HELCOM PLUS

1.0

Technical Document

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 16.03.2016 | 0.1 | PLUS Technical Document | Sriram Sethuraman |
|  |  |  |  |
|  |  |  |  |

Table of Contents

1. Introduction 3

1.1. System Overview 3

1.2. Acronyms and Abbreviations 3

1.3. Points of Contact 3

2. System Architecture 4

2.2 Database and Application Software Utilities 5

3. PLUS Quality Assurance Process and Procedures 5

4. Database Specifications 18

4.1. Physical Design 18

4.2. Physical Structure 19

4.3. User Access Management for PLC Database 19

# Introduction

The HELCOM PLUS project aims to modernize the HELCOM waterborne pollution load compilation (PLC) database, and develop a web application to upload and access the data. The new design changes implemented to the PLC Database would provide a more efficient data system both for reporting and retrieving data derived from pollution discharges into the Baltic Sea.

## System Overview

| System Overview | Details |
| --- | --- |
| System name | HELCOM PLUS |
| System type | Web Application |
| Operational status | In development |
| Database Name | PLC Database |

## Acronyms and Abbreviations

| Acronym / Abbreviation | Meaning |
| --- | --- |
| HELCOM | Helsinki Commission |
| PLUS | Pollution Load User System |
| PLC | Pollution Load Compilation |
| DBA | Database Administrator |

## Points of Contact

| Role | Name | Email | Telephone |
| --- | --- | --- | --- |
| Data Manager | Pekka Kotilainen | [pekka.kotilainen@ymparisto.fi](mailto:pekka.kotilainen@ymparisto.fi) | +358 29 525 1317 |
| Application Developer | Alexander Sokolov | alexander.sokolov@su.se | +46 8 674 7586 |

# System Architecture

The PLUS system has been developed as a web application. Client side is written in JavaScript and html5 which are generated using Java Server Page technology. Client part runs in a web browser and communicates with the server side software consisting of a set of Java servlets communicated with the database.

The PLC Database is developed in MS SQL. The PLUS application interfaces with the PLC Database using JDBC connector (Java Database Connectivity) .The figure below shows the architecture of the PLUS Application.

SU firewall

User machine

Web browser

HTML 5

CSS 3

JavaScript

Application server

Tomcat

helcom\_plus

SYKE server

daily backup

Database server

MS SQL Express

PLC database

Admin machine

MS SQL Server

Management Studio

dbVisualizer Pro

midday backup

midnight backup

Figure 1. Architecture Diagram

## Database and Application Software Utilities

| Vendor | Product | Version | Comments |
| --- | --- | --- | --- |
| Microsoft | MS-SQL Server | Express Edition 2014 | Database Management System |
| Apache | Apache Tomcat |  | Web Server |
| Oracle | Java |  | Application Software |

Table 1: Database Software Utilities

A more complete list of 3rd party softwares is listed [here](http://nest.su.se/helcom_plus/about.html)

# PLUS Quality Assurance Process and Procedures

The PLUS Application provides a Quality Assurance system to obtain additional information on the quality of the reported data. Using this information the contracting parties are able to verify, correct and improve the quality of data, help to ensure a consistent data set. The diagram shown below indicates the various stages of the QA process.

PLC data are reported Annually and Periodically as defined in the templates and in [Annex 2](https://portal.helcom.fi/workspaces/PLC-6-9/Shared%20Documents/PLC-6_Guidelines_Oct2015_fin.pdf) and [Annex 3](https://portal.helcom.fi/workspaces/PLC-6-9/Shared%20Documents/PLC-6_Guidelines_Oct2015_fin.pdf) of the [PLC water guidelines](https://portal.helcom.fi/workspaces/PLC-6-9/Shared%20Documents/PLC-6_Guidelines_Oct2015_fin.pdf)

The Annual and Periodic reporting templates for each country can be downloaded from the PLUS application as a zip file

Periodic Reporting Template

<http://apps.nest.su.se/helcom_plus/data/PERIODIC_REPORTING_2014_20151124.zip>

Annual Reporting Template

<http://apps.nest.su.se/helcom_plus/data/ANNUAL_TEMPLATE_2013_20150226.zip>

Additionally, a sample of the reporting templates can also be found in a zip file at the locations specified below.

Annual Reporting Template

[https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FAnnual%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View={E807520A-BED8-43EB-9868-EBA8A05FFFDA}](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FAnnual%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View=%7bE807520A-BED8-43EB-9868-EBA8A05FFFDA%7d)

Periodic Reporting Template

[https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FPeriodic%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View={E807520A-BED8-43EB-9868-EBA8A05FFFDA}](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FPeriodic%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View=%7bE807520A-BED8-43EB-9868-EBA8A05FFFDA%7d)

Users and Data Manager will fill and approve reporting template before uploading them using the application

**QA Level 0 . Manual format and content verification using the reporting templates**

The QA Level 0 will involve format and content verification by the national experts according to the annual and periodic reporting template, before reporting the data. The templates for reporting data need to be filled according to the guidelines mentioned for annual and periodic reporting.

The template is designed in a specific way that it accepts only a set of predefined options or values for some specific parameters.

For e.g. the subcatchment type should be one of the following:

R = River catchment, L = Land, S = Sea, A = Aggregated subcatchment or Z = unknown (default)

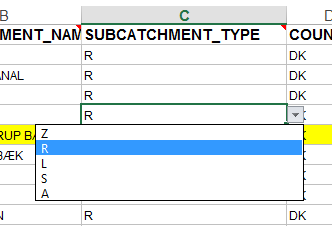


Figure 2. Sample Reporting Template

**Quality Assurance in the PLUS application**

After the data have been filled in the reporting template, it can be uploaded via the upload functionality provided in the PLUS application. Once the upload functionality is performed, uploaded files are stored in the PLC database in the original format (Excel file) in the table TBL\_REPORTED\_DATA as a BLOB (Binary Large OBject, implemented as VARBINARY(MAX) in MS SQL server) together with additional information.

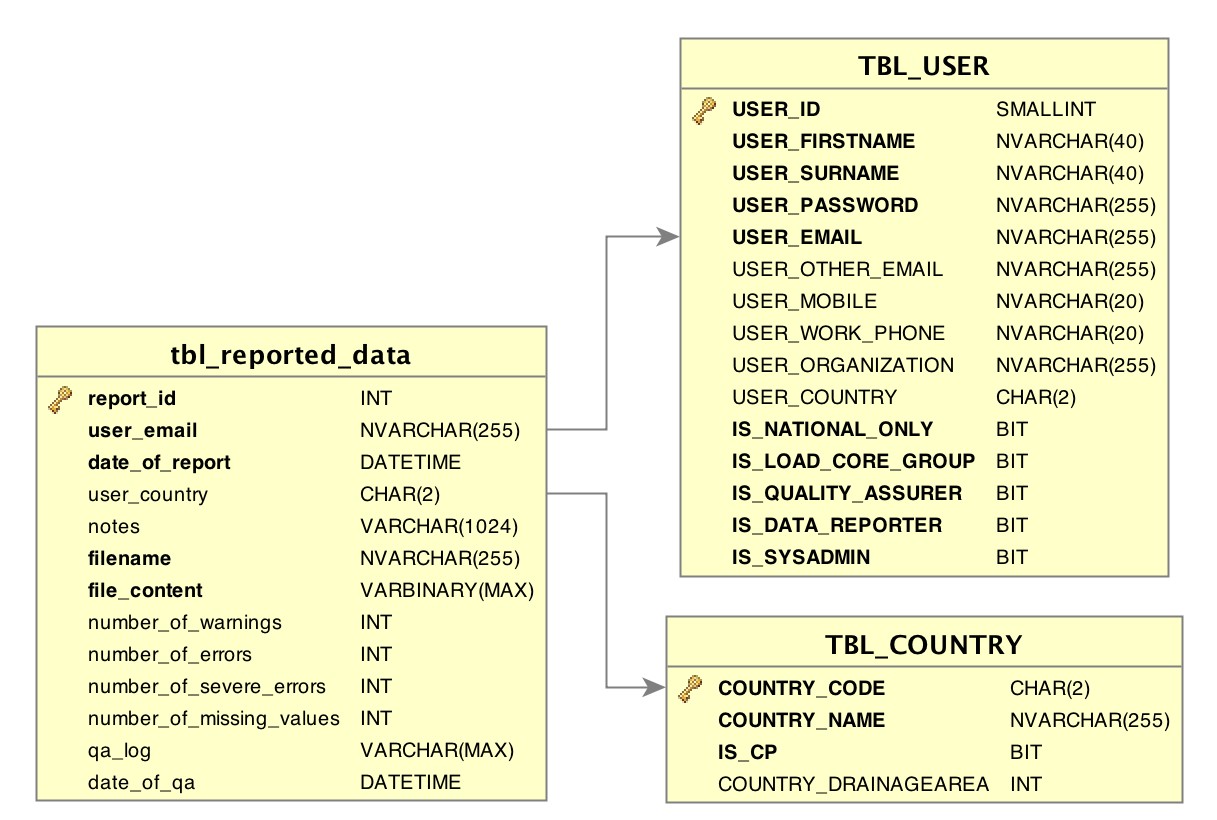
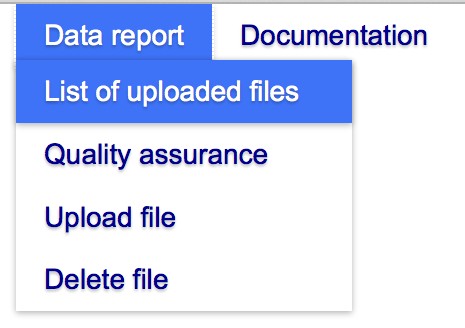


Figure 3. Table Structure of Reported Data

List of uploaded data can be shown using command “List of uploaded files”



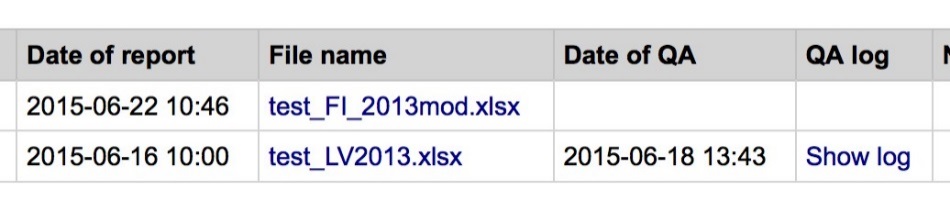


Figure 4. Sample from the application to show a list of uploaded files

**Quality Assurance functionality can be performed using any of the following 3 menu options as shown in the figure 4**

1. Check data
2. Check and insert data
3. Check and insert/update data

Each of these functionality checks for the QA level 1 and QA level 2 procedures, which are described in the next pages.

“Check data” functionality performs QA verification procedures QA 1 and QA2 on the desired data and report the errors, as shown in the sections describing QA 1 and QA2.

“Check and insert data” performs the same QA verification procedures as “Check data”.

This process also inserts new data into the database. Before insertion, the process also performs additional checks on the integrity of the data by checking across the sheets. For e.g. if there is lack of background information or erroneous background information the value will not be inserted.

As an example one can see a section below (QA level 1). All parameters reported as mandatory (value 9) need to be reported according to the specified QA mask.

Any new data inserted into the database will be flagged according to the rules described in the section devoted to the QA Level 2. If the data already exists in the database, there will be an error message stating the existence of data and no update will be conducted .

It is also possible to perform partial insertion of any new data. The partial insertion can be performed for an entity (e.g. station/subcatchment/point sources).

“Check and insert/update data” performs the same QA 1 and QA 2 verification procedures as the “Check data”. This functionality can be used to correct data which are already in the database. The procedure differs from the “Check and insert”, that it allows to overwrite the data, which already exist in the database.

“Check and insert/update data” allows complete or partial update of data. Partial update could vary from one row of data from a sheet to multiple rows and multiple sheets. As an example, it would be possible to insert or update a specific parameter ( e.g. Total Nitrogen load) for a specific subcatchment. Re-reporting of the data is also possible using this option.

The errors at QA 1 need to be corrected to ensure that the specific row of data will be inserted when the insert data is performed.

Also, the “Check and insert data” functionality only flags suspicious or distrust data, but allows them to be inserted into the database. So it is highly recommended to correct the errors at QA 2 before performing a “Check and insert data”.

In case of data insertions or updates, the application also provides the results of insert or insert/update at the end of the notifications in terms of

"Processed nr of rows, rejected nr of them"

As an e.g., of a Commented Excel file, one can see the following data.

*RU\_PERIODIC\_REPORTING\_2014\_DIRECT\_DATA\_20160215094134\_20160215100613.xlsx is available.*

*NOTIFICATION - STATION\_FLOW\_CONCENTRATION: Row 17 (STATION\_CODE=CRU0009, PERIOD\_ID=27, PARAMETER\_ID=20) is already in the database. Update is not allowed.*

*NOTIFICATION - STATION\_FLOW\_CONCENTRATION: Row 18 (STATION\_CODE=CRU0009, PERIOD\_ID=27, PARAMETER\_ID=23) is already in the database. Update is not allowed.*

*NOTIFICATION - Insert/update database: Processed 164 rows, rejected 139 of them.*

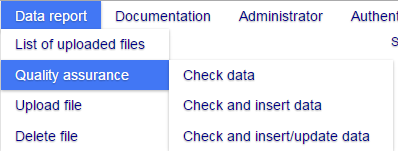
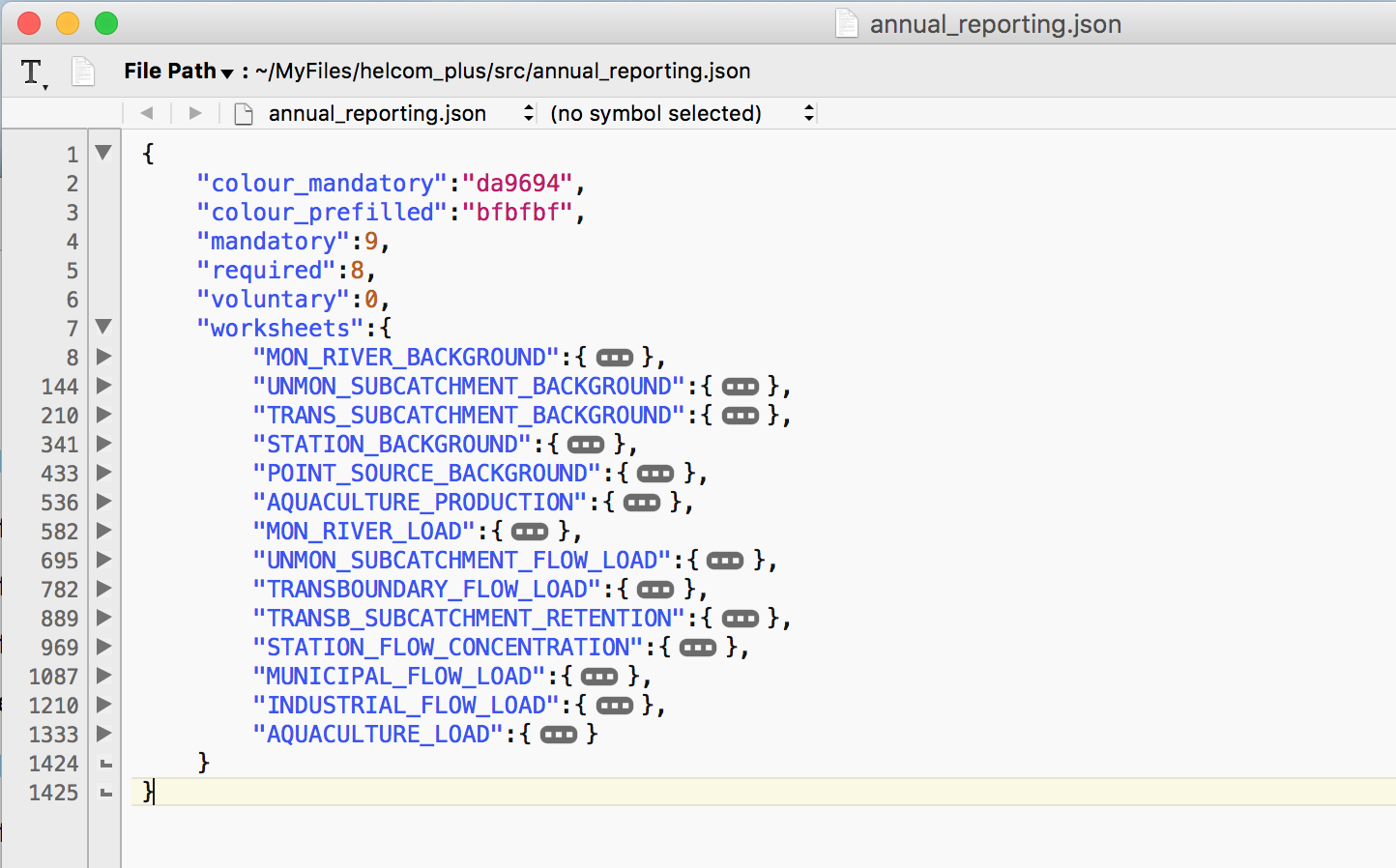
****

Figure 5. Sample From Application depiciting options for Quality Assurance

**QA Level 1. Automatic verification of the format and conformity of the data with the database structure by the application**

QA level 1 involves a sheet wise verification of the annual and periodic data reported in the [annual](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FPeriodic%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View=%7bE807520A-BED8-43EB-9868-EBA8A05FFFDA%7d) and [periodic](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Forms/AllItems.aspx?RootFolder=%2Fworkspaces%2FPLUS-58%2FShared%20Documents%2FProject%20Documents%2FReporting%20Templates%2FPeriodic%20Template&FolderCTID=0x0120009FC51963E16613458543F2D5A8EB517B&View=%7bE807520A-BED8-43EB-9868-EBA8A05FFFDA%7d) templates according to the [reporting guidelines (Annex 2 and Annex 3)](https://portal.helcom.fi/workspaces/PLC-6-9/Shared%20Documents/PLC-6_Guidelines_Oct2015_fin.pdf) and report revealed errors.

All data in the reported template (Excel files) are checked against rules specified for each cell of the Excel file in a JSON file annual\_reporting.json, a tiny part of which is presented below (Fig. 5). The PLUS application checks conformity of data type, are these data obligatory, conformity to the pattern, to the specified numerical intervals.



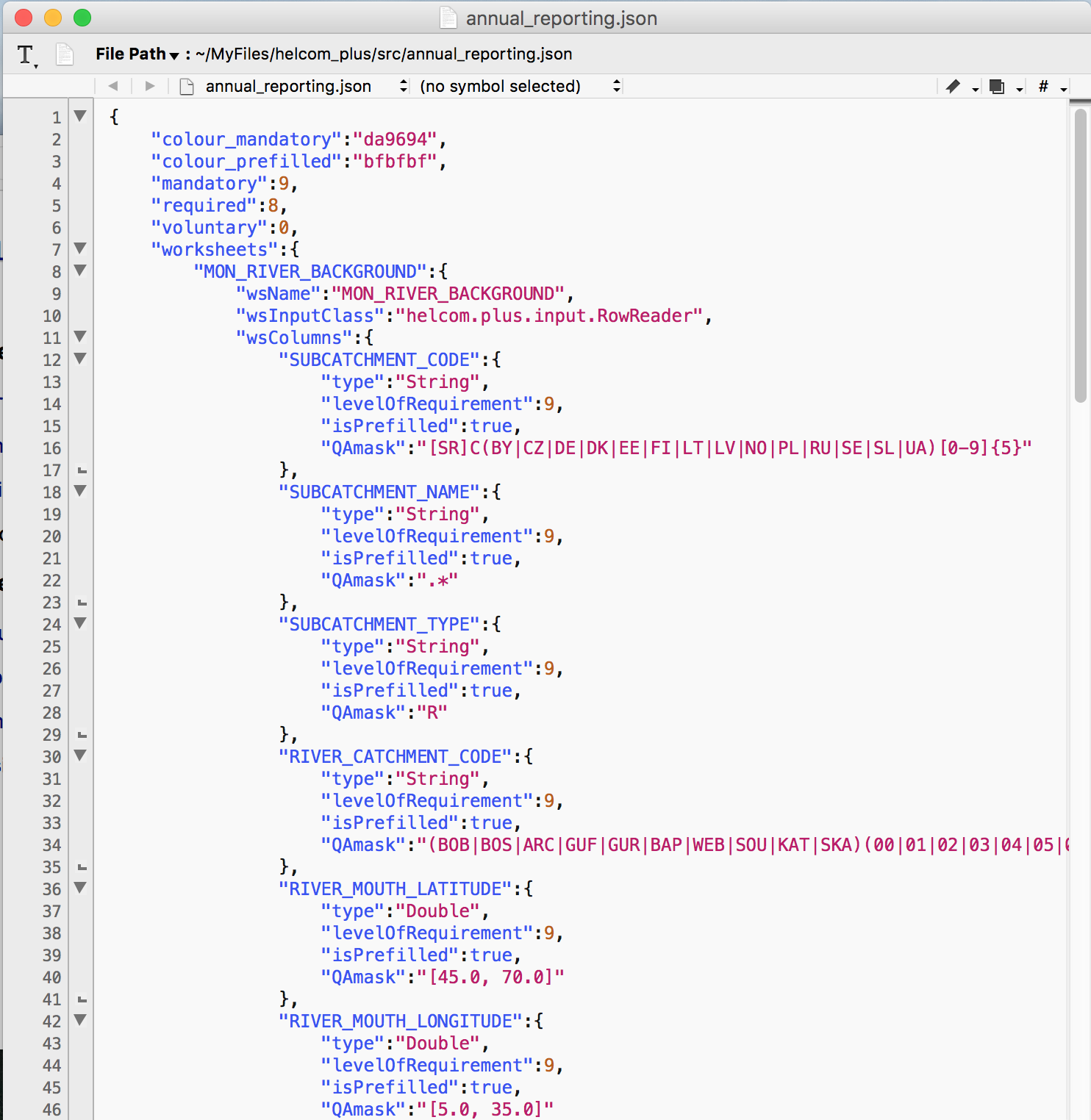


Figure 6. Sample From JSON File depicting rules for Quality Assurance

An example of a single cell checking criteria are given below:

"SUBCATCHMENT\_CODE": {

"type": "String",

"levelOfRequirement": 9,

"QAmask": "SC(BY|CZ|DE|DK|EE|FI|LT|LV|NO|PL|RU|SE|SL|UA)[0-9]{5}"

},

The example shows verification rules of the parameters in the column “SUBCATCHMENT\_CODE”. The subcatchment code needs to be a string and mandatorily reported ( levelofRequirement = 9) and needs to be in the format specified in the QAmask. A valid value for this would be SCDK00010.

A complete template sheet wise listing of QA 1 checks in the JSON file can be found [here](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Project%20Documents/PLUS_QA1_List.xlsx).

A mapping document containing a parameter wise mapping from the reporting templates to the database table names and columns can be found [here](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Project%20Documents/MappingAnnualTemplate.xlsx)

On clicking the “Check data” option, the application will verify and report errors in the log file.

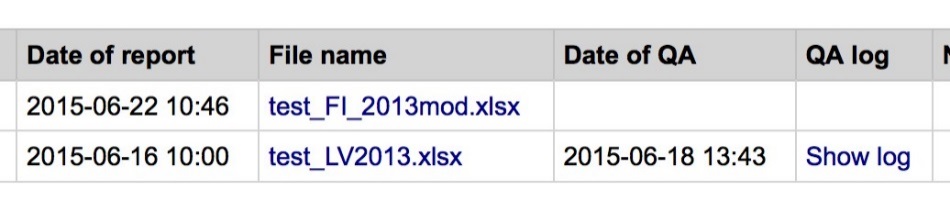


Figure 7. Sample From Application listing the uploaded files

There are 4 types of system responds at level QA1 and this information can be obtained by clicking the “Show log” . On clicking the “Show log” the page with total number of errors, number warnings, number of notifications and number of missing required values will be shown as presented below.

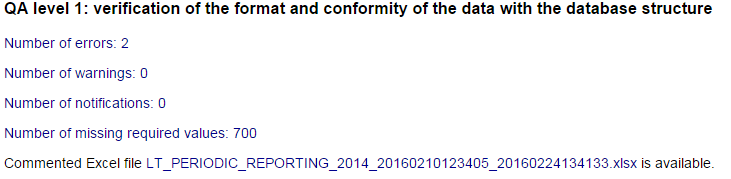
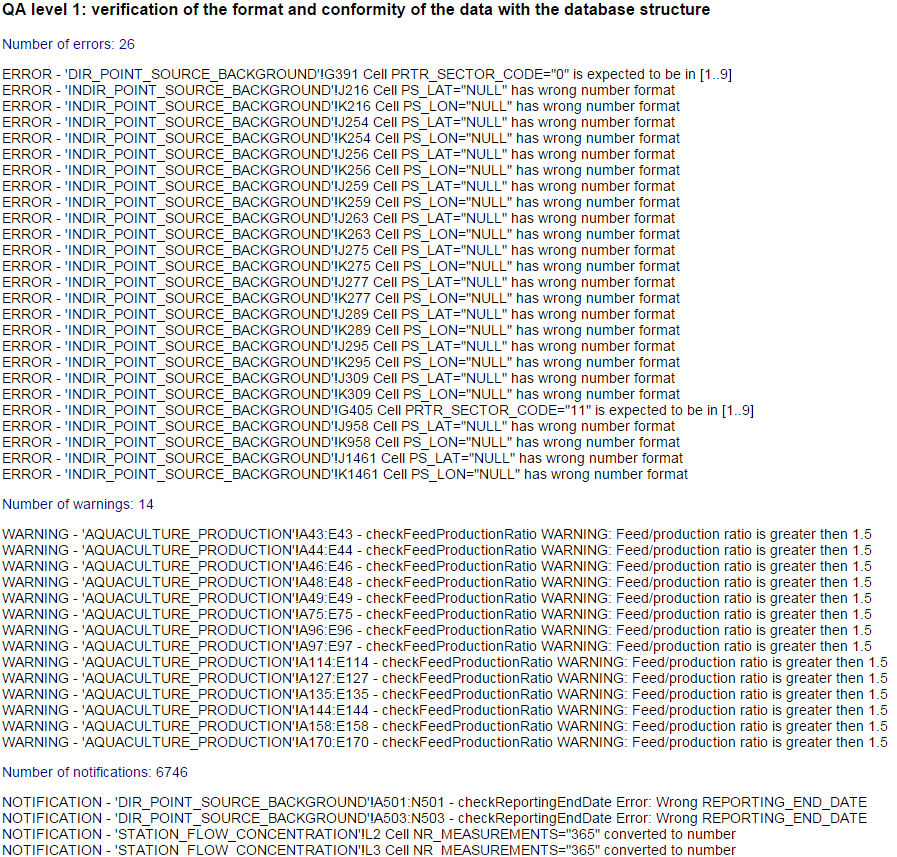


Figure 8. An example of the system respond on QA level 1.

Details of the specific error, warning or notification can be obtained by clicking on the corresponding item: Number of errors, Number of warnings, Number of notifications or Number of missing required values. This results in a detailed listing of all the errors, warnings and notifications.

An error message is reported when a specific data format does not comply with the QA mask or when the values reported are not within a defined range. Such errors need to be corrected before performing an insertion of the data into the database. An example for Error message can be seen on Figure 7.

There are warning and notification messages that are reported in specific cases. Examples of these can be seen in the Figure 7



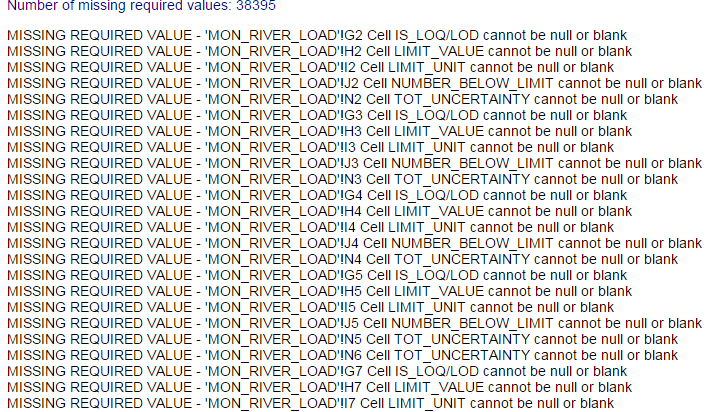


Figure 9. An example of detailed list of QA Level 1 error messages

The errors are also written to the Excel file, which can be downloaded.

**QA level 2: verification of the content for questionable data values**

On the QA level 2 statistical checking procedures are applied to verify data values.

The statistical procedure utilized in the QA 2 is based on the function of Z-score or box-and-whiskers parameters.

The procedure is applied on the VALUE field for the following tables

VAL\_SUBCATCHMENT\_LOAD

VAL\_STATION\_FLOW\_CONCENTRATION

VAL\_MUNICIPAL\_FLOW\_LOAD

VAL\_INDUSTRIAL\_FLOW\_LOAD

VAL\_FISH\_FARM\_LOAD

TBL\_RETENTION\_NEW

The statistical procedure checks the VALUE data against its historical counterparts in the database and flags the specific row of data with Quality Assurance status

(QA\_FLAG\_ID) as NQ (no quality assurance) or AC(accepted) or QU(questionable) depending on the outcome of the statistical procedure.

The QA\_FLAG\_ID is available for all the tables mentioned above.

The statistical procedure which is based on the function of Z-score or box-and-whiskers parameters is described in the section below

|  |  |  |
| --- | --- | --- |
| QA\_FLAG\_ID= NQ | QA\_FLAG\_ID= AC | QA\_FLAG\_ID=QU |
|  | ) and (() | ) and (( ) |
|  |  |  |

is a minimal size of data for using box-and-whiskers

is a minimal size of data for using Z-score ()

is a reported value,

is a mean value,

is a standard deviation,

and are quartiles,

IQR denoting the interquartile range.

The assigned values stand for:

NQ – no quality assurance,

QU – questionable and

AC – accepted.

As an example let us have 5 values

7, 3, 9, 4, 5

Ordering them gives you 3, 4, 5, 7, 9

5 -> median

Q25 = 4

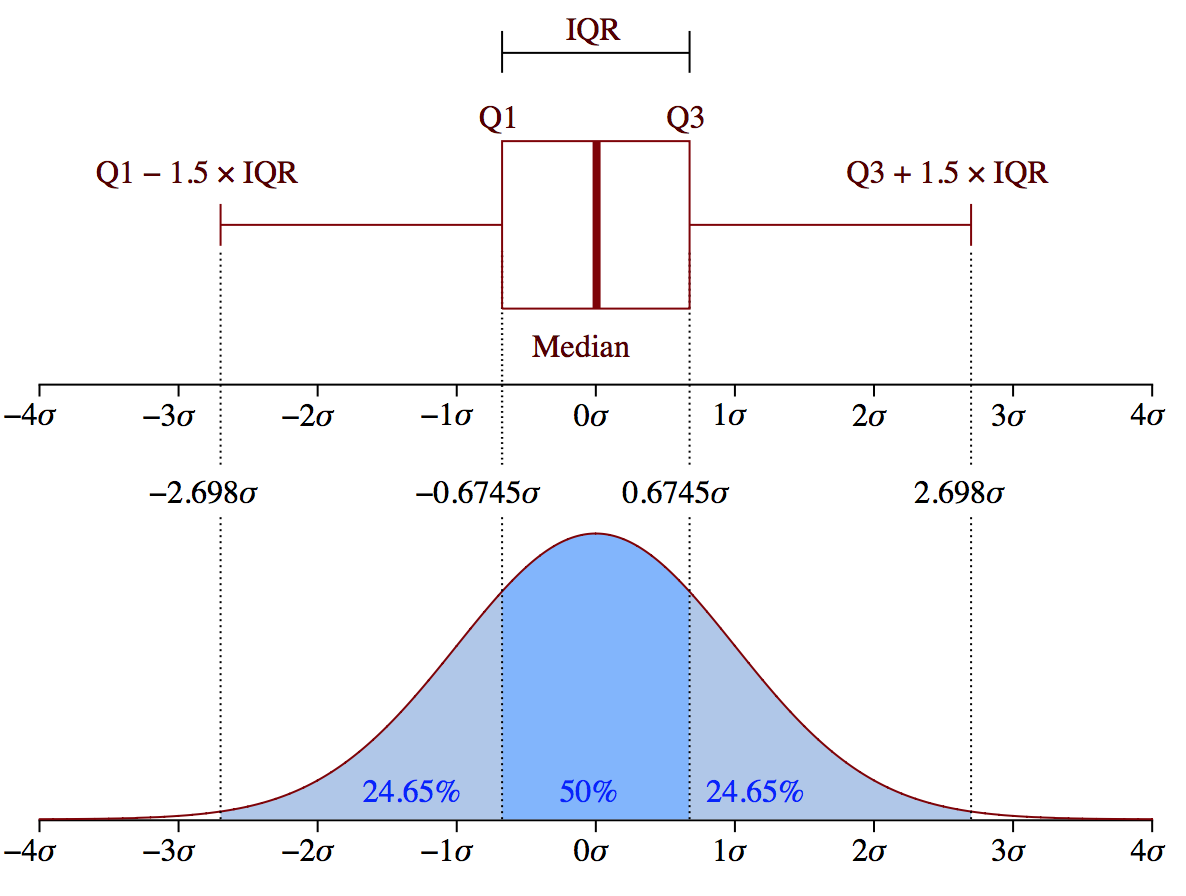
Q25 = 7

Applying the logic , if the reported value is between 0.5 and 11.5, then the value is accepted. If the reported value is outside the range from 0.5 and 11.5, then it is questionable.

At least 5 historical values (e.g. annual data reported 5 times) are needed in order to perform the statistical evaluation. If we have less than 5 years of annual data for the reported value, then the value is marked with no quality flag.

Applying the above defined logic, if the standard deviation is within +/- 2

then the value will be marked as AC



QA level 2 provides the following different categories of questionable data values

Values are marked as distrust if they lie outside of 99.7% band around mean value and suspicious if they lie outside of 95.4% band around mean value

The data reporter can get more details on the distrust, suspicious and unchecked values by clicking the links.

**QA level 2: verification of the content for questionable data values**

[Number of distrust values: 117](http://apps.nest.su.se/helcom_plus/)

[Number of suspicious values: 163](http://apps.nest.su.se/helcom_plus/)

[Number of unchecked values: 6227](http://apps.nest.su.se/helcom_plus/)

[Number of distrust values: 117](http://apps.nest.su.se/helcom_plus/)

DISTRUST VALUE - STATION\_FLOW\_CONCENTRATION row=416: STATION\_CODE=CSE0006 parameter\_id=6 value=37.270 lies outside of 99.7% band around mean value (16.589) (stDev=2.946, n=15)  
DISTRUST VALUE - STATION\_FLOW\_CONCENTRATION row=417: STATION\_CODE=CSE0006 parameter\_id=6 value=38.860 lies outside of 99.7% band around mean value (16.589) (stDev=2.946, n=15)

[Number of suspicious values: 163](http://apps.nest.su.se/helcom_plus/)

SUSPICIOUS VALUE - MON\_RIVER\_LOAD row=208: SUBCATCHMENT\_CODE=SCSE00005 parameter\_id=12 value=1.280 lies outside of 95.4% band around mean value (2.913) (stDev=0.782, n=20)  
SUSPICIOUS VALUE - MON\_RIVER\_LOAD row=410: SUBCATCHMENT\_CODE=RCSE00015 parameter\_id=19 value=2.850 lies outside of 95.4% band around mean value (1.609) (stDev=0.609, n=19)  
SUSPICIOUS VALUE - MON\_RIVER\_LOAD row=452: SUBCATCHMENT\_CODE=SCSE00089 parameter\_id=22 value=13456.000 lies outside of 95.4% band around mean value (8080.632) (stDev=2582.194, n=19)  
SUSPICIOUS VALUE - MON\_RIVER\_LOAD row=507: SUBCATCHMENT\_CODE=RCSE00061 parameter\_id=20 value=191.000 lies outside of 95.4% band around mean value (81.870) (stDev=37.948, n=20)

Figure 10. An example of detailed list of QA Level 2 error messages

The distrust and suspicious values are flagged as “QU” and unchecked values are flagged as NQ.

**QA level 3: Manual verification of the content by the National Data Reporters**

QA Level 3 is yet to be implemented in the PLUS Application.

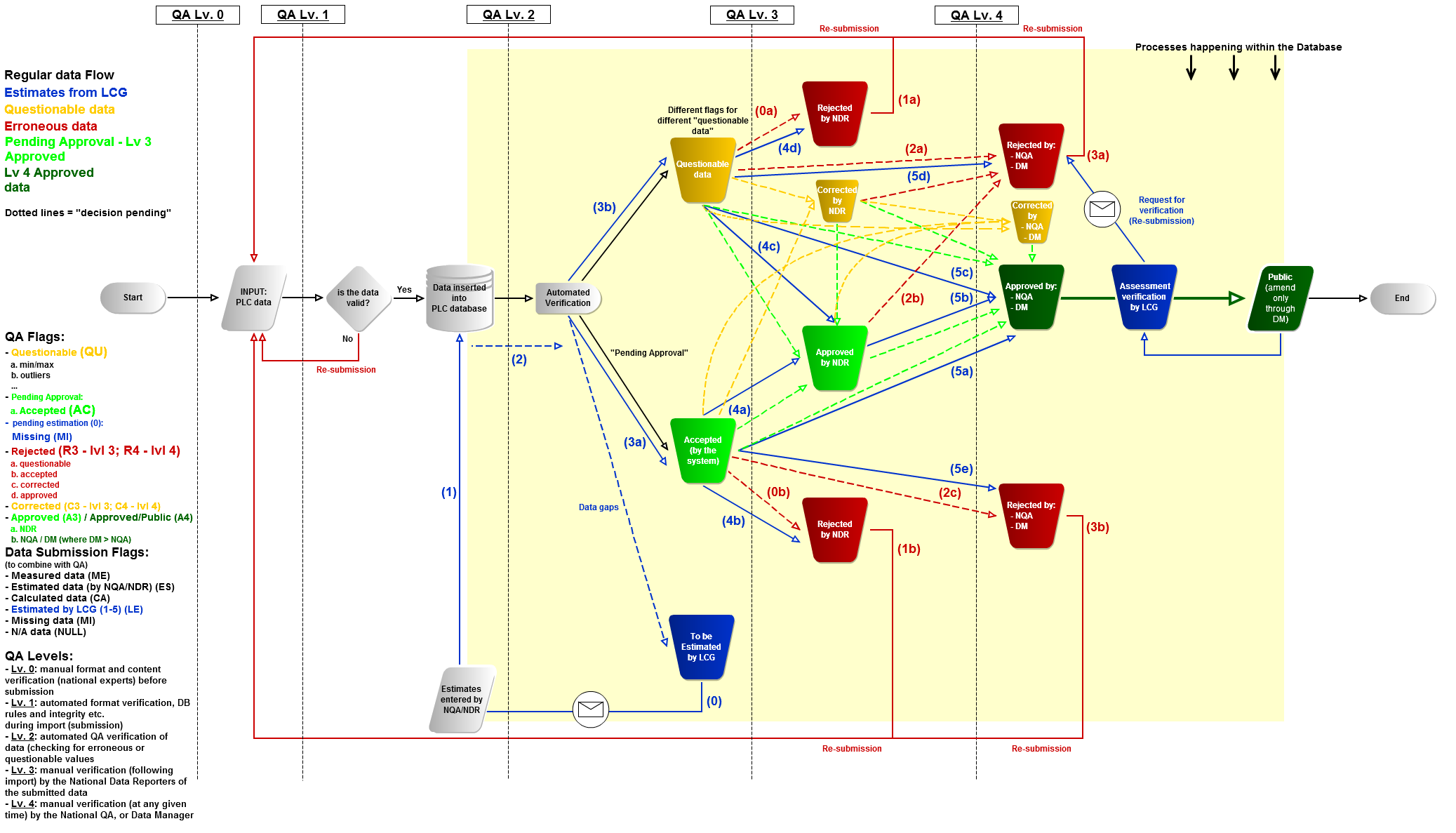
However, if the National data reporters want to manually verify the data they can do so by going through data manually and request the Data Manager to flag the data as “Approved by the Data Reporter” or “Rejected by the Data Reporter” in the database.

**QA level 4: Manual verification of the content by the National Quality Assurer Data Manager**

QA Level 4 is not implemented in the PLUS Application

At this level the National Quality Assurer / Data Manager can manually verify the data and approve or reject them.

The data can be flagged as “Approved by the National Quality Assurer / Data Manager” or “Rejected by the National Quality Assurer / Data Manager” by the Data Manager in the database.

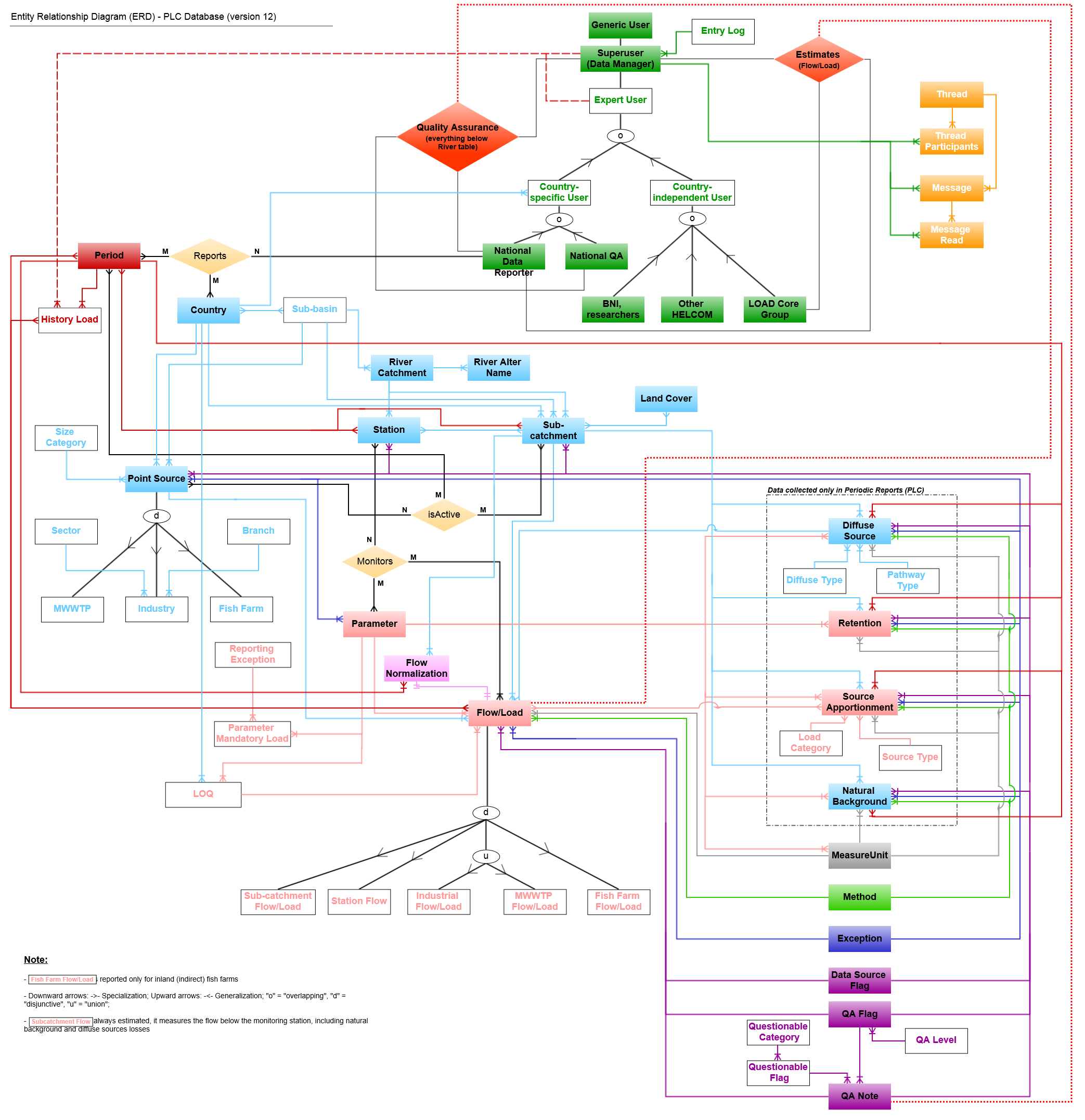


[Figure 11. PLUS QA Process](#_Toc443315131)

# Database Specifications

## Physical Design

Below is the entity relationship diagram, which shows the physical design of the database



[Figure 12. Entity relationship diagram](#_Toc443315131)

## Physical Structure

Link to the database specifications (physical structure of the database, with the tables, relationships and description of the tables and fields) can be found [here](https://portal.helcom.fi/workspaces/PLUS-58/Shared%20Documents/Project%20Documents/Final%20Documents%20for%20PLUS/Data_definitions_FEb_01_2016_PK.xlsx).

## User Access Management for PLC Database

The HELCOM PLUS Application shall support different options for users to view and modify the data depending on the user profile.

The User categories and profiles are stored in the PLC database in table TBL\_USER.

The TBL\_USER contains the following details regarding the user

User id (automatically generated by the database)

First name(s) of the user

Last name(s) of the user

Email address of the user

Alternative email address of the user

Mobile number of the user

Work phone number of the user

Organization to which the user belongs

Country of the user

Currently, there are 2 key user profiles

System Administrator

Data Reporter

The System Administrator is the user with the highest privileges (administrative).

He has privileges to add and delete users. He also has rights to view all the Data uploads.

The Data Reporter is able to upload data, view the data uploads for his own country and insert the data into the database.

Following is the list of other user profiles that will be implemented in the future.

National Only

Red Core Group

National Quality Assurer

National Data Reporter

In addition to this the table will help in identifying the user profile as well.

For more details, please refer to the data model in Section 4.2