

EMODnet – Chemistry

ATLANTIC

Data Qualification processes for French Coastal Data in Q²

Jun 2015

Morgan Le Moigne and Emilie Gauthier , DYNECO/VIGIES, Ifremer

Content

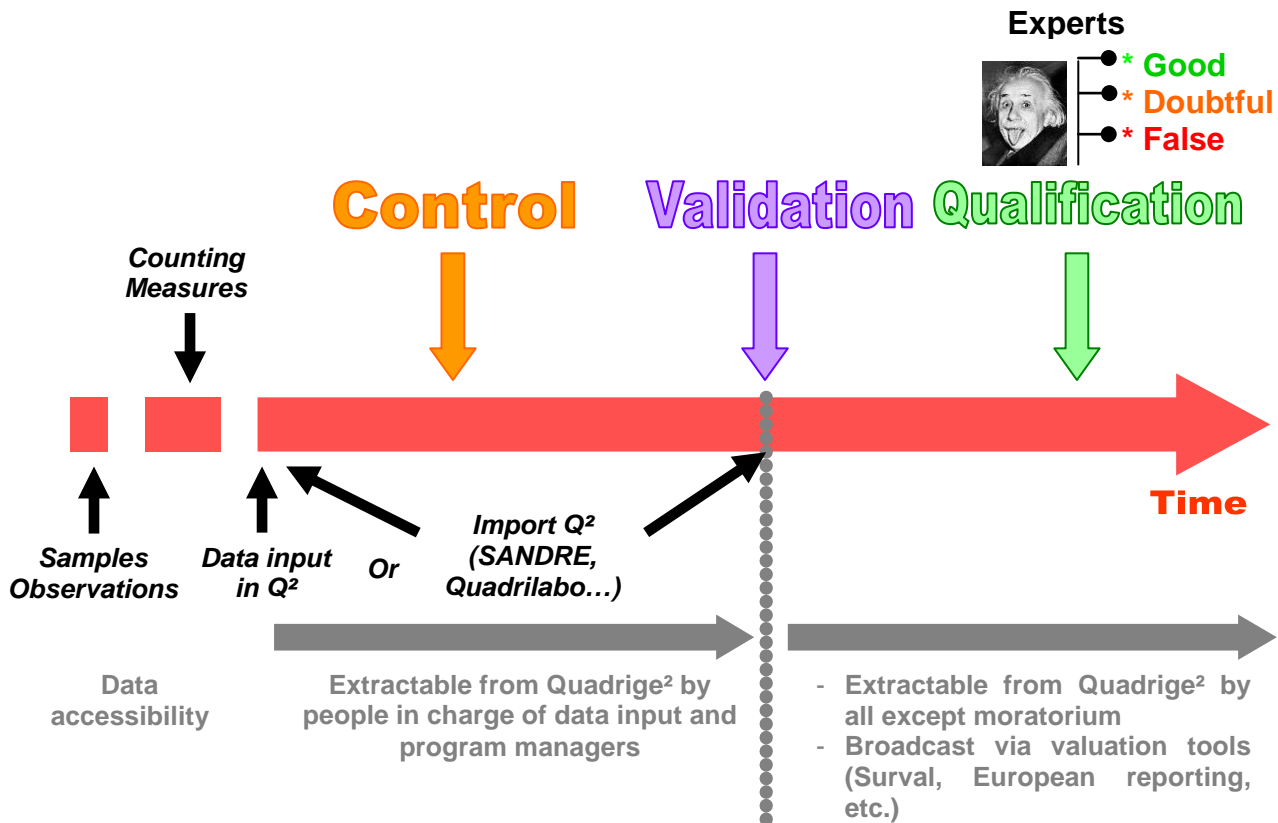
1. Introduction.....	3
2. Qualification processes	4
2.1 “Automatic” qualification.....	5
2.2 “Expert” qualification	5
3. Single data qualification or dataset qualification ?	6
3.1 Occasional qualification.....	6
3.2 Qualification of a homogeneous dataset	7
3.2.1 Integration of dataset.....	7
3.2.2 Expertise on specific datasets	7
3.3 Routine qualification of data regularly integrated in Q ²	8
4. List of quality levels and definitions	8
5. Conclusion	9

Figures

Figure 1 : Diffusion process for Ifremer coastal data	3
Figure 2 : Example of a graph for DDTpp' (RNO network) for expert qualification. The big points, ringed with black, are the data identified as « outliers » (potentially doubtful or false).	6
Figure 3 : Access to Q ² qualification tool.....	6
Figure 4 : Q ² data qualification grid.....	7
Figure 5 : Example of qualified data symbols in Q ² : purple, validated data only (Not qualified) and green qualified data.	9
Figure 6 : Role distribution in the qualification process.....	9

2. Qualification processes

Quadrigé² data have a life cycle common to all themes:



Data are collected on the field and/or on laboratory and input into the Quadrigé² database through the application of the same name. **Control** is under the responsibility of people in charge of data input and/or people with access to field records and laboratory sheets. They make a data output (results and metadata) and check their consistency with the field sheets.

Once the control and corrections have been done, **data are validated** by these same operators:

1. Confirmation of the **technical validity** of the data (correspondence with the result of the analysis)
2. Data are **locked** (it cannot be changed, even by people in charge of data input)
3. **Dissemination** of the data: validated data are downloadable by all Q² users with access to the database, and disseminated via Surval (unless the data is protected by a moratorium).

Qualification is realized after that first data verification process. Qualification involves:

- Research of doubtful data or outliers from a scientific point of view,
- Correction of data when possible,
- Attribution of a qualification level to the data. This level is:
 - o **good** : data makes sense, their analysis will be relevant,
 - o **doubtful**: data may be wrong: they may bias the analysis that will be made,

o **false**: data are aberrant or has a known problem (e.g. bad analytical series and impossibility to remake). They will not to be integrated with data analysis.

Qualification level corresponds to the confidence level in the data. Only data qualified "good" and "doubtful" are disseminated via Surval.

Qualification is divided into two main steps: an **"automatic" qualification** and **"expert" qualification**.

2.1 "Automatic" qualification

Obvious or easily identifiable errors are detected (e.g.: parameter or analytical support error, error in the sample : 100 °C instead of 10 °C) or inconsistencies (e.g.: data entered on the level "surface" with a depth of 20 m) . These errors can be detected by computer by defining simple control rules (e.g. immersion < 2 m).

Automatic qualification involves awarding a level of quality data possibly temporary (good, doubtful or false).

Only good or doubtful qualified data are used for expert qualification.

2.2 "Expert" qualification

The responsible for this qualification are thematic experts who have the scientific knowledge needed to interpret the data. It consists to highlight the statistical outliers via appropriate methods (time series, statistical tests ...) (see the example Figure 2).

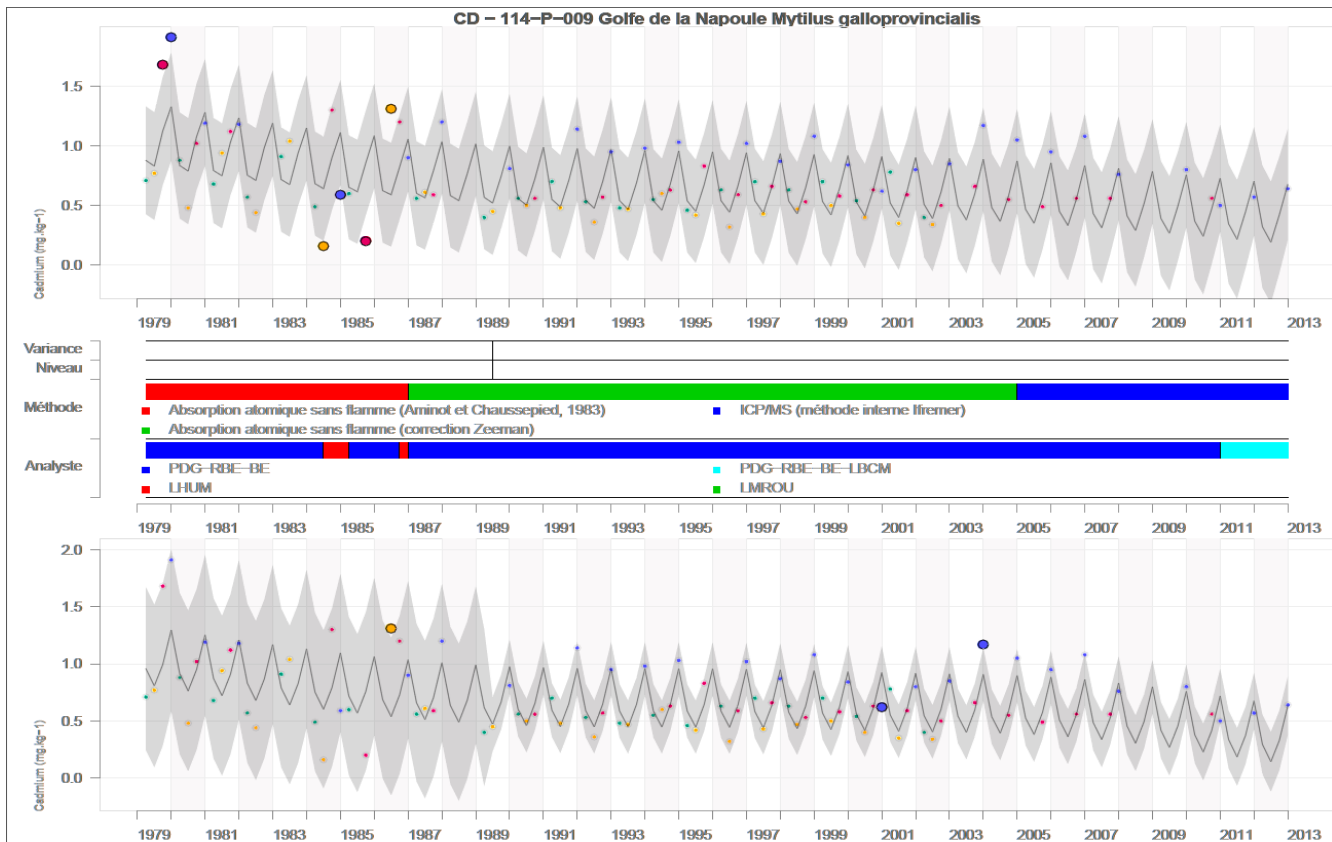


Figure 2 : Example of a graph for Cd (RNO network) for expert qualification. The big points, ringed with black, are the data identified as « outliers » (potentially doubtful or false).

At the end of this expert qualification, data are described as followed:

Initial level \ Final level		Expert Qualification		
		GOOD	DOUBTFUL	FALSE
Automatic Qualification	GOOD	X	X	X
	DOUBTFUL	Possible, but rare (statistical analysis can remove doubt)	X	X
	FALSE	not applicable		

3. Single data or dataset qualification ?

Each data are qualifiable in Q² :

- Metadata : location/time (called « survey »), sampling operation and sample. Each level of metadata can be qualified. For example, informations on geolocation, date, time, depth of water under the boat of a sample can be described as "Good", but a problem with the sample conservation can qualify it as "Doubtful".
- Results : all results of analysis or observation, whether physical, chemical, biological, or even as a file (map layer, files from automatic sensors, photos, etc.) are qualified individually. Each result carries its own quality.

However, qualification process allows to attribute these quality levels by batch of data (usually annual batches), without entering manually each level of quality one by one.

The expert qualification aim at analyze data statistically in **larger lots** (time series for example) because the analysis of the entire history makes it easier to identify outliers. Only **results** are discussed in expert qualification.

3.1 Occasional qualification

For example when a sample has been poorly preserved and the results of the analysis are doubtful or when we found an analytical problem and we know that the results are wrong, these data can be qualified via specific Q² interfaces (**Erreur ! Source du renvoi introuvable.3**).

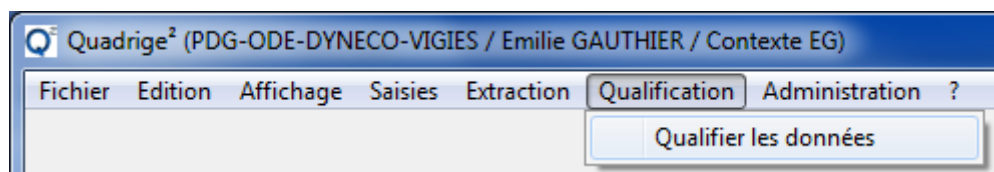


Figure 3 : Access to Q² qualification tool

This tool is used to select data to qualify with query criteria and then display these data in a grid to assign them a level of quality (Figure 4).

Données à qualifier

Commentaire global :

Passer à "bon" les données non qualifiées de la grille

Niveaux

Sélection Récursive Passage Prélèvement Echantillon
 Population Initiale Lot

Passage	Prélèvement	Echantillon	PSFM	Niveau de qualité	Commentaire de qualification
Dannes - 13/02/2013 - 1399	Emergé - Main			Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL		Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	MS%-Bivalve-Chai...	Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	CU-Bivalve-Chair t...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	NI-Bivalve-Chair to...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	ETVPTAIL-Bivalve-...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	CR-Bivalve-Chair t...	Douteux	Méthode analytique non optimale
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	INDVTAIL-Bivalve-...	Non qualifié	

Figure 4 : Q² data qualification grid

This menu is accessible only by people in charge of the quality checks, i.e. monitoring network coordinators who manage their data in Q². They qualify data usually on the request of a data producer.

3.2 Qualification of an homogeneous dataset

3.2.1 Integration of dataset

Some datasets are integrated in Q² as large batches, via computer scripts. According to the source of these datasets, the supplier / producer of the data can define the level of quality of the whole dataset in agreement with the Q² team that manages data integration.

In the two following examples, the level of quality is determined in the computer migration script Q²:

1. *Marine mammal monitoring data, which have been the subject of specific expertise by the producer of data: data can be integrated into Quadriges² with a quality level of "Good".*
2. *Integration of historical data, some important information is missing (the analytical protocol for example): integration of these data with a level of quality "Doubtful."*

3.2.2 Expertise on specific datasets

When a thematic expert works on a specific dataset, whether for a study report, a scientific publication, or other statistical analysis, he can transmit the results of its expertise to the Q² team, and quality levels can be allocated to data.

In this case, the expert defines the scope of dataset with Q² team. He defines the quality levels to be assigned to the whole dataset by the Q² team, (computer language: SQL).

e.g : Qualification of zooplankton data collected under the IGA program (Impact of Large Facilities), thesis work on a theme for a given period, etc.

3.3 Routine qualification of data regularly integrated in Q²

This qualification is for data collected in permanent monitoring networks (REPHY, REMI, ROCCH) and is made **automatically** since 2009 to qualify data every year.

The principle is as follow:

- 1) The qualifiers = thematic experts, define "anomalies" to search into the data (e.g. temperature outliers [out of 0; 30 ° C]).
- 2) Q² team performs extraction of data to qualify (.csv file) and initiates computer programs (software developed under R) to search for these anomalies in the dataset.
- 3) Data without anomalies are immediately qualified as "Good" in the database.
- 4) Data with anomaly (potential) are sent to the data producers (coastal laboratories) for correcting / qualification (.csv).
- 5) Feedback from data producers are centralized by the Q² team that sends them to QC operator(s) for validation (.csv).
- 6) QC operator(s) returns the .csv file of bug fixes / qualified to Q² team that incorporates the corrections / qualifications in database via a script R (running SQL queries.)

Following this "automatic" qualification, an "expert" qualification is performed by analyzing the results using statistical models defined with biostatisticians team. These statistical analysis are performed via R programs, editing graphics in PDF format and data tables associated with the .csv format.

The principle is as follow:

- 1) The QC operators and biostatisticians from DYNECO/VIGIES define the statistical analysis needed to identify potentially doubtful or false data.
- 2) Q² team performs extraction of data to be qualified (.csv) and launches the R computer graphics editing program outputs and data tables to qualify.
- 3) The QC operators analyze these files and assign a level of quality to the data (either they are confirmed good, or they are qualified "false" or "doubtful" with a comment explaining why). Qualified data files (.csv) are referred to the Q² team.
- 4) The Q² team integrates quality levels in the database.

4. List of quality levels and definitions

Q² database contains 4 quality levels :

- 0 = "not qualified" (when data is loading)
- 1 = "Good" (protocol respected and no errors detected)
- 3 = "Doubtful" (non-compliance with the protocol, not confidence in the recorded value)
- 4 = "false" (error detected at the end of test)

The qualification of a data is composed of three informations:

- Quality level (see above)
- Qualification date
- Qualification Comment : mandatory if the level is "Doubtful" or "False".

In Q² application, we recognize the qualified data by a green square next to their symbol (Figure 5).

Note: the green color doesn't indicate that the level of quality is "good", but only a quality level has been assigned.

- ▲  Men er Roue - 11/01/2010 - 12:35:00 - eau
 - ▲  Surface (0-1m) - Bouteille type Niskin tous volumes
 -  Eau filtrée
 -  Masse d'eau, eau brute
- ▷  Men er Roue - 25/01/2010 - 11:45:00 - eau
- ▷  Men er Roue - 08/02/2010 - 11:55:00 - eau
- ▷  Men er Roue - 08/03/2010 - 10:15:00 - eau
- ▷  Men er Roue - 23/03/2010 - 11:10:00 - eau
- ▷  Men er Roue - 06/04/2010 - 11:15:00 - eau
- ▷  Men er Roue - 19/04/2010 - 10:10:00 - eau
- ▷  Men er Roue - 21/04/2010 - 10:00:00 - TOX
- ▷  Men er Roue - 26/04/2010 - 15:00:00 - TOX

Figure 5 : Example of qualified data symbols in Q² : purple, validated data only (Not qualified) and green qualified data.

5. Conclusion

For coastal French data, control / validation / qualification scale is initially local then national:

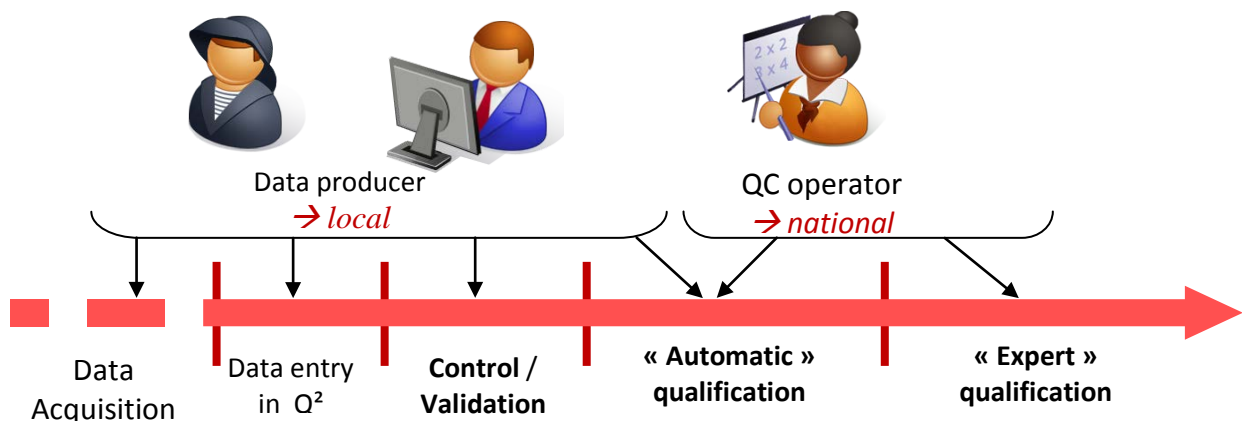


Figure 6 : Role distribution in the qualification process

Data producers are Ifremer laboratories, academia, design office, associations, State decentralized services, and any other structure that collects environmental monitoring data. In the case of "routine" data qualification for Ifremer monitoring, it is mainly the Environment Resources Laboratories (LERs) which are concerned.

The QC operators are the coordinators of monitoring networks, assisted by thematic experts from research laboratories. These experts are recognized internationally.

Data qualification is initiated by the thematic responsible of the data (e.g. monitoring network coordinators). Data Integrated in Q² can thus be described as one of the processes mentioned here above.