed application was determined to be the best available option to counter future demands on cost efficiency, data quality and requirements of data collection from other components of the ecosystem.



4.6.2 Technology

The application is built to be independent of hardware manufacturer. Different brands of rugged tablets, callipers, scales etc. have all some limitations. For a future safe system it is important to be able to switch product when something better becomes available on the market.

4.6.3 User expectations

The intended users of the system have been a crucial part in the process of the designing the graphical interface for the application. Allowing the end-users into the development process at an early stage has facilitated the final implementation.

4.6.4 Implementation

The initial cost for developing an in-house system is higher compared to most commercially available products. Time to implementation and return of investment is also longer. However, when the system is in place the cost for maintenance is likely to be lower (no licenses to pay). Since it will be developed and fitted to the specific need of the owner it will likely also be easier to implement and give a higher return in efficiency.

4.7 EDC – Marine Scotland

The reader is referred to the case study in Section 5.2.3 for more detailed overview of the RFID boards utilised on Marine Scotland surveys.

4.8 RFID Tag & other technologies

As referenced in the introduction to Section 4, one of the earliest technologies employed for electronic measuring boards in fisheries was the bar code. This was the technology implemented also by CEFAS for their original Electronic Data Capture (EDC) system, still currently in use at the Marine Institute (MI), Ireland. The MI procured the bar code EDC boards in 2000 and have measured almost 3 million fish with it on their annual Q4 IBTS survey with periodically another trawl survey annually over that period. With the average being c.950 fish per haul being measured the system proved itself to be stable, cost-effective and quite efficient. The technology is dated now however and has been replaced by CEFAS with the RFID version of the EDC system now also deployed by Marine Scotland (see Section 5.2.3 below).

The strengths and weaknesses of bar codes, RFID tags and some other technologies were discussed and a summary is given below.

Bar Codes - The original light pen/barcode solution had technical weaknesses:

- Lenses were easily damaged, although straightforward to replace (pens now much cheaper ca. €90 and more robust).
- Large amounts of opaque material on the board made the system less responsive (cuttlefish/squid ink or sticky mud).
- Repetitive Strain Injuries (RSI) became a concern in intensive sampling operations.
- They were not IP67 and pens required in-house modifications to waterproof and connect (can use directly as USB but connections not waterproof/secure).
- The wands could however also be used to read barcodes from custom labels on otolith boxes or other sample labels of course.

Various routes explored in the search for a replacement technology for Barcodes before selecting RFID as the solution (Note, the reasons below relate to technology the time and latest technologies may not have the same issues):

Magnetic/Inductive reed switches:

- weight of board once populated plus requirement for additional power supply as well as one for a tablet PC/laptop to receive the data. This is ok for semi-permanent installations e.g. a survey vessel, but cumbersome for short term or mobile sampling operations
- concern over vulnerability of the components in the board especially in mobile operations
- lack of flexibility and extendibility to other measuring tasks
- In-the-field repairs were not straightforward

Laser/Infra-Red scanners:

- Expensive
- Not portable
- Contamination on scanner housing and board caused accuracy issues

- Still required a method of user intervention to indicate a measurement should be taken
- In-the-field repairs (other than cable breaks) were impossible

RFID is not a perfect solution:

- RFID readers can be power hungry
- In the event of a tag failure, replacement of an individual tag is quick and simple, but, depending on the method used, it can take time for fixing compound to cure and for the board to be back in action
- There is a balance between having an acceptable reading sensitivity and avoiding spurious reads
- There are still sensitive electronics at the 'dirty end' of the sampling process

RFID was eventually selected by CEFAS as the hardware successor to Barcode boards given the technical advantages at the time gave significant benefits over the other technologies available. As the system was developed, other advantages came to light however.

The function of any tag on a board can be context sensitive, for example in meta-data collection mode (haul, species, category) a tag can be designated with a species code, while in measuring mode the same tag can be designated as a length, 15cm for example.

Where sampling was mostly undertaken with callipers, we could interface the calliper directly to the system and just supply a small board (ca. 20cm x 10cm) with enough tags to enable entry of the meta-data for the sample. (implemented in port sampling system).

Given some base 'libraries', custom software and board implementations could be delivered with much shorter time-scales (an idea not considered previously).

And, of course, some disadvantages, mostly of the hardware/support system rather than RFID.

- Windows is not an ideal OS for this type of work
- Automatic updates
- Viruses
- System overhead, a lot of the system requirements are there to support Windows, but the system doesn't actually use many features of Windows

Users found the number of cables (power, wand, headphone, display) problematic in mobile operations. Selection of materials and construction method for the boards into which the RFID tags are embedded has to be optimised.

4.9 Other electronic capture systems

Finally, a couple of alternate systems were highlighted during the workshop, but no participants had much experience with them so are mentioned here for reference only.

- Bioscribe[6] is an inductive (touchscreen) technology board, developed by Hallprint, New Zealand. It can be used stand alone or attached to rugged PC.

[6] https://www.hallprint.com/fish-tag-products/2014/8/26/ogborjudblbsvxgn1y54jdf0mpqknd

- Scielex[7] produce a simple linear sensor measuring board with limited functionality, but lower cost than some of the alternatives and simple to use.

Natural Resources Institute Finland (Luke) made contact during the workshop once aware of the meeting. They provided some details of their RUFCO fish measuring system which integrates an electronic measuring board and weighing scales to acquire both data types at the same time. One person can measure and weigh catch samples fluently with RUFCO system which has been in production since 2005. The RUFCO system for measuring samples from commercial catch consists of RUFCO DL2 data logger with software (latest update 11/2016) and an electronic measuring board situated on top of a bench scales. The RUFCO software used on the BIAS survey since 2014 is designed for networked workflow: 3 RUFCO devices in wet lab (length distributions for all species, length stratified subsampling for herring and sprat) and 5 RUFCO data loggers in dry lab to punch in the individual data.

[7] http://scielex.com.au/products/products.asp?ID=3&Category=

4.10 Overview Table of current technologies

	Smartfish	OpenSMB	Osb_pelakamp + icrOS	Allegro	Big Fin Scientific	swedish solutions E-reg	scotland: CEFAS rfid system
type	measurement board	measurement system	measurement system	measurement system	measurement system	part of data collection system	workflow system
status	production since	beta-testing	production	production	commercial product	production 09/2017	production 2013
hardware	x	x	x	(Big Fin)	x	waterproof tablet	passive board from cefas
sensor	magnet	magnet	tactile keyboard	-	magnet		pit tags
embedded computer	-	Beaglebone	Adurino / RaspPI	-	Propeller (Hardware)		PC
power	external battery	internal battery	PoE	-	internal battery		cable
data connection	bluetooth	WLAN / LAN	RS232 over LAN	virtual serial	Bluethooth & USB		network
software	external (Windows)	internal (Linux)	external (Linux)	external (Windows)	external (Android)	windows	windows (XP)
database	MS SQL-Express	PostgreSQL	PostgreSQL	?	Android & Cloud solution	MS SQL-Express	
data exchange		API		generic format file	Generic formats (CSV, JSON)		
multiple devices for one task		yes	yes		Software – Yes Hardware - No	yes	yes
basic data validation	yes	yes	yes	yes	yes	yes	yes
extended data validation	yes	backoffice	yes	yes	not for european surveys	backoffice	yes/backoffice
web frontend	no	yes	no	no	yes	no	yes
link to other devices	windows software	direct (LAN, serial)	?	windows software	Software – Yes Hardware - No	depends on connectors	

5 Data quality control & Management

5.1 Introduction

This section discusses the quality control and interoperability of collected data. In this context quality control is used in the more general sense of overall quality management including both QA and QC highlighted in Section 3.2. Three short case studies based on existing Data QC systems are presented and then some of the wider issues of data interoperability are discussed.

5.2 Data QC Case Studies

The case studies presented for FishTrawl/RoME, the IGFS survey, and the Scottish demersal survey systems aim to answer the following questions:

- How are data collected?
- Are sampling targets set and measured? (This should be related to the accomplishment of the survey. etc)
- What specific QC processes are applied to the data?
- Is historic or spatial data used during the QC processes (e.g. comparing measured data to historical ranges or with other data collected in the same area)?
- How is data displayed or reported?
- What data formats are used?

5.2.1 FishTrawl /RoME

5.2.2 Background to MEDITS programme

The MEDITS project started in 1994 within the cooperation between several research Institutes from the four Mediterranean member States (Spain, France, Italy and Greece) of the European Union. The target was to conduct a common bottom trawl survey in the Mediterranean in which all the participants were using the same gear, the same sampling protocol and the same methodology.

At the beginning of the project, one of the main challenges was the adoption of a common standardized sampling methodology. The basic protocol has been adopted early in 1994, just before the first international survey. This protocol included the design of the survey, the sampling gear (feature and handling), the format of the information to be collected, the management of the data and the production of common standardized data analysis . Before the first survey, the common protocol was issued as “Manual of protocols” and published later on (Anon. 1998) to ensure a wider distribution. The protocol has been subsequently amended when necessary to take into account the experience gained during the surveys and the scope in terms of new geographical areas and information to be collected. In 2012 revision 6 was issued, which included substantial modifications to the MEDITS manual, though not affecting the main characteristics of the protocol regarding the sampling scheme, methods and gear. This new version included changes in the list of target species and groups, which were both expanded. In addition, the protocol for otolith sampling and biological parameters measurements was included, while adjusting the storage data formats accordingly. Furthermore additional standardized checks in open source routines were prepared and distributed. Even then monitoring of marine litter was included in the protocol of the MEDITS survey.

The last version of the manual[8] is available at the following link: http://www.sibm.it/MEDITS%202011/principaledownload.htm.

In the manual the common data exchange formats are described in details:

- TA - haul data: start and end position (lat, long, depth), date, time, vertical and horizontal net opening, etc.;
- TB – catch data: information on the number of individuals and total weight by species for each haul;
- TC – length data: aggregated number of individuals by length, sex and maturity by species;
- TE – individual data on weight, sex, maturity stage and age;
- TL – litter data

These relational tables are shared among all Countries involved in the MEDITS survey and have been defined since 1994, when the programme started and have been further detailed in the following years. This data exchange format is used to provide the data to the main end-user (European Commission through DCF).

Each laboratory involved in MEDITS organized its work for MEDITS data collection by developing in-house applications and software to manage these exchange formats.

Since 2011, when the RoME[9] routine was developed for the first time, data checks should now be performed using this shared tool, which is written in R and open source. Updates of the routine are freely distributed to all the MEDITS partners and each new release issued is accompanied by a recommendation at JRC level.

The update of the reference list of the species is annually presented during the International Coordination meeting and a single scientist is in charge of this repository, supported by a working group.

Sampling targets are set according to the DCF National Programs following the methodology of the MEDITS Handbook. Each year the MEDITS group, formed by participants from all the countries involved in the survey, meets (generally in March, a couple of months before the start of the survey) and presents the results of the past survey year, including presentation of the time-series of abundance and biomass indices of the most important taxa. The group discusses the achievements, highlights relevant issues if necessary and provide recommendations, which are also reported to the following RCMMed and BS. The timing ensures that data can be reviewed and discussed before the delivering of the dataset to the data call which is generally issued in June.

5.2.2.1 FishTrawl web application

FishTrawl is one of the more complete applications designed to specifically manage MEDITS data collection; it is a web application developed within the MAREA Specific Contract 9 (Horizontal Service) envisaging different functionalities:

- data storage, through a relational database in MySQL;
- data entry, by using an interactive Graphical User Interface (GUI) including a set of syntactic checks;
- data import and export, through specific exchange formats;
- extraction of subsets of data;

---

[8] MEDITS-Handbook. Version n. 9, 2017, MEDITS Working Group : 106 pp

[9] Bitetto et al. 2017. RoME_1.3.2_User_Manual – 2017. Link: http://www.sibm.it/MEDITS%202011/principaledownload.htm

- data check (RoME based);
- data analysis (abundance and biomass indices, maturity parameters, length-weight relationships, etc.).

The data format used by Fishtrawl is the MEDITS common exchange formats. The application can also work as locally installed software as well as accessed using a website.

The samples are collected and measured on board during the survey or successively in laboratory; all the biological data are recorded on paper by most of the countries involved in the programme, then the data are entered in specific databases and/or software (e.g. Fishtrawl). Only IFREMER (Gulf of Lyon) currently carries out the collection on board by electronic data capture.

The quality checks carried out by Fishtrawl during the data entry are:

Haul

- Survey, Country, Area, Vessel code, Gear code, start and end quadrant, bottom shape, stratum code, codend closing, part of the codend, observations, recorded species, geometrical precisions, instrument, method, course, weather conditions, sea conditions, wind direction, shooting and hauling depth, distance, wing opening, bridles length, warp length and diameter, bottom temperatures and salinity are validated against the related tables in data configuration;
- The average depth is checked against the minimum and maximum depth of the stratum;
- The haul number is checked to be unique inside the chosen area and survey;
- The start time is checked to not be later than the end time;
- The coordinates are checked to be in the selected quadrant and in the selected GSA;
- A warning is issued if the date is not in the same year as the survey name.

Catch

- The species is checked against the reference list;
- The total number of individuals caught in the haul must be equal to the sum of the total numbers per sex;
- A warning is issued if the total numbers are not equal to the sum of biological data.

Biological Data

- Sex is validated against the list [F, M, I, N];
- Maturity Scale is validated against the maturity scale codes table in data configurations;
- A sex must be selected if the species group is G1;
- The length must be between the min and max length defined on the basis of historical data;
- If the sex is I or N the maturity must be either 0 or ND and the sub maturity must be either blank or ND;
- If the sex is F or M the maturity and sub maturity are validated against the maturity scale stages defined inside a maturity scale code in data configuration.

Individual Biological Data

- Sampled is validated against the list [Y, N, NR];

- Read is validated against the list [Y, N, NR];
- Hard Structure is validated against the defined hard structures table;
- The individual weight cannot be lower than 0.1 g and higher than 999999 g;
- The total number of individual biological data can't be higher than number in biological data.

FishTrawl includes a series of checks and tests when a file is imported in one of the MEDITS exchange formats:

Checks on TA (haul data) and validation

- All the fields, except to HYDROLOGICAL_STATION and OBSERVATIONS, must be not empty for valid hauls;
- WING_OPENING, WARP_DIAMETER and VERTICAL_OPENING fields have to be not equal to 0;
- The DURATION, SHOOTING_TIME and HAULING_TIME fields have to be consistent;
- The DURATION and DISTANCE fields have to be consistent;
- The field BRIDLES_LENGTH can assume value 100 between 10-200 m of depth or 150 between 200-800 m;
- The difference between start depth and end depth should be not greater than 20%;
- Start depth and end depth of each haul should be in the same stratum;
- The shooting quadrant and the hauling quadrant should be the same;
- The WING_OPENING and VERTICAL_OPENING fields have to be expressed in dm;
- The distance has to be consistent with the coordinates at the start and at the end of the haul (validation algorithm included);
- Check of the dictionary of specific fields (e.g. validity can be only V or I) ;
- There must not be duplicated records.

Checks on TB and validation

- All the fields must be not empty;
- If total weight is different from 0, total number must be different from 0 (only if the category of the species is different from "E") and vice versa (for all faunistic categories);
- Correctness of species MEDITS code and faunistic category according to the reference list;
- Total number must be equal to the sum of the numbers by sex;
- Check of the dictionary of specific fields;
- There must not be duplicated records.

Checks on TC and validation

- Correctness of LENGTH_CLASSES_CODE;
- All the fields must be not empty;
- Consistency of maturity stages, according to the faunistic category, sex and species;
- Check of the dictionary of specific fields.

Checks on TE and validation

- All the fields, except to OTOLITH_CODE, must be not empty;
- All the specimens present in TE must be present in TC;
- Check of the dictionary of specific fields.

FishTrawl allows quality checking to take advantage of historical data for the checks on the plausibility of length measures and of mean weight using appropriate ranges by species.

The plausibility of the hauls position can be visually checked in real time (Figure 1).

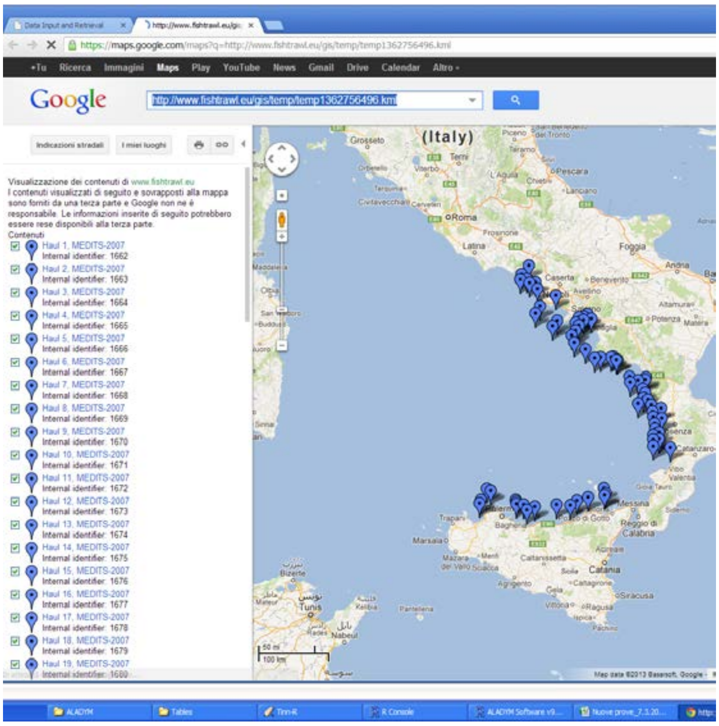


Figure 1 Check of the plausibility of haul positions in Fishtrawl

Other checks are related to the consistency between depth and bottom temperature (Figure 2) and between the length and the maturity stage of the sampled individuals (Figure 3).

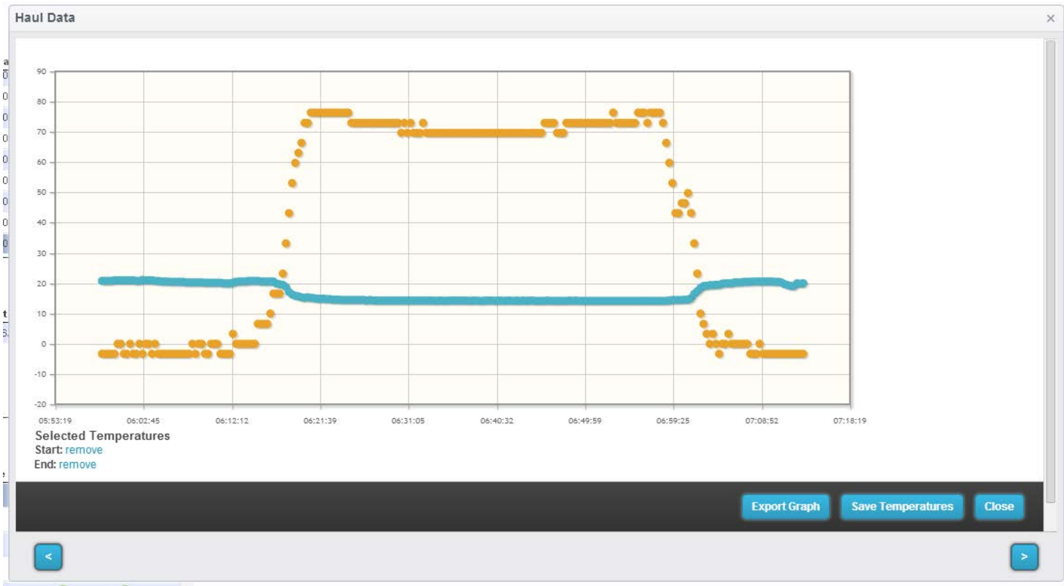


Figure 2 Consistency check between depth and bottom temperature

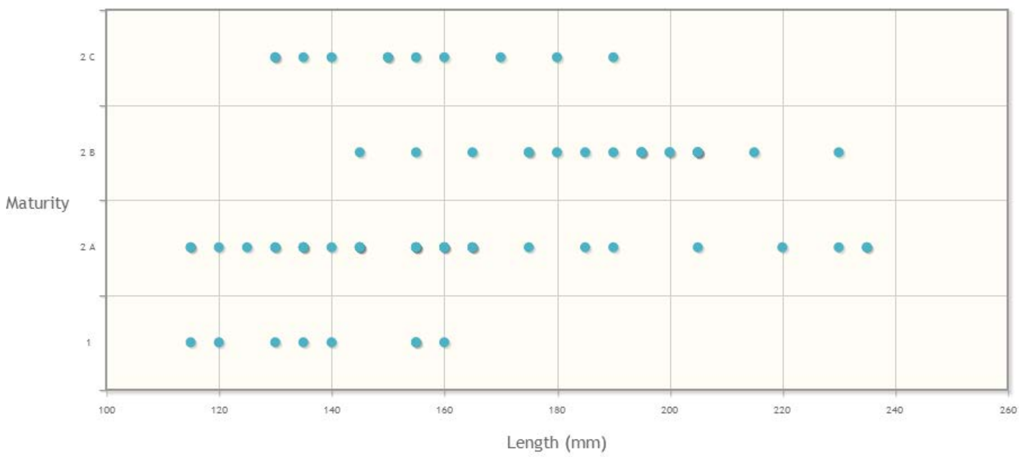


Figure 3 Consistency check between length and maturity stage

The completeness of the information among the tables (TA, TB, TC, TE and TL) is verified by a dedicated tab (Figure 4):

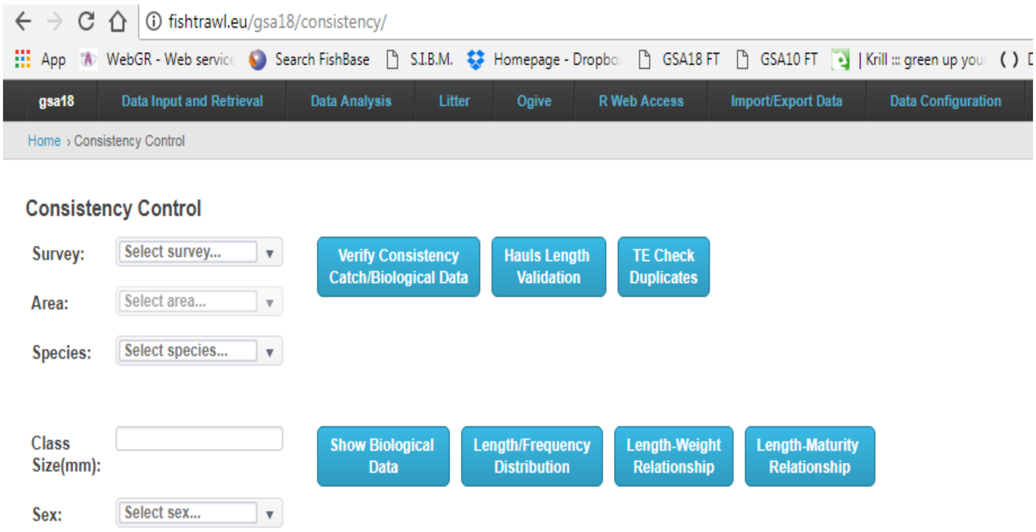


Figure 4 Consistency Control Tab implemented in Fishtrawl. The same tab is also used to display the data (with the buttons showing biological data, length–frequency distributions, etc.).

5.2.2.2 RoME data quality tool (R code for checking MEDITS Survey data)

RoME (Bitetto *et al.* 2017) has provided a common standardized tool within MEDITS coordination to perform data checks since 2011. RoME 1.3.2 is distributed as a package to be loaded directly in an R console and can be downloaded directly from MEDITS website: http://www.sibm.it/MEDITS%202011/principaledownload.htm.

The JRC (Joint Research Center) has the responsibility to check data sent by Member States for the European Commission, and has based its MEDITS data quality on RoME checks[10].

---

[10] *"MEDITS data quality. Since December 2012 JRC has developed quality checks with SQL routines in the MEDITS Postgres database of JRC to do cross table consistency tests and conformity to the survey manual checks. In total 26 routines where developed, these share a similar*

RoME works directly on the common exchange data formats (TA, TB, TC, TE) in .csv and returns a set of visual checks and a log file containing the result of every check.

- *What specific QC processes are applied to the data?*
- *Is historic or spatial data used during the QC processes (e.g. comparing measured data to historical ranges or with other data collected in the same area)?*

The order of the checks in RoME was implemented in a defined sequence to avoid cascade errors due to the correction of a previous error. No automatic correction is implemented in RoME and the software stops if an error occurs; then the user has to correct the error and run again the code to continue with the other checks. This is to increase the awareness of the user and to ease the detection of the more recurring errors.

RoME inspired all the checks listed in the Fishtrawl. Below additional checks performed using RoME are listed:

TA:

- Relationship between depth and warp length;
- Relationship between warp length and wing opening.

TB:

- Consistency between the total number of individuals and the total weight in the haul (Figure 5).

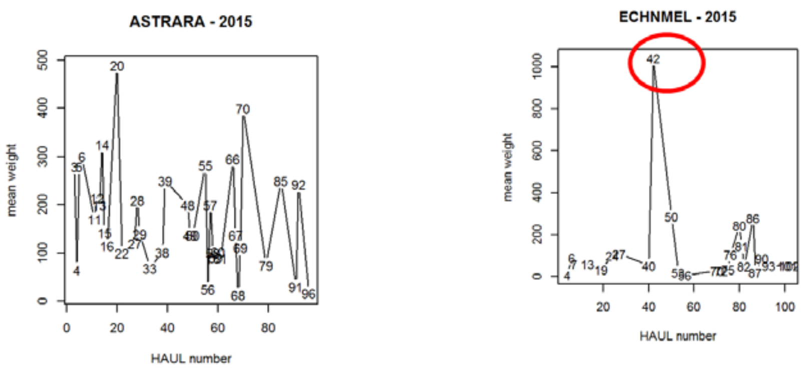


Figure 5 RoME: qualitative check to test the coherency between number of individuals and total weight by species in each single haul

TC:

- Detection of anomalies in staging according to literature information on spawning season and size at first maturity;
- Consistency of the subsamples (>10%);
- Use of information of sex inversion for hermaphrodite species.

Cross checks among the tables

---

philosophy to the ROME routines (Spedicato and Bitetto 2012) and when ROME is used before data upload the JRC routines correctly show no error patterns".

- All the hauls in TA must be in TB.
- All the hauls in TB, TC and TE must be in TA.
- All the target species in TB must be in TC.
- All the species in TC and TE must be listed in TB.
- All the hauls in TC and TE are in TB.
- In case of subsampling in TC, the Total number and the number per sex in TB must be raised correctly.
- TA, TB, TC and TE must have the same year and area.

The data are displayed through the graphical output produced by RoME. A specific check on individual data returns a summary table about the sampling coverage (Figure 6).

Species	LENGTH_CLASS	LENGTHS	WEIGHTS	OTOLITH
MERL MER	50	2	2	2
MERL MER	70	5	3	3
MERL MER	75	8	2	2
MERL MER	80	17	3	3
MERL MER	85	32	2	2
MERL MER	90	78	3	3
MERL MER	95	133	2	2

Figure 6 Summary table containing information of the number of weights and age data collected in order to evaluate the sampling coverage (RoME).

References

Bitetto I., Facchini M.T., Spedicato M.T., 2017. RoME (version 1.3.2): R code to perform multiple checks on MEDITS Survey data (TA, TB, TC, TD, TT and TE files). http://www.sibm.it/MEDITS%202011/principalereports.htm

Casciaro L., Bitetto I., Spedicato M.T., Lembo G., Chieti V., Chieti A., Segato C., 2015. A Scientific Surveys Web Platform. USER MANUAL, Release 1.0 – April 2015

5.2.3 Ireland IGFS survey data collection

This case study discusses some of the quality checks that are done on the Irish Ground-Fish Survey (IGFS) but it also focuses on the ideal case scenario i.e. where and when these checks should ideally take place. This section focuses on the checks for data collected in the wetlab / fishroom where all sorting and sampling of the catch is carried out. However, similar principles apply to fishing operations data collected on the bridge.

Some general points:

- All checks / feedback should ideally take place as the data are entered. However, it is also important to do some of the same checks at different stages, e.g. once the data from a haul is complete, at the end of a working day or shift and at the end of a survey. This is because many outliers only become apparent in the context of the distribution as more data are collected during the survey. In addition, there may not be enough time to fully evaluate every check during every sampling event.

- Graphical checks generally convey information faster than numerical, tabular or report validation. It is easier to spot outliers on a plot than deal with too many or too few warnings of values that are outside a specified range.
- Many checks rely on a certain amount of redundancy in the data that is collected, e.g. to quality check the distance between the shoot and haul position we can divide the vessels speed by the haul duration. Having either speed or duration of course we could simply derive one value from the other. It is important to allow this type of redundant data to be collected and not make the data collection so efficient that no 'useful' proxies exists that might help in overall quality management.

Three main groups of fishing survey data are collected on IGFS

- Fishing information: station positions; gear parameters and environmental parameters (weather, sea conditions etc). Data are collected on the bridge.
- Catch information: catch weights by species (and optionally: size category and sex). Data are collected on a central computer in the wetlab.
- Sample data: Length and biological data (individual weight, sex, maturity, etc). Data are collected in a number of measuring boards in the wetlab.

Wetlab data checks

- Wrong haul number. Currently there is no formal QC for this.
  - Before catch data are entered, the user has to select the current haul number (incremental number). Ideally this should be linked to haul information entered on the bridge, so the user can see the haul and shoot times already entered on the bridge. Also, the user should receive a warning if they select a haul number for which catch data already exists or a haul number for which no bridge data exists.
  - The measuring boards could be linked to the computer on which the catch data are entered and only allow data to be entered into the 'current' haul number. The user should be warned when they try to enter data for a haul that already has sample data or for which no catch data exists.
- Wrong catch weight. Currently there is an option for the user to check this but it is not enforced. It can be very difficult to identify mistakes if the catch is subsampled.
  - Once all catch data are entered, the user could receive feedback with the total weight of the catch (in kg and/or boxes) as well as a breakdown of the species in the catch (pie or bar plot; e.g. Figure 7). The software should ask the user if it tallies with their impression of the catch or else force users to make an estimate of the bulk catch before sorting.
  - The user could also receive feedback on the raising factors resulting from subsampling. This information can be displayed graphically, in combination with the previous point (Figure 7).
  - Validation on the catch weights entered. For large catches, a number of weights will be entered as the catch is being weighed in individual fish boxes. A warning should be given if any of these exceeds the maximum weight of a full fish box (40 or 50kg).
- Wrong species / size category / sex. Once a sample has been weighed, the sample weight is entered on a central computer in the wetlab, while the measurements take place on individual electronic measuring boards. It is possible that the sample details (species, size category, and sex) on the central computer do not match those entered into the measuring boards.

  - The measuring boards could only allow users to enter samples for which weights already exist.
  - The measuring boards should warn users if they enter sample details that other users have already entered at a different measuring board.
  - The samples could be labelled with a barcode or equivalent.
- Wrong Length. Currently no validation.
  - A warning could be given by the measuring board software if a length measurement falls outside an expected range for a species.
  - After a sample has been measured, (or while the sample is being measured) the user should see a length frequency distribution plot. This will allow the user to identify outlying lengths. It will also show the user whether data should be collected in mm, ½cm or 1cm size classes.
- Wrong sample weight
  - The sample weight can be compared to an estimated weight of the measured fish from a length-weight relationship. If this is not available for all species, then Fulton's condition factor (K) can be estimated by assuming allometric growth:

$$K = \frac{100 \cdot W}{\sum L^3}$$

  - Where W is the sample weight (in grammes) and L is the length of individual fish in the sample (in cm). If the value of K is plotted for a number of samples of the same species, outliers can be identified (see Figure 8). This should be evaluated directly after the sample has been measured and perhaps again (by a central person) after all samples have been measured and finally, at the end of the survey.
- Wrong individual weight
  - This can be checked in the same way as the sample weights: either from a length-weight relationship or Fulton's K. Similar plots to Figure 8 can be presented to the user straight after the weight measurement; after all samples have been measured and at the end of the survey.

Other feedback / reports for the wetlab

- Sample size
  - The 'ideal' sample size depends on the ultimate use of the data but there is a general rule-of-thumb that the number of fish in a sample should be around 10 times the number of size classes in the sample (Gerritsen et al., Fish Bull 105(1), 2007). This can be difficult to judge beforehand but users can train themselves by examining feedback of samples where too many or too few fish were measured. This feedback can also include information on how size categories were used. See Figure 9 for an example - in that example the haddock catch was divided into three size categories. The smallest size category (turquoise) had a raising factor of 3.1, resulting in 110 fish measured. This category spans 9 size classes, so the 'ideal' sample size is 90 fish. Category 2 (white) was the most abundant and was most heavily subsampled; the raising factor was 20.5 with 148 fish measured. Again this is close to the 'ideal' sample size of 120. The final category (pink) was least abundant; all fish in this size category were measured. This is an

example where the catch was very efficiently handled, resulting in considerable time savings.

- Otolith reports.
  - o If otoliths are collected in trays (rather than envelopes) it is possible to visualise this tray to assist in checking that all otoliths are in the correct cell or tray and to help fix mistakes. Figure 10 and Figure 11 illustrate this.

Figures

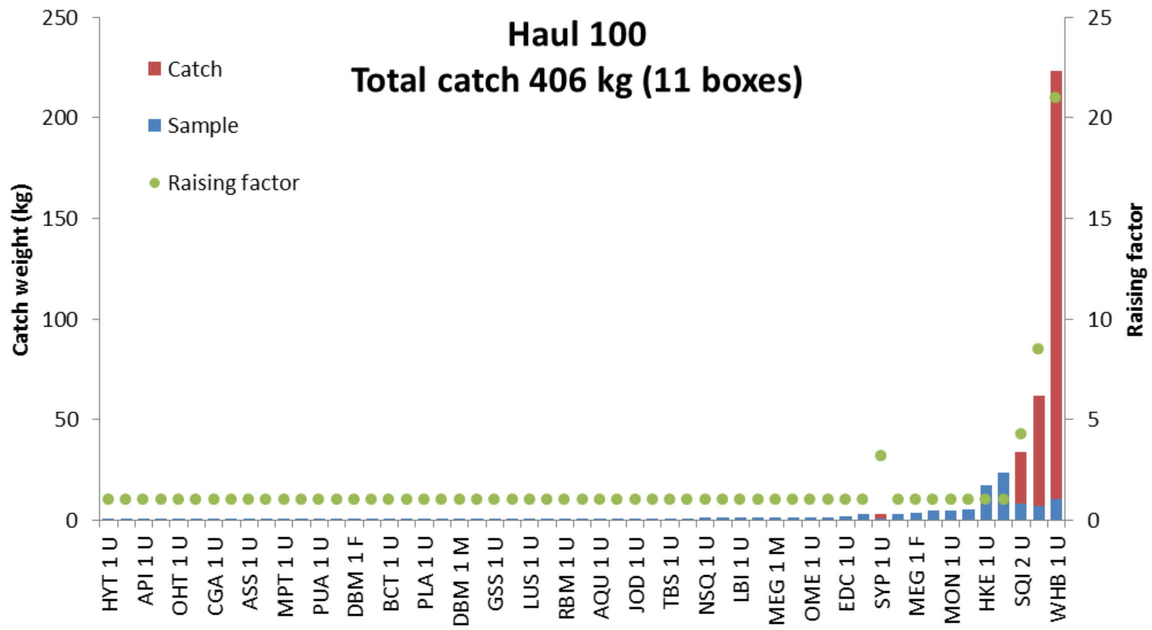


Figure 7 Example of feedback on the catch composition (to allow the user to identify mistakes in the catch weights) as well as the sample size and resulting raising factors (an excessive raising factor could also point towards incorrect catch weights).

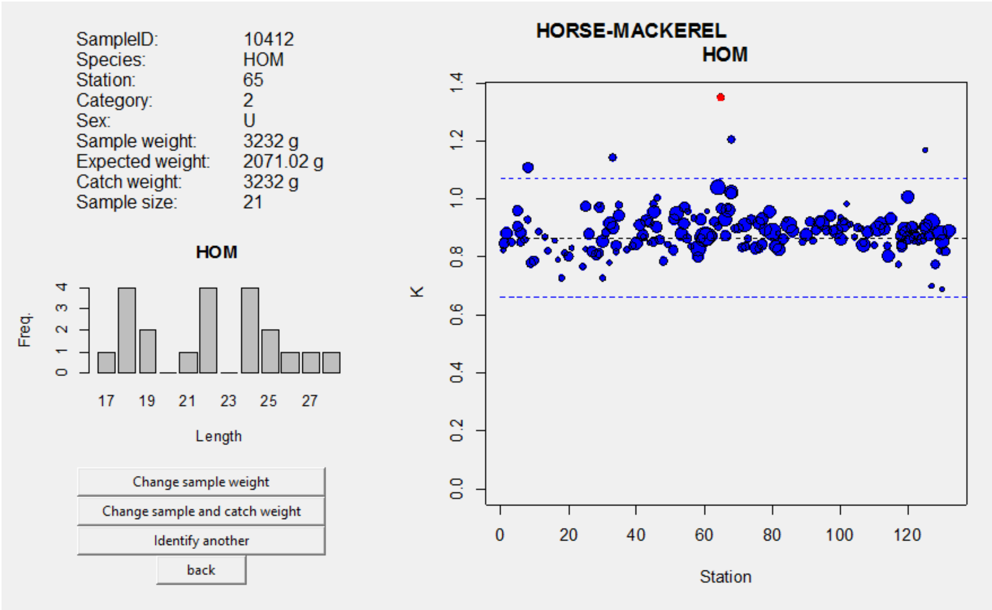


Figure 8 Fulton's condition factor (K) for length samples. The red point is an outlier; based on the mean value of K, the expected weight for this sample was 2071g while the recorded weight was 3232g.

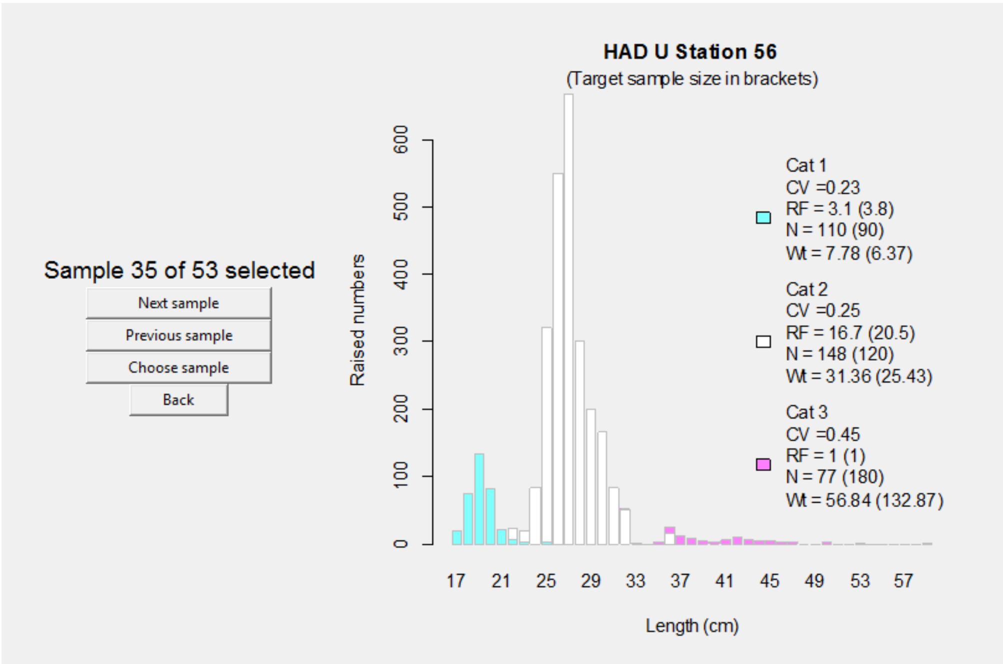
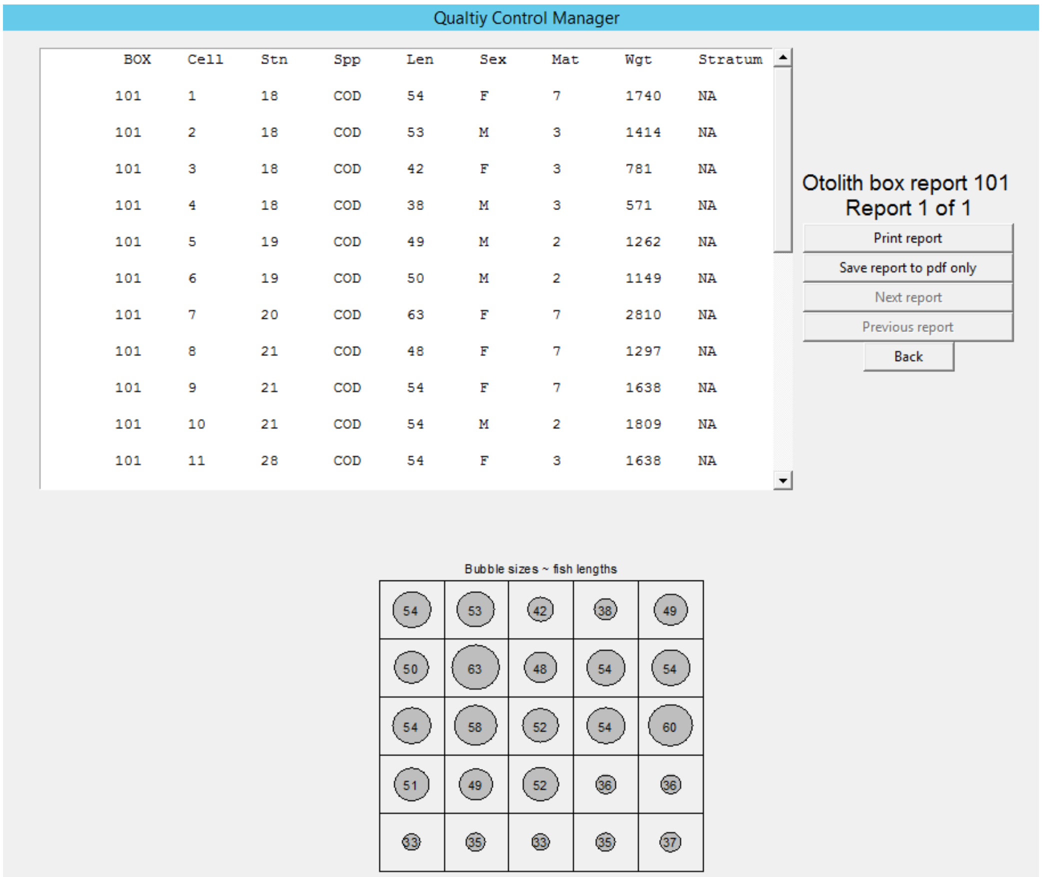


Figure 9 Feedback on the use of size categories and sample sizes.



BOX	Cell	Stn	Spp	Len	Sex	Mat	Wgt	Stratum
101	1	18	COD	54	F	7	1740	NA
101	2	18	COD	53	M	3	1414	NA
101	3	18	COD	42	F	3	781	NA
101	4	18	COD	38	M	3	571	NA
101	5	19	COD	49	M	2	1262	NA
101	6	19	COD	50	M	2	1149	NA
101	7	20	COD	63	F	7	2810	NA
101	8	21	COD	48	F	7	1297	NA
101	9	21	COD	54	F	7	1638	NA
101	10	21	COD	54	M	2	1809	NA
101	11	28	COD	54	F	3	1638	NA

Figure 10 Example of an otolith report with a visual representation of the otolith tray where the size of the circles corresponds to the length of the fish. This allows users to check that otoliths are in the correct tray and can help fix mistakes.

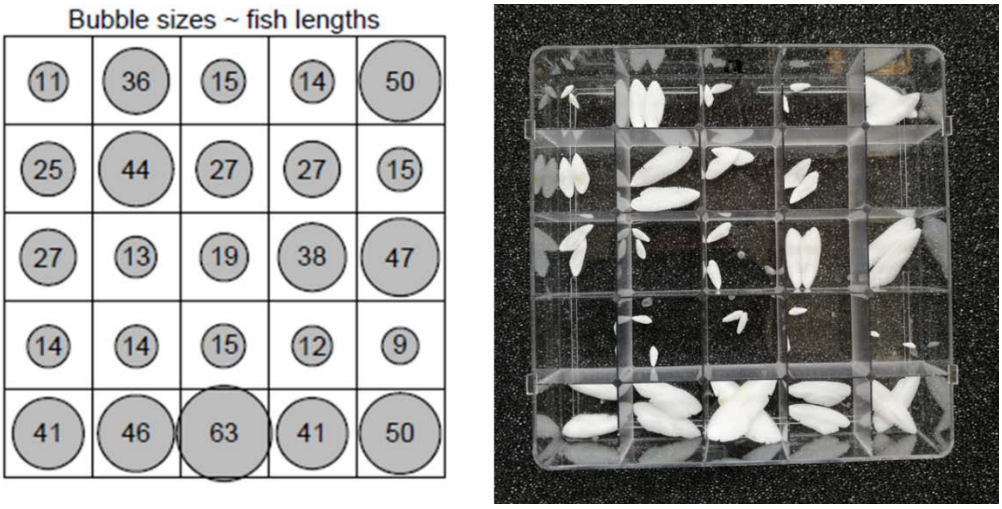


Figure 11 Graphical representation of an otolith tray (left) and the actual tray (right). It is easy to see that the otoliths correspond to their expected location and relative size the tray.

5.2.4 Scotland demersal survey data collection

How are data collected?

Scotland utilises an electronic data collection (EDC) system on-board the MRV Scotia on all IBTS surveys, the Rockall Haddock survey, the Herring Acoustic survey and the Scotia leg of the SIAMISS survey. The EDC system and database software was purchased from CEFAS, England in 2012 and has been utilised since with only minor changes to the hardware. It is comprised of 6 Windows PCs with measurement software to which RFID readers are attached. The measurement boards have embedded RFID tags to correspond with a measurement. The system allows length and biological measurements such as weight, sex, maturity along with survey specific parameters, to be collected electronically.

Are sampling targets set and measured?

Sampling targets are set using an EDC survey configuration program. This usually is completed pre-survey however the targets can be modified while the survey is underway. The program provides a GUI to generate reference tables that are then uploaded to each unit via a network before being utilised by the EDC to recognise sampling requirements. Biological sampling targets that can be applied for any species include length interval targets e.g. 1/cm/haul, stratum based targets e.g. 10/area and size range based targets e.g. 1/cm for 0-39cm, ≥40cm 2/cm. For a species that biological data are required, the system will ask each time the species is about to be processed either what percentage of the target is to be fulfilled or how many individuals per measurement interval are required.

What specific QC processes are applied to the data?

During the sample processing stage the software displays a number per length bar chart to visualise the measurements already obtained. This can be useful in determining when to subsample a species and also highlights lengths recorded accidently e.g. an 80cm herring. When a biological sample is required for an individual the software requires a whole weight along with the sex and maturity of the fish before the record can be saved ensuring that the most basic biodata are recorded. When all samples have been processed the measurement and biodata are downloaded onto a central PC. Here it has to match catch weights that were recorded pre-processing. If an incorrect species, gear type, haul number, sex etc. has been entered on the EDC it is flagged up at this stage. After this stage the data are uploaded to the ships server where a web-based frontend is used for the QC processes.

The web-based frontend has a multitude of checking tools for both metadata and measurements. QC checks are performed on date and time, geographical positions compared with distance towed/time towed, haul duration, depth range between deployment and recovery, gear parameters including door spread, wing spread, headline height plotted against each other and species catch weight compared with calculated weight (either survey specific or combined survey historical data). QC checks on biological data include species and sex specific length-weight charts, whole weight over gutted weight and length-at-age (after age data has been added).

Is historic or spatial data used during the QC processes (e.g. comparing measured data to historical ranges or with other data collected in the same area)?

Historical data are used during the QC processes to compare catch weight with a calculated weight for the numbers measured/raised per species per haul. This can also be survey series specific to take into account the variations in fish weight throughout the year. The ICES gear parameter relationships are used to highlight any issues that occur with the net parameters by plotting the relationship against data from the net monitoring system.

How is data displayed or reported and what data formats are used?

The survey data are primarily displayed using the web-based frontend as discussed earlier however the data can be accessed to edit using the FSS software provided by CEFAS. The reporting format such as the DATRAS format is generated using R scripts. The reporting of the data is very flexible due to the use of R scripts to deliver the data in the format that is required.

5.3 Data Interoperability - Reference Data

One of the key requirements for data interoperability is that a common vocabulary is used – this can be enabled by the use of standard reference data. Using standard reference data has a number of benefits:

- It ensures heterogeneous databases can still use a common vocabulary – this makes data sharing and integration easier.
- The work of maintaining the accuracy of the reference list is removed from an single institute or country – standard reference lists will also probably be more accurate than local reference lists.

5.3.1 Species

There are a number of marine species reference sets available. Some of the most well-known are described below.

5.3.1.1 World Register of Marine Species (WoRMS)

WoRMS contains details of over 240,000 marine species at the time of writing. It was built through combining the Aphia species database developed by the Flanders Marine Institute (VLIZ) with other authoritative species databases – it is hosted by VLIZ and is continually updated and edited by an international steering committee and editorial board. Aphia contains valid species names, synonyms and vernacular names, and extra information such as literature and biogeographic data.

WoRMS can be accessed via a web interface (http://www.marinespecies.org/); users can request a complete copy of the database; and it can also be accessed by web services (http://www.marinespecies.org/aphia.php?p=webservice).

5.3.1.2 ASFIS List of Species for Fishery Statistics Purposes

The FAO Fisheries and Aquaculture Statistics and Information Branch (FIAS) collates world capture and aquaculture production statistics at the species, genus, family or higher taxonomic levels in 2 269 statistical categories (2015 data) referred to as species items. The 12,721 species items were selected according to their interest or relation to fisheries and aquaculture. For each species item stored in a record, codes (ISSCAAP group, taxonomic and 3-alpha) and taxonomic information (scientific name, author(s), family, and higher taxonomic classification) are provided. The data can be downloaded via the web (http://www.fao.org/fishery/collection/asfis/en) or searched and retrieved online through the FAO Aquatic Species Portal (http://www.fao.org/faoterm/en/).

5.3.1.3 FishBase

FishBase is a global database of finfish that currently includes details about 33,600 species at the time of writing – it can be accessed at http://www.fishbase.org. FishBase shares its taxonomic information with WoRMS.

5.3.2 Gears

5.3.2.1 International Standard Statistical Classification of Fishing Gear (ISSCFG)

The International Standard Statistical Classification of Fishing Gear (ISSCFG) was originally adopted during the 10th Session of the CWP (Madrid, 22-29 July 1980). The revised Classification - ISSCFG Revision 1 (Annex M II) - has been endorsed and adopted for CWP Member's implementation by the CWP at its 25th Session (Rome, 23-26 February 2016)[11]. The ISSCFG Revision 1 gear list can be found at http://www.fao.org/fishery/docs/DOCUMENT/cwp/handbook/annex/AnnexM2fishinggear.pdf

5.3.3 Métiers

A métier is a group of fishing operations targeting a similar species assemblage, using similar gear, during the same period of the year and/or within the same area and which are characterised by a similar exploitation pattern. The EU Data Collection Framework (DCF) defines métiers according to a hierarchical structure using six nested levels: level 1, activity (fishing/non-fishing); level 2, gear class (e.g. trawls, dredges); level 3, gear group (e.g. bottom trawls, pelagic trawls); level 4, gear type [e.g. bottom otter trawl (OTB), bottom pair trawl]; level 5, target assemblage based on the main species type [e.g. demersal fish vs. crustaceans or cephalopods]; level 6, mesh size and other selective devices. An example of métier notation is "OTB_DEF_70-99_0_0" which represents a vessel using a bottom otter trawl with a mesh size in the range 70-99mm to target demersal fish – no selectivity devices are used. The métier classification scheme can be found In the Appendix IV of the 2008/949/EC Commission Decision at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:346:0037:0088:EN:PDF#page=16

The method of defining target assemblage is described as "*The retained part of the catch should be classified by target assemblage (crustaceans, cephalopods, demersal fish, etc.) at a trip level or at a fishing operation level where possible, and sorted by weight or by total value in the case of valuable species (e.g.Nephrops, Tunas). The target assemblage that comes up at the first position should be considered as the target assemblage to be reported in the matrix*.". In practice there can be significant differences in how each country calculates the target assemblage from the fishing logbook records [12].

There are 6 Regional Coordination Groups (RCGs) within the framework of the DCF, 4 of which cover regional areas (Baltic, North Sea and Eastern Arctic, North Atlantic, and Mediterranean and Black Sea) and 2 to cover Long Distance Fisheries and Large Pelagic Fisheries (currently held jointly with the Mediterranean and Black Sea RCG). Reports from these groups can be found at https://datacollection.jrc.ec.europa.eu/docs/rcm/.

Since métiers are based on data recorded by fishers if data has been input incorrectly then it will generate a métier that doesn't exist in reality. To ameliorate this problem the RCGs have defined métiers which are believed to be valid fishing activities. The combination of métier and spatial area can also be validated so that some métiers are

---

[11] CWP Handbook of Fishery Statistical Standards. Section M: FISHING GEAR CLASSIFICATION. CWP Data Collection. In: *FAO Fisheries and Aquaculture Department* [online]. Rome. Updated 17 January 2017. [Cited 8 November 2017]. http://www.fao.org/fishery/cwp/handbook/M/en

[12] Nicolas Deporte, Clara Ulrich, Stéphanie Mahévas, Sébastien Demanèche, Francois Bastardie; Regional métier definition: a comparative investigation of statistical methods using a workflow applied to international otter trawl fisheries in the North Sea, *ICES Journal of Marine Science*, Volume 69, Issue 2, 1 March 2012, Pages 331–342, https://doi.org/10.1093/icesjms/fsr197

only valid in certain places. Unfortunately there is not currently a repository to easily access these RCG métier validation lists.

5.3.4 Spatial reference systems

The FAO defines 27 major fishing areas, 8 of which are inland and 19 are marine[13].

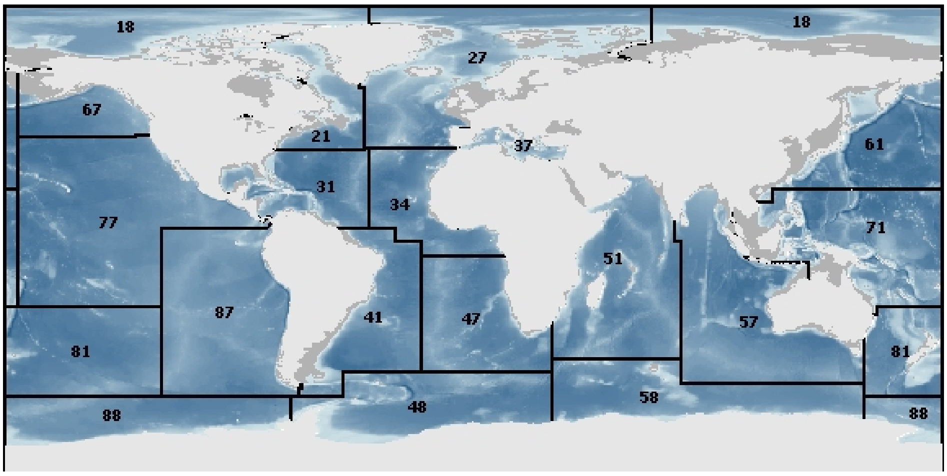


Figure 12 FAO Marine fishing areas: (http://www.fao.org/fi/figis/ontology/data/assets/images/fao_fish_area.jpg)

The areas of main interest to European countries are Areas 27 and 37.

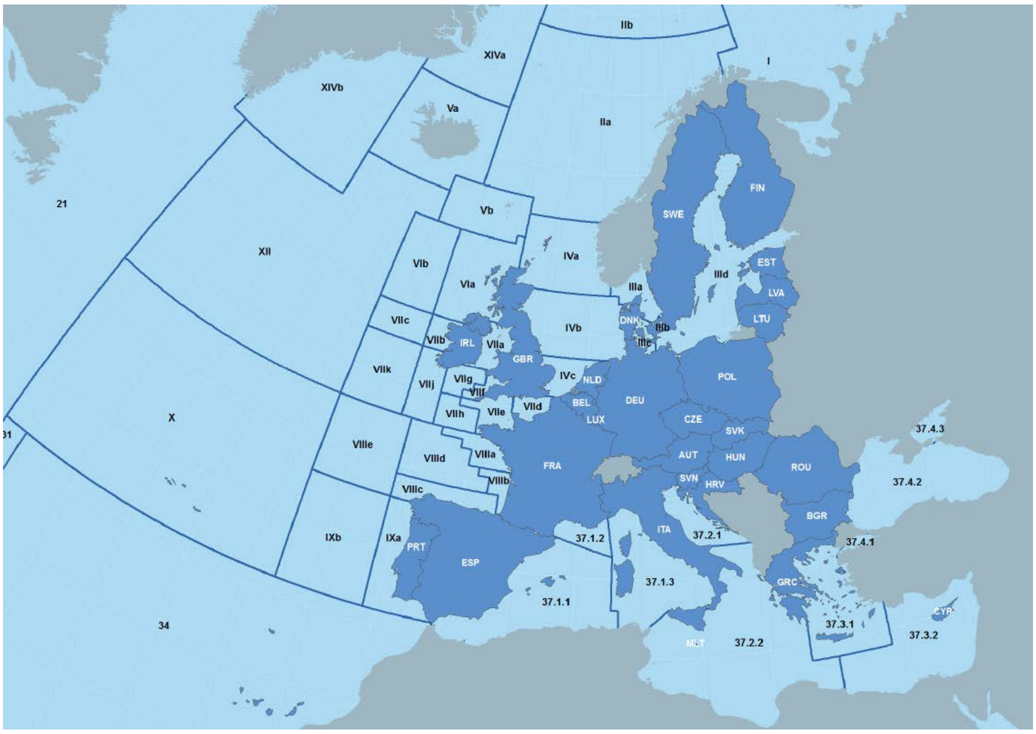


Figure 13 Sub-areas and divisions of FAO areas 27 and 37 (https://ec.europa.eu/fisheries/sites/fisheries/files/docs/body/fishing_areas_en.pdf)

[13] http://www.fao.org/fishery/cwp/handbook/h/en

FAO Area 27 corresponds to the Northeast Atlantic and is subdivided according to the ICES spatial reference system.

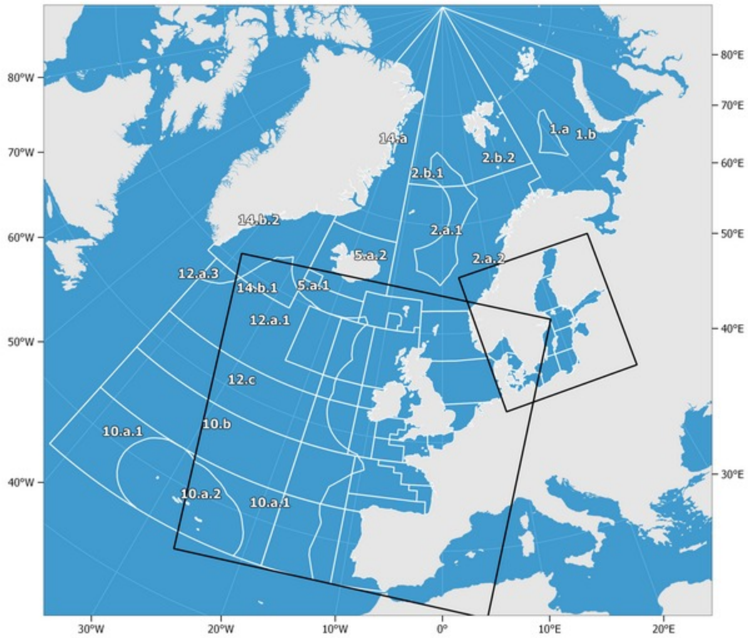


Figure 14 ICES fishing areas: (http://www.fao.org/fi/figis/area/data/assets/images/faoarea27_1.jpg)

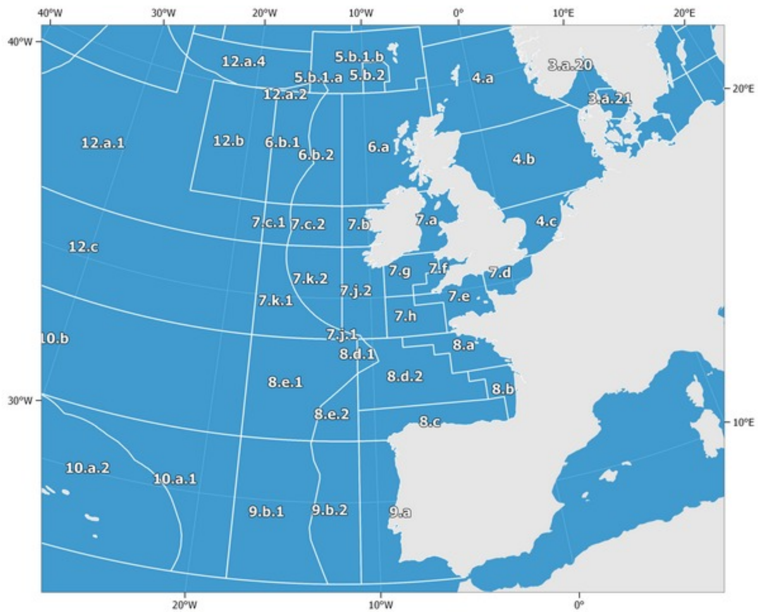


Figure 15 Detailed boundaries of the ICES Subareas 27.4, 27.5, 27.6, 27.7, 27.8, 27.9 (http://www.fao.org/fi/figis/area/data/assets/images/faoarea27_2.jpg)

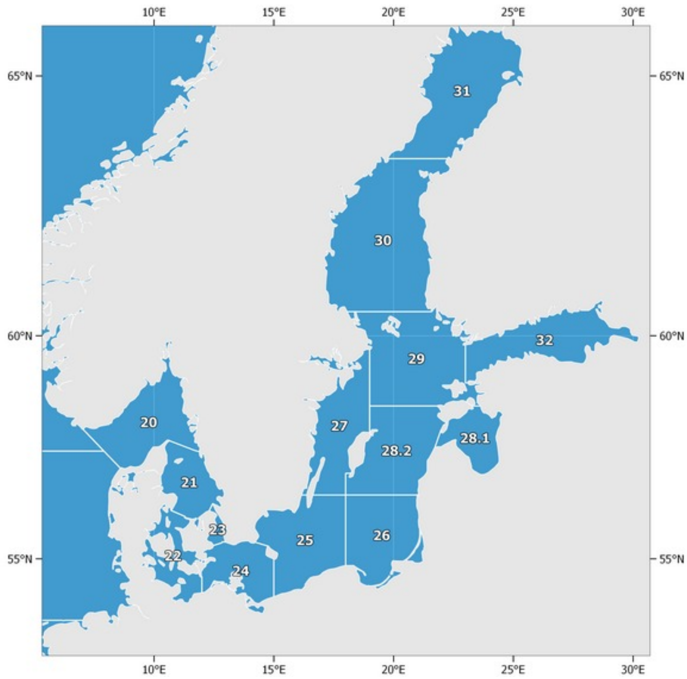


Figure 16 Map of the Baltic Sea showing the subdivisions of the Belt, the Sound, and the Baltic for the reporting of catch statistics: (http://www.fao.org/fi/figis/area/data/assets/images/faoarea27_3.jpg)

ICES also defines a grid within this region – these are known as statistical rectangles with each one covering a latitude of 30′ and a longitude of 1°. These statistical rectangles can also be further divided into sub-rectangles[14].

FAO Area 37 corresponds to the Mediterranean and Black Sea and is divided into 4 subareas: Western Mediterranean (Subarea 37.1), Central Mediterranean (Subarea 37.2), Eastern Mediterranean (Subarea 37.3), and Black Sea (Subarea 37.4). The subareas are further divided into divisions[15].

[14] http://www.ices.dk/marine-data/maps/Pages/ICES-statistical-rectangles.aspx

[15] http://www.fao.org/fishery/area/Area37/en

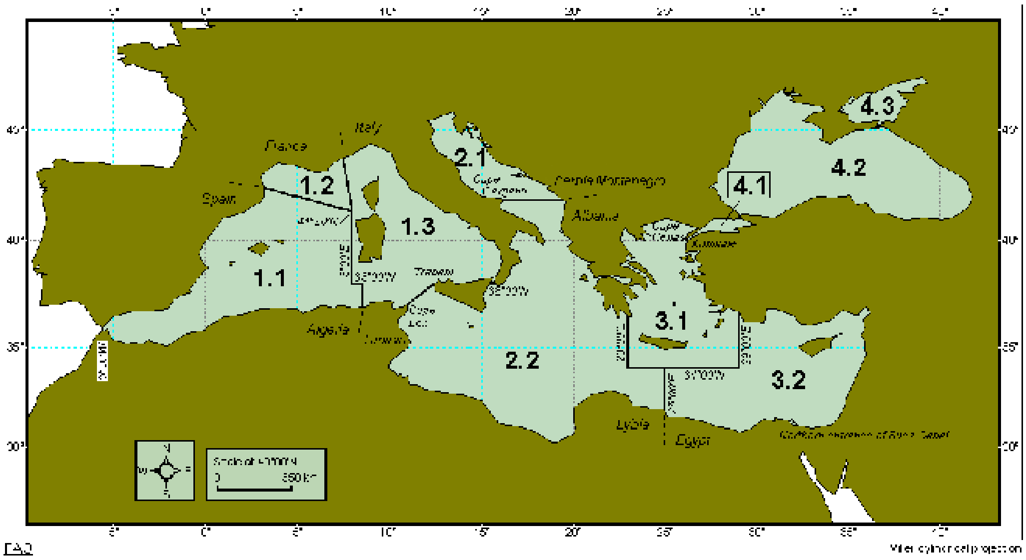


Figure 17 Subareas with FAO Area 37: (http://www.fao.org/fi/figis/area/data/assets/images/Area37.gif)

The General Fisheries Commission for the Mediterranean (GFCM) also define 30 Geographical subareas (GSAs)[16].

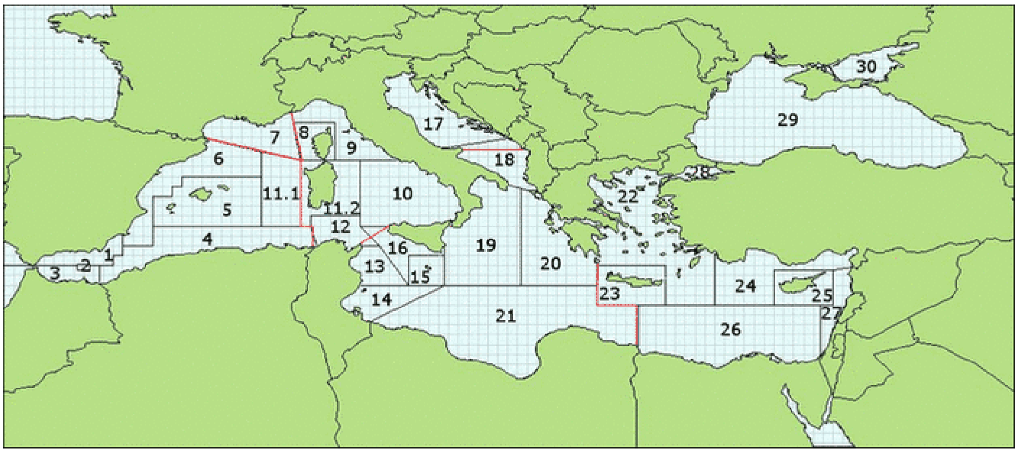


Figure 18 GFCM GSAs (http://www.fao.org/typo3temp/pics/3bf32307ba.png)

A statistical grid is also defined by the GFCM[17].

5.3.5 Ports

The United Nations Code for Trade and Transport Locations (UN/LOCODE) assigns codes to locations used in trade and transport including ports – the codes used consist of 5 characters with the first 2 characters being the ISO 3166-1 alpha-2 country code of the location. The complete list is available to download[18].

Each record in the UN/LOCODE list has the following structure[19]:

---

[16] http://www.fao.org/gfcm/data/map-geographical-subareas/en/

[17] http://www.fao.org/gfcm/data/map-statistical-grid/en/

[18] http://www.unece.org/cefact/locode/service/location

[19] https://www.unece.org/fileadmin/DAM/cefact/locode/Service/LocodeColumn.htm

Column name	Column description
Ch	Record change indicator
LOCODE	5 character code
Name	Place names are given, whenever possible, in their national language versions as expressed in the Roman alphabet using the 26 characters of the character set adopted for international trade data interchange, with diacritic signs
NameWoDiacritics	The names of the locations which have been allocated a UN/LOCODE without diacritic signs
SubDiv	Column "Subdivision" is intended to contain the ISO 1-3 charac-ter alphabetic and/or numeric code for the administrative divi-sion of the country concerned (state, province, department, etc.).
Function	This column contains a 8-digit function classifier code for the location. The value "1" specifies that the location is a Port.
Status	This column is intended to indicate the status of the entry by a 2-character code, e.g. whether approved by Government, by Customs etc.
Date	The last date when the location was updated/entered.
IATA	The IATA code for the location if different from location code in column LOCODE.
Coordinates	The geographical coordinates (latitude/longitude) of the location.
Remarks	General remarks.

The EU also provides a similar location list which uses a similar 5 character code but specifies more locations for EU countries than the UN/LOCODE list[20]. Where the EU location already exists in the UN/LOCODE list it uses the same code e.g. the Irish port of Dunmore East has the code IEDNM in both the EU and UN lists but the EU list also defines IEUNI as Union Hall which is not listed in the UN/LOCODE list.

5.3.6 EU Master Data Register

CIRCABC (Communication and Information Resource Centre for Administrations, Businesses and Citizens) is an application used to create collaborative workspaces where communities of users can work together over the web and share information and resources[21]. Within the "Maritime Affairs and Fisheries" category there is a Master Data Register which contains lists of fisheries codes to be used in electronic information recording and exchanges among Member States and for Member States' communica-tions with Norway with the purpose to record and report fishing activities[22].

---

[20] https://circabc.europa.eu/d/a/workspace/SpacesStore/d47b93ca-c54c-4b64-9d9c-cdf983566028/Code-Location-v2.0.xls

[21] https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp

[22] https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp?FormPrincipal:_idcl=FormPrincipal:left-menu-link-inf-closed&FormPrincipal_SUBMIT=1&javax.faces.ViewState=ab-nxtj5XqtsuqVWygZiy573cj2Xs0mk40VpRYe1zidXXGTDy%2FEhDEj%2Bzd2Z0GtgLuwbZI-uIQz%2BzN5hF03KZeaZY99765PLYOJUJxS3oxETgrHc8z8b6Y2j0%2Fe-bunzsuTbJNuL4X2ihaso5ylcWgUpvOMnFo%3D

This contains a number of lists of reference data which is already in use within system such as the ERS Logbooks system and the FLUX (Fisheries Language for Universal EXchange) standards.

5.4 Data Interoperability – Common Data Structures

There are a number of common data structures that are commonly used to exchange data.

5.4.1 DATRAS

DATRAS is an online database of trawl surveys coordinated by ICES expert groups that uses a standard data format. When DATRAS data are exchanged each file contains three record types with the relevant headers:

- HH: Haul meta-data
- HL: Species length-based information
- CA: Species age-based information

Detailed description of these records and corresponding fields can be found under DATRAS menu "Reporting Format". It is possible to relate different record types by the first 11 fields. So, HL- and CA- records would belong to a HH- record with the same quarter, country, ship, gear, station number, haul number, and year. A detailed description of the format can be found at ICES[23].

5.4.2 MEDITS

As discussed earlier the MEDITS project created a common data exchange format consisting of the following tables:

- TA - haul data: start and end position (lat, long, depth), date, time, vertical and horizontal net opening, etc.;
- TB – catch data: information on the number of individuals and total weight by species for each haul;
- TC – length data: aggregated number of individuals by length, sex and maturity by species;
- TE – individual data on weight, sex, maturity stage and age;
- TL – litter data

In addition, an open source package (SDEFQuality)[24] was developed to check submitted data against this, or any other format which is defined and supplied to the application in an Excel workbook.

5.4.3 Regional Database (RDB)

The RDB is hosted by ICES and stored fisheries dependant data from the 3 RCGs (North Sea and Easter Arctic, Baltic, and North Atlantic)[25]. It consists of 3 types of data:

- CS – disaggregated commercial fisheries sampling data
- CL – aggregated commercial fisheries landings data
- CE – aggregated commercial fisheries effort data.

The CS data structure is further broken down into the following types of records:

[23] http://dome.ices.dk/datsu/selRep.aspx

[24] https://github.com/ldbk/fishPifct

[25] http://www.ices.dk/marine-data/data-portals/Pages/RDB-FishFrame.aspx

- TR – Fishing trip
- HH – Haul header
- SL – Species list
- HL – Length frequency data
- CA – Sex-Maturity-Age-Weight-Length data

5.4.4 FLUX – Fisheries Language for Universal Exchange

FLUX has been developed by an expert group of the United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT) as a standard for the exchange of fishery messages for the sustainable management of fisheries. It is split into a "Transportation Layer" which defines how to send and receive messages, and a "Business Layer" which defines the content of those messages.

FLUX defines a language for exchanging data on fisheries – its scope is limited to data from catch to the first sale (including processing on board, landing, transhipment, transport, first sale, inspection, vessel position, vessel information, and vessel licenses).

FLUX doesn't define a database structure instead it defines a message structure. FLUX is split into "domains" – where a domain is a family of messages about a common subject.

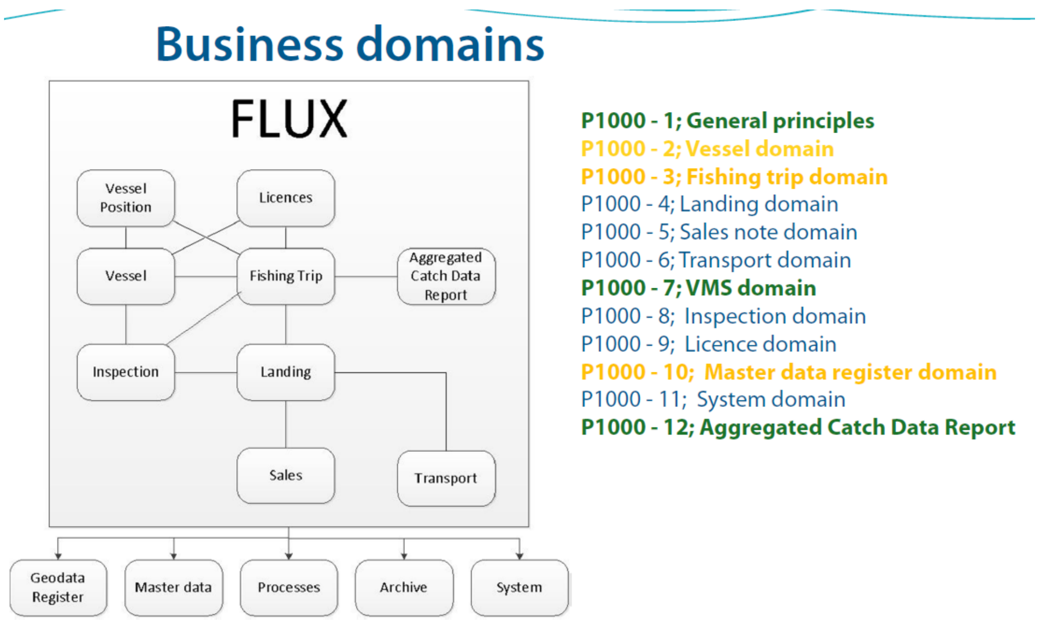


Figure 19 FLUX Business Domains

(https://ec.europa.eu/fisheries/sites/fisheries/files/docs/body/annex-2.9_en.pdf)

5.4.5 ICES Acoustic Database format

ICES have defined a format for an acoustic database which consists of two parts – an Acoustic and a Biotic part[26].

The Acoustic part of the format consists of six record types: the five metadata record types Instrument, Calibration, Data Acquisition, Data Processing, and Cruise; and one

[26] http://www.ices.dk/marine-data/Documents/Acoustic/ICES_Acoustic_data_format_description.zip

data record type Data (which is the combination of Log, Sample and Data entities in Figure 20).

The Biotic part of the format consists of four record types: one metadata record Cruise; and three data record types Haul, Catch, and Biology.

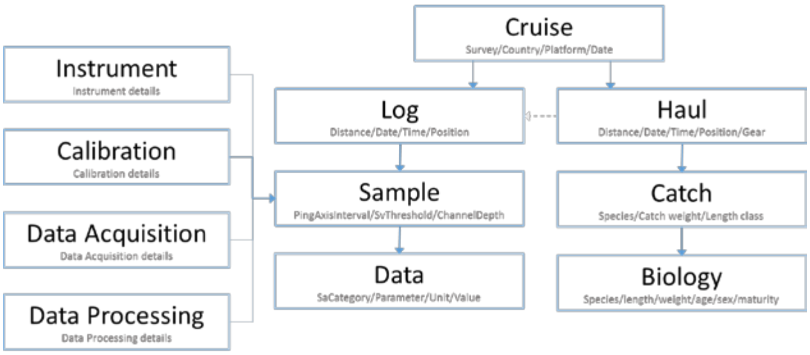


Figure 20 ICES acoustic database format

5.5 Discussion

It has been shown in the 3 case studies considered that a large amount of work has been put in to developing QC checks. In particular there is a lot of experience in which QC checks are important to trap measurement errors and how results can best be presented (e.g. use of graphical output can be more expedient than tabular or text output).

The QC processes and checks developed by Ireland and Scotland essentially have been developed independently and are limited to working within their respective country's EDC system (which happen to be very similar) – this makes it difficult to directly share code or scripts with the wider community. The FishTrawl/RoME project shows the benefit of developing QC processes within a collaborating group of countries. In this case the QC functions are developed to work with the common MEDITS data structure so that any institute which uses this structure could potentially benefit.

There already exist a number of sources of reference data for common concepts – if institutes can either directly use these reference data sources or map their own data to them then it becomes much easier for people to work together.

Similarly there are already a number of common data formats defined – if code and scripts are designed to work with one of these formats using standard reference data then it becomes much more practical for institutes to collaborate on EDC and QC tool development.

5.6 Conclusion

A common QC framework built to process standard data formats and reference data would allow the already existing QC knowledge, experience, processes and code to be shared much more widely than is done currently and benefit many institutes.

6 Conclusions on Collaborative potential

6.1 Electronic Data Capture Hardware

Section 4, Review of Current Approaches, presents overviews for the approaches to sourcing and/or customizing the electronic equipment required for fisheries data capture presented at the workshop. Budget, ease of use, in-house support, warranties and connectivity are all important considerations in which approach to take. In terms of lifespan, Institutes with existing long-term electronic sampling in place report operating comfortably for 10yrs or more with an existing system without significant overhaul or upgrade. Therefore return on investment over the life cycle of the technology gives a truer context to the potentially significant outlay initially. Options for leasing arrangements or annual renting as part of data collection programs was briefly discussed to offset the initial outlay, but not concluded so postponed for a later workshop.

Except where a novel technology might be sought, the hardware components and basic software are generally available and straightforward to specify and compare costs. Development has been required for most of the boards presented, but essentially they all emulate standard keyboard type entry usually integrated with a pre-defined number range for efficient entry of length data. However, the potential range of functionality of the software for data quality management is an order of magnitude greater and therefore the focus for the rest of the discussion below.

6.2 Electronic Data Capture Process and Interfaces

There are a number of steps that are generally common in the process of electronically capturing data from a biological sample regardless of user or data model (see Figure 1):

1. Metadata about the sample needs to be recorded. This could include details such as the survey name, station number and position, or for commercial sampling the sampling place, vessel sampled, and origin of the catch. It will also include details about the fish such as species and size category. The metadata might be entered via a central system and "pushed" to the sampling workstations; it could be entered entirely at the workstations, or a mix of both of these.
2. Measurements of the sample are then taken. These could be data such as length measurements captured by an electronic measuring board, weight measurements captured by scales, or size measurements captured by electronic callipers. There will usually be a series of measurements taken for each sample (e.g. 50 fish might have their lengths measured). There can also be multiple types of measurements captured electronically for the same sample – an example of this is length stratified sampling where a certain number of fish in each length class will also have their weight sampled. In this case some length measurements will be followed by a weight measurement. The application recording these data could be something like an Excel spreadsheet or a complicated custom application with QC functions, or it could be application embedded in the measuring device.
3. Often in the case of surveys there will be more than 1 sampling workstation recording data from a single survey station/haul (sampling event). The catch will be sorted and divided among the sampling workstations for recording – this could result in 2 sampling workstations recording data on the same component of the catch (e.g. same species, size category, sex). If there are multiple sampling workstations then the data recorded from these must be combined.
4. Usually the electronic data capture process cannot rely on being constantly connected to a computer network so this means data will be stored locally for a certain

length of time. This could be for a relatively long time (e.g. for a whole survey leg) or a relatively short time (e.g. only until a sample has been completed). This local database can be simple or complicated and might contain QC functions.

5. At an appropriate point the data will be transferred from the local database to a database located within the sampling institute's network. This network database will usually be more sophisticated than the local database and allow more rigorous analysis and QC. It might also allow for extra data to be appended to the sample data at a later point (e.g. age readings).
6. Once the institute has completed its data collection programme and QC procedures it will normally have to upload the data to an international database – the timing of this will vary depending on the data. Some examples of these international databases include DATRAS and the RDBs. The upload procedure for these international databases will often be manual (e.g. the user extracts data from their national databases and uploads it via a web portal) and rely on the data passing a series of defined QC tests and rules.
7. QC processes tend to be spread throughout all the stages of the collection.

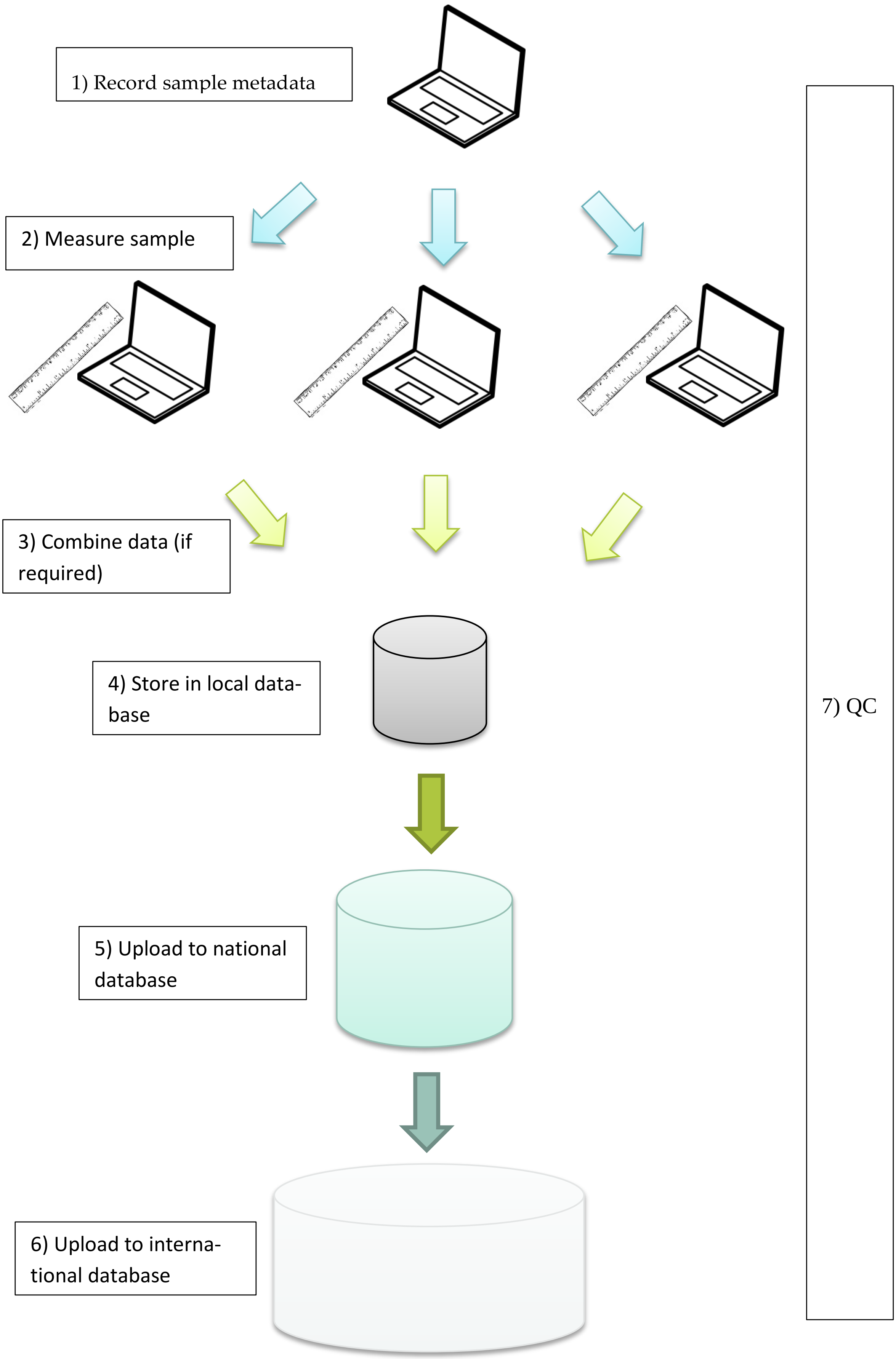


Figure 21 Electronic data capture process

Although many countries and institutes follow a broadly similar process when considered abstractly, in practice the details of the measurement devices, applications and procedures that each uses will differ significantly. This means that currently each country essentially has to modify or develop its own end-to-end process including the design and building of its own measuring boards and writing custom in-house applications and database systems. Really the only common point is that they all have to upload to the same international database so are required to be able to output their data in the agreed international format and are subject to the same final QC checks. The processing and procedures that are applied to the data before this point are essentially the responsibility of each country (while following agreed survey protocols and data collection plans).

The current situation leads to duplication (multiple countries have to develop essentially the same systems) and a lack of inter-operability (each institute's system is tightly integrated into their own processes so cannot be used easily by any other institutes). This is both inefficient in terms of the usage of limited resources and doesn't facilitate the best tools and techniques to be shared among the international community. Most importantly it has resulted in many countries postponing or avoiding electronic data capture altogether - the result is an unavoidable delay implementing automated broad spectrum quality assurance checks at an early point in the process whilst the samples are still available to correct any errors.

Attempting to design a single system and forcing everybody to use it would be unsuccessful since there are often good reasons for differences in systems due to each institute's infrastructure and processes. A single system would also probably end up being overly complicated as it tried to fulfil all users' requirements. A more effective and agile approach to inter-operability and collaboration is to modularise the electronic data capture process. Interfaces between these modules can then be specified. As long as each module is able to fulfil the requirements of the interface then the details of how it does that can be left to each institution. Modules produced by different people can also be interchanged since they all are capable of fulfilling the interface requirements. This is similar to apps on a smart phone – because each app is written to interact with the Android/iOS operating system interfaces they can run on a variety of different phones made by different manufactures using different hardware. These apps can also interact with each other through interfaces without needing to know in advance exactly which other apps will be installed on a user's phone. During install many Apps will request access to the phone's location data, camera, address book and so on without any further set up information being required – modular 'Plug and Play' hardware/software.

Figure 2 defines the location of these interfaces within the data capture flow. These interfaces allow different parts of the electronic data capture process to be encapsulated from each other. For example the "Measurement Application" does not need to know any details about how any of the measuring devices take their measurement or their proprietary operating systems as long as all measuring devices agree to output their data via a common interface – in this case the interface would probably define a communication protocol (e.g. serial connection over Bluetooth) and a common format for the data format (e.g. perhaps something similar to NMEA sentences). The Measurement Interface could also define methods for the "Measurement Application" to interact with a measuring device (e.g. the application might be able to signal a "Bad Measurement") without needing to know exactly what type of device it was – the device would receive the "Bad Measurement" signal and do something appropriate (e.g. beep or flash a red LED).

The correct identification and definition of interfaces between modules/functions would remove the need for different parts of the system to be tightly integrated and allow more inter-operability and sharing between hardware and software.

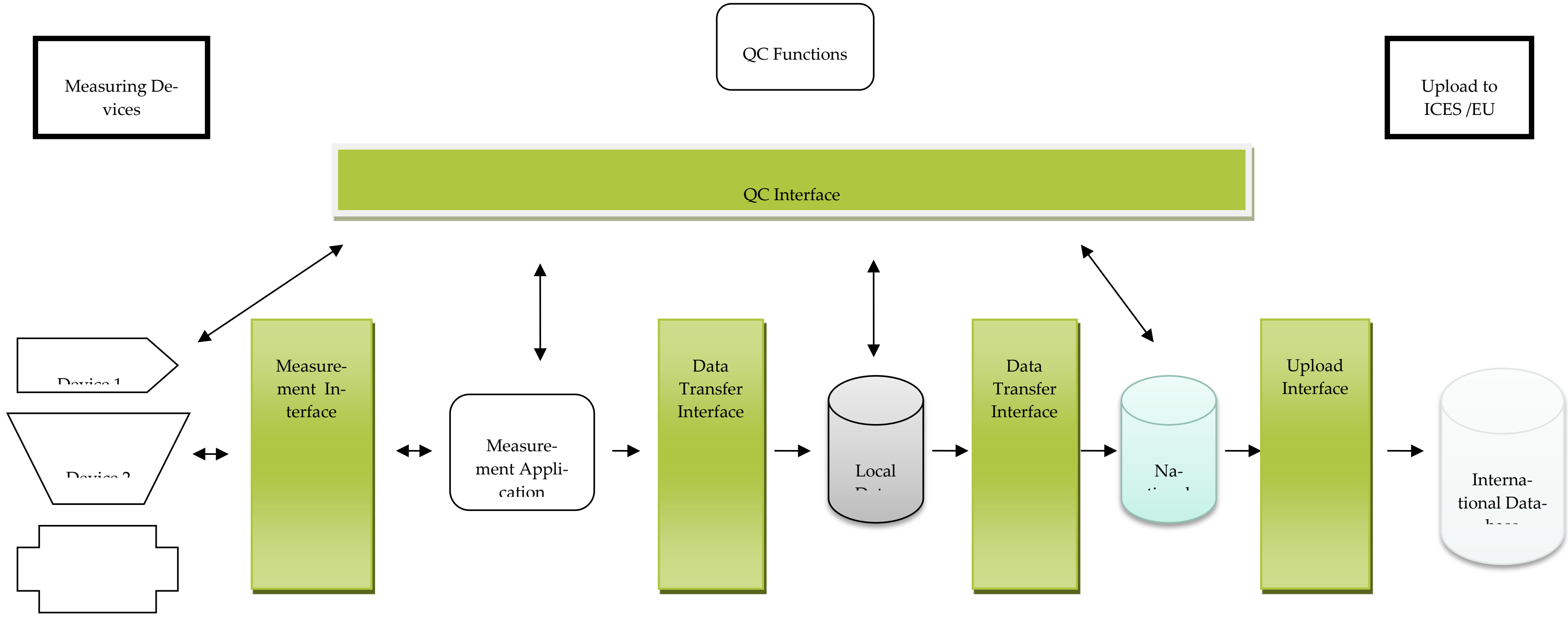


Figure 22 Interfaces between functions/modules are shown in green.

Each interface should be defined by specifying the protocol that it will use and a data format. It would be most useful if these protocols and data formats already exist since that will ease adoption and use of the interfaces. It is also useful to make the protocols chosen platform-agnostic so that they do not overly restrict users in how they implement them. An example of this would be HTTP as the protocol for an interface since essentially all Operating Systems should be able to make HTTP calls and receive the results. Figure 3 shows a simplified version of how a client and server can interact indirectly using interfaces – in this case referred to as APIs (Application Programming Interfaces).

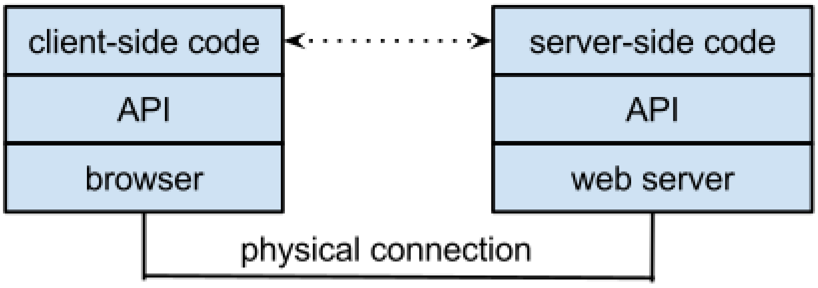


Figure 23 Client/Server API (By Lubaochuan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=34946978)

The "Upload Interface" is probably already specified since this will be the method of uploading data to an international database such as DATRAS - in this case the upload protocol would manually use the website and the common data format would be the DATRAS exchange format.

6.3 Proposal for an interface between individual survey software and an open source "quality check toolbox"

6.3.1 Definitions

Individual software:

Most organisations have developed workflows and software programs to process data during the task of data acquisition (Figure 4). The software solutions have different stages of complexity. There is a range from typing data from pen and paper into user interfaces up to electronic measuring devices producing complete datasets.

Functions (for processing quality checks):

The individual software solutions contain individual functions / programs / workflows (hereafter called 'functions') to perform quality checks (QC). They are integrated into workflows with different levels of complexity and programmed in different programming languages. An organization therefore can't easily implement and maintain the QC of other organization(s) so standardized QC throughout the workflow is problematic.

Graphical User Interface (GUI) for processing QC:

The GUIs behave in a similar way to functions. They are an integrated part of the programs and specialized to fit a specific database. They rarely work together with alternate hardware or software solutions.

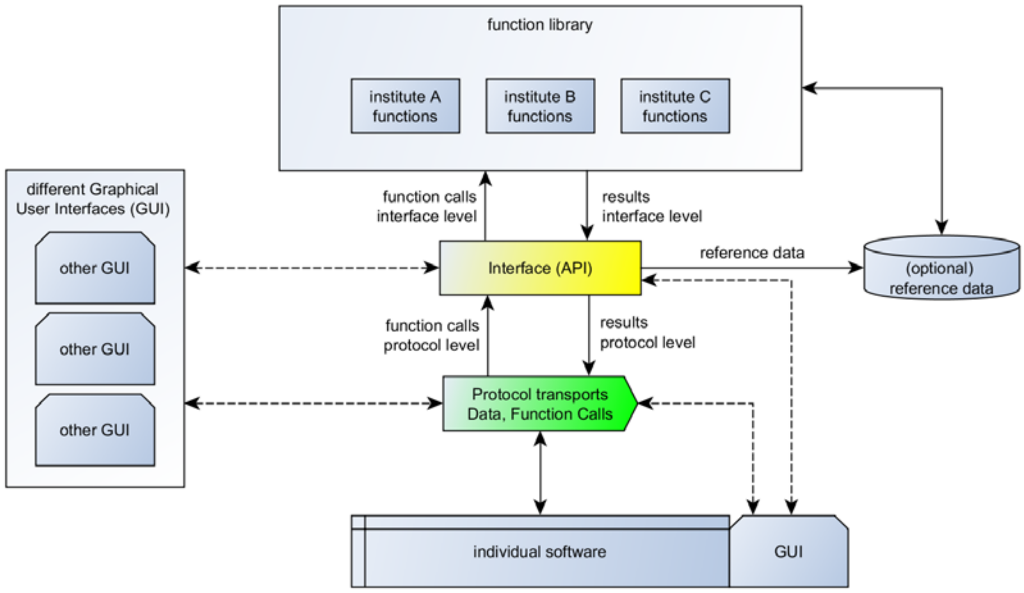


Figure 4 Schematic view of how an Application Program Interface (API) can sit between multiple independent applications managing the flow of function calls and responses as data and information flow in both directions.

Reference data:

To perform QC on a higher level (e.g. weight-length in a specific region) some reference data are needed by the functions. These datasets are "private" property of the organizations. So a specific region can't be QC by functions from a different region for example if they don't have a set of reference data.

Interface (yellow box):

On the input side the interface level accepts data, predefined instructions and optionally reference data. It forwards data to a specific function in a local function library. The function responds with a result based on the given data. The result is forwarded to software making the call and/or directly to a GUI.

Protocol (green box):

The protocol is a well-defined common communication standard. It allows individual software (with small changes in the code) to communicate with foreign functions and GUIs in a local environment.

6.3.2 Problems

On a basic level of data acquisition is no common standard established for QC-ing "fresh" data. So the different teams in the field produces datasets with very different quality levels. This case leads to problems on higher abstraction levels (e.g. RDB, international databases). For finding the source of these discrepancies, they have to be tracked down all processing layers with costs for time and resources.

6.3.3 Recommendation

The community should avoid too many different quality levels during the data generation process. To optimise synergies effects and resources a flexible interface layer (yellow box) with a common communication protocol (green box) is proposed.

This communication layer is the prerequisite for establishing an open source toolbox for quality checking. This toolbox can contain QC functions from different organisations and can be applied at any point in the sampling process once the data are in electronic format. Even closed source programs can be connected to various applications, as long as they "speak" via the common protocol.

To support collaboration across Institutes and integration of functions within a toolbox, the interface should be able to call functions from different platforms and programming languages.

6.3.4 Benefits

With a well-defined interface and protocol every team can perform similar real-time checks, based on a common standard. Cross-checks with different QC functions from different organizations can be done. As a side effect a QC of different functions is possible.

Organizations with a simple (or no) data capture method which haven't developed specific QC functions can easily plug into the toolbox.

Functions and GUIs can freely be combined to get the best solution for your own needs. Resources like time and money can be saved.

A common quality standard for quality assuring data at source in real-time can be then developed in the community rather than standard checks being applied during the final upload screening process. This ability to share and 'plug into' a common QC toolbox addresses the issue in Section 3 of Institutes not always being sure what to expect from a data capture system. Where much of the data QC higher functionality might already be available and easily shared the scoping of a data capture module itself should become quite straightforward.

6.3.5 Example

A. has no solution for a specific QC function
B. has developed a very smart, but slow function
C. invented a quick function, but it has less accuracy
D. visualizes QC results in an easy to understand way

A can combine functions from B or C with the GUI from D, because the interface/protocol is the same for all modules. Perhaps A is using C's function for a quick check done by the technicians in the fish lab during the data collecting process. When the team moves on to the next haul and time is no limiting factor, A's cruise leader double checks data with B's function.

6.3.6 Necessary next steps

1. A group of quality controllers and software developers define a common communication protocol. They also define a small set of easy to implement real-time QC functions.
2. Software developers define a multiplatform interface to connect QC functions written in different languages on different operating systems.

3. Implementation of a reference interface to demonstrate how the protocol is processed and connected to QC functions.
4. Implementation of a basic QC function library.

Encourage organisations to equip their individual software solutions with a protocol interface.

Annex 1: List of participants

Name	Address	Phone/Fax	E-mail
Chris Carroll	Big Fin Scientific, Austin Tx, USA		chris@bigfinscientific.com
Isabella Bitetto	COISPA, Bari, Italy		bitetto@coispa.it
Henrik Degel	DTU Aqua, Copenhagen, Denmark		hd@aqua.dtu.dk
Jorge Tornero	IEO, Cadiz, Spain		jorge.tornero@ieo.es
Vincent Badts	IFREMER, Nantes, France		Vincent.Badts@ifremer.fr
Kevin DeCoster	ILVO, Ostende, Belgium		Kevin.DeCoster@ilvo.vlaanderen.be
Wim Allegaert	ILVO, Ostende, Belgium		Wim.Allegaert@ilvo.vlaanderen.be
Jens Rasmussen	Marine Scotland Science, Aberdeen, Scotland		J.Rasmussen@marlab.ac.uk
Ruadhán Gillespie-Mules	Marine Scotland Science, Aberdeen, Scotland		r.gillespie-mules@marlab.ac.uk
Anders Svenson	SLU Aqua, Lyeskil, Sweden		anders.svenson@slu.se
Mikael Ovegard	SLU Aqua, Lyeskil, Sweden		mikael.ovegard@slu.se
Dave Stokes (co-chair)	The Marine Institute, Galway, Ireland	+ 353 (0)91 38 7200 Ext 7426	david.stokes@marine.ie
David Currie	The Marine Institute, Galway, Ireland	+ 353 (0)91 38 7200	David.Currie@Marine.ie
Hans Gerritsen	The Marine Institute, Galway, Ireland	+ 353 (0)91 38 7200	hans.Gerritsen@Marine.ie
Daniel Stepputtis	Thünen-Institut, Rostock, germany		daniel.stepputtis@thuenen.de
Marcellus Rödiger (co-chair)	Thünen-Institut, Hamburg, Germany		marcellus.roediger@thuenen.de
Richard Ayers	Seascope Fisheries, Scarboruogh, UK		richard_ayers@hotmail.com

Leif Thomsen	DTU Aqua, Copenhagen, Denmark	lt@aqua.dtu.dk

Annex 2: Agenda

Tuesday 12th Sept 2017			
Time / Location	**Item** / Person	**Process**	**Output**
10.00 – 10.45 Atlantic Room	**Welcome**, Confirmation of agenda, **Overview** Dave	Introductions & Presentation	Opening, adoption of the agenda, background to workshop.
10.45 – 11.30	**Smartfish** Wim Allegaert & Kevin DeCoster	Presentation & discussion	TOR (b)
11.30-12.00 Coffee			
12.00 – 13.00 Conference room	**OpenSMB** Marcellus Rödiger & Daniel Stepputtis	Presentation & discussion	TOR (b)
1300-1400 Lunch			
Time / Location	**Item** / Person	**Process**	**Output**
14.00 – 14.45	**ICROS** Jorge Tornero	Presentation & discussion	TOR (b)
14.45 – 15.30	**Tutti/Allegro?** Vincent Badts	Presentation & discussion	TOR (a,b)
15.30-16.00 Coffee			
16.00 – 17.00	**DCS - BigFin** Chris Carroll	Webex/Skype Presentation	Review and discuss format to address TOR (b)
17.00 -	**Close/Report writing?**		

Wednesday 13th Sept 2017			
Time / Location	**Item** / Person	**Process**	**Output**
09.00 – 09.10 Atlantic Room	**Review & comments day 1** Marcellus	Discussion	Feedback and comments for TOR (b)
09.10 – 10.00	**Swedish solution** Mikael Ovegård & Anders Svenson	Presentation & discussion	TOR (b)
10.00 – 10.40	**Pit Tag & other Approaches** Richard Ayers	Discussion	Review of other methods and Summary discussion for TOR (b)
10.40-11.00 Coffee			
11.00 – 11.45	**Measuring boards vs measuring systems** Dave	Presentation & discussion	Opening discussion for TOR (a)
11.45 – 12.30	**CEFAS EDC Board & Extensions** Jens Rasmussen	Presentation & discussion	Review to address TOR (b, a)
12.30 -13.30 Lunch			
13.30 – 14.45	**Rome/Fishtrawl** Isabella Bitetto	Presentation & discussion	Review to address TOR (a)
14.45 – 15.00	**Pros & Cons of Data Visualization & QC in Open Source methods such as R** Hans Gerritsen	Webex\Skype Presentation & discussion	Review to address TOR (a,c)
15.00-15.30 Coffee			
15.30 – 16.15	**RDB/DATRAS data models – history & potential for collaboration** Henrik Degel	Presentation & discussion	Opening discussion for TOR (c)
16.15 – 17.00	**Plenary** All	Discussion	Agree format and tasks for summarising material for TORs (a,c).
17.00 -	**Report writing & close**		

Thursday 14th Sept 2017			
Time / Location	**Item** / Person	**Process**	**Output**
09.00 – 09.15 Atlantic Room	**Review & comments day 2** Dave	Text & discussion	Feedback and comments for TOR (b)
09.15 – 09.45	**1. Report Overview Dave/Marcellus**	Review text & Discussion	Benefits and impediments to paperless sampling (financial/time cost, expertise, technical shortfall?).
09:45 – 10:15	**2. Requirements** **??**	Review text & Discussion	Who are the users? How many survey/commercial/freshwater samplers? Are we replacing a pencil or bringing 1st phase data management into the field?
10.15 – 10.30	**3. Current Systems** **??**	Review text & Discussion	Review in terms of functionality, cost, flexibility, ease of use, tech know-how in the field, extendibility and data/QC management & security, etc.. Information gaps? Report/table format?
10.40-11.00 Coffee			
11.00 – 11:45	**4. Data QC & Management** **??**	Review text & Discussion	What are useful checks during a sampling event? Useful sources of code, reference tables (e.g. Spp, metiers), sampling targets, historic and spatial data for context, reporting functionality to showcase sampling program on the frontline.
11.45 – 12.30	**5. Collaborative potential** **??**	Review text & Discussion	Would a "field" version of e.g. RDB/DATRAS with near "realtime" QC/reporting functions be useful 1st step? Sharing of expertise through common data model, exchange format, ongoing coordination of developments, standards, reference tables etc. Funding proposal?
12.30 -13.30 Lunch			
13.30 – 14.30	**6. Recommendations Marcellus/**	Review text & Discussion	Current gaps and proposed solutions with guide costs and realistic timelines to feed into recommendations.
14.30 -	**Discussion & project proposal(s)**	Discussion & outline for	Objectives, participants, costs, timelines

Annex 3: Resolutions

WKSEATEC – Workshop on Technical Development to Support Fisheries Data Collection

The **Workshop on Technical Development to Support Fisheries Data Collection 2 (WKSEATEC2)**, will make recommendations on technical solutions for the collection and quality assurance of fisheries data at sea and in ports. The workshop will be co-chaired by Dave Stokes, Ireland, and Marcellus Rödiger, Germany, will meet on 11 – 13 September 2018, in ICES Headquarters, specifically to:

a) Review and support progress on electronic measuring board projects underway and presented at WKSEATEC2017;

b) Review additional electronic data capture technologies such as electronic callipers, scanners beyond scope of WKSEATEC2017;

c) Address the key recommendation from WKSEATEC2017 by agreeing on a roadmap to defining a common Fisheries Data Language (FDL) and the development of an Application Program Interface (API).

WKSEATEC will report by **29 October 2018** to the attention of the EOSG Committee.

Supporting Information

Priority	Substantial resources are expended on fisheries data collection annualy with much of the data screening occuring often weeks or months after sampling is complete. Electronic data capture provides the opportunity to review data in realtime while samples are still available thus facilitating the correction of data rather than its removal after the fact where issues arise. It is critical therefore that fisheries data collection be supported to utilize the technologies avaialble to maximise quality assurance during the narrow window where sampling process is actually live.
Scientific justification	**Justification by topic area** **a) – Update on Board Development** Several countries are in the process, or recently completed electronic measuring board development and would benefit from updates following significant exchange of ideas at WKSEATEC2017. **b) – Review of additional data capture technologies** The 2017 workshop ostensibly limited itslf to measuring board technologies in the first year to ensure this multi-disciplenary and multi-project topic was addressed in reasonable detail. Application of a number of other data capture technologies such as electronic callipers, scanners, various tags, cameras for example is being actively pursued by many member states. The effectiveness and application of these in both teleost and non-teleost sampling programs is of equal relevance to data quality management and would therefore benefit from a comparative review. **c) – FDL & API** The ambitious, but key outcome from the 2017 workshop was the concept of a common Fisheries Data Language (FDL) in conjunction with an Application Program Interface (API). Both concepts are proven in other fields, but were seen as potential 'game changers' in supporting the integration of technology and open source "data tool boxes" for fisheries data collection. An FDL in itself would enhance technology integration and data exchange by extending the familiar concept of exchange files to include additional data types not already covered by DATRAS, RDB for

	example. An API would operationalise this static format so incoming data from a range of hardware could be automatically recognised through this common language, once hardware and software are connected through the API. Both of these concepts need further development - the workshop will agree on the specific outcomes and milestones that are required, who will be involved in this development, and a timeline. If possible, simple implementations could be developed or presented during the 2018 workshop.
Resource requirements	A 3 day workshop to work on TORs and report recommendations.
Participants	The Group is normally attended by some 15–20 members and guests.
Secretariat facilities	Admin support and communication with other relevant groups/meetings where sampling data quality and planning is a term or reference.
Financial	No financial implications.
Linkages to advisory committees	EOSG (SGIEOM), SCICOM, ACOM..
Linkages to other committees or groups	Members of IBTS, MEDITS, ICES Data Center/DIG, PGDATA and WKINVITED, FishPi2.
Linkages to other organizations	TBC.

Annex 4: Recommendations

Recommendation	Adressed to
1. WKSEATEC 2017 recommends a follow up WKSEATEC II workshop is arrange for Q4 2018 to address outstanding work (see Annex 3).	ICES Secretariat
2. WKSEATEC 2017 recommends collaboration with WKINVITED hackathon workshop in Copenhagen on 29-30 May to evaluate some QC visualization.	WKINVITED
3. WKSEATEC 2017 recommends collaboration with PGDATA to prioritise and coordinate the QC procedures most useful and effective during the sampling process (“*in the field*”) as part of PGDATA work on a Quality Assurance Framework.	PGDATA
4. WKSEATEC 2017 seeks feedback from DIG as to where the work of SEATEC might integrate with data collections and various QC and assement tools being developed at the ICES Data Centre.	DIG