Document title	Database Functional Design Document
Code	4-1
Category	DEC
Agenda Item	4 – Database Functional Design Specifications
Submission date	20.2.2014
Submitted by	HELCOM PLUS Project Team

Background

The attached document contains the PLC Database Functional Design Specification which lists all the functionalities that the PLUS Project Team proposes to implement in the first release of PLUS.

Action required

The Meeting is invited to review and finalize the Database Functional Design Specification document.

The Meeting might focus on the automatic quality assurance functionalities, scrutinizing if they are sufficient and applicable.

HELCOM PLUS

1.0

Database Functional Design Document

Oct 2013

Revision History

Date	Version	Description	Author

Table of Contents

1. Introduction

The HELCOM PLUS project aims to modernize the HELCOM waterborne pollution load compilation (PLC) database, and develop a web application to access the data. The new design changesimplemented 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.

1.1. Purpose

This Functional Database Design document provides detailed infomation of the PLC data model implemented to support the functional requirements for HELCOM PLUS target database management system with consideration to the system's performance requirements.

The document describes, how the database that will support the [Application] Data Model with details of the logical and physical definitions.The document provides the functional and non-functional usage of the tables, considerations and requirements.

Further, the document would briefly describe the integration aspects of the Database with the Web Application . The Web Application would provide the users with easy access to PLC data.

1.2. Scope, Approach and Methods

The Database Design for the [Application] is composed of definitions for database objects derived by mapping entities to tables, attributes to columns, unique identifiers to unique keys and relationships to foreign keys.

During design, these definitions may be enhanced to in order to support the requirements of the PLUS application listed in the *Requrements Traceability Matrix* .

The document shall also describe the database changes pertaining to the requirements listed in the *Requrements Traceability Matrix* and also briefly describe, how the specific requirements will be designed and implemented structurally in the database.

1.3. System Overview

System Overview	Details
System name	HELCOM PLUS
System type	Client Server Application
Operational status	In development
Database Name	PLC Database

1.4. Acronyms and Abbreviations

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

1.5.Points of Contact

Identify the points of contact that may be needed for informational purposes.

Role	Name	Email	Telephone
Project Manager	Sriram Sethuraman	Sriram.sethuraman@helcom.fi	
Data Manager	Pekka Kotilainen	pekka.kotilainen@ymparisto.fi	
System Specialist	Marco Manzi	Marco.Manzi@ymparisto.fi	
Database Administrator			

Table 1: POC Contact Information

2. System Overview

The diagram shown below indicates the Data Flow Diagram for the PLUS Application. As one can see, the National Data Reporters use the Web Application to input the data to the PLC Database. This is done using a standard reporting template in the form of an Excel file. The data passes through a set of QA processes, before it is finally available in the database. The QA process involves a series of steps , including data verification from National QA's, and providing estimates for data gaps At the end of the QA process, the data is finally approved.

As for the visualization aspect, the approved data is made available to the end users (NGO's, scientific institutes, decistion makers etc.) in the form of tables, graphs and reports. Users can access the data via web interface from a public URL accessible via the HELCOM website.

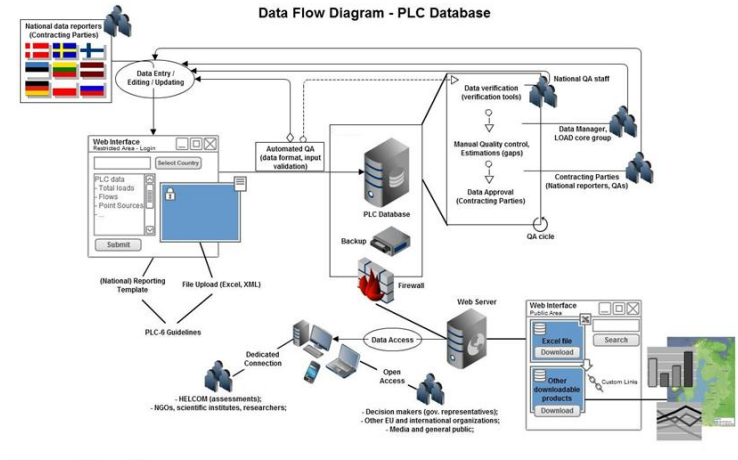

Figure 1. Data Flow Diagram

2.1. Quality Assurance Process

The PLUS Application will provide a Quality Assurance system to ensure a minimum level of quality to the reported data. The diagram shown below indicates the various stages of the QA process. The QA Level 0 will involve manual format and content verification by the national experts, before reporting the data.As shown in the figure, QA level 1 will verify automatically the format and conformity of the data with the database structure (logical schema). QA level 2 will verify the content for questionable data values, meaning possible outliers or other values which could be potentially incorrect. QA level 3 will provide the National Data Reporters with the option of manually verifying, correcting and approving the data. QA level 4 will involve in a similar fashion the verification from National Quality Assurers, including the final approval of data to be used for assessments and made accessible to the public.

For more details on the QA process, please refer to section 4.1 of the document containing Requirement Id 30 and Requirement Id 31.

Comment [MM1]: At later stage a separate document of QA should be made.

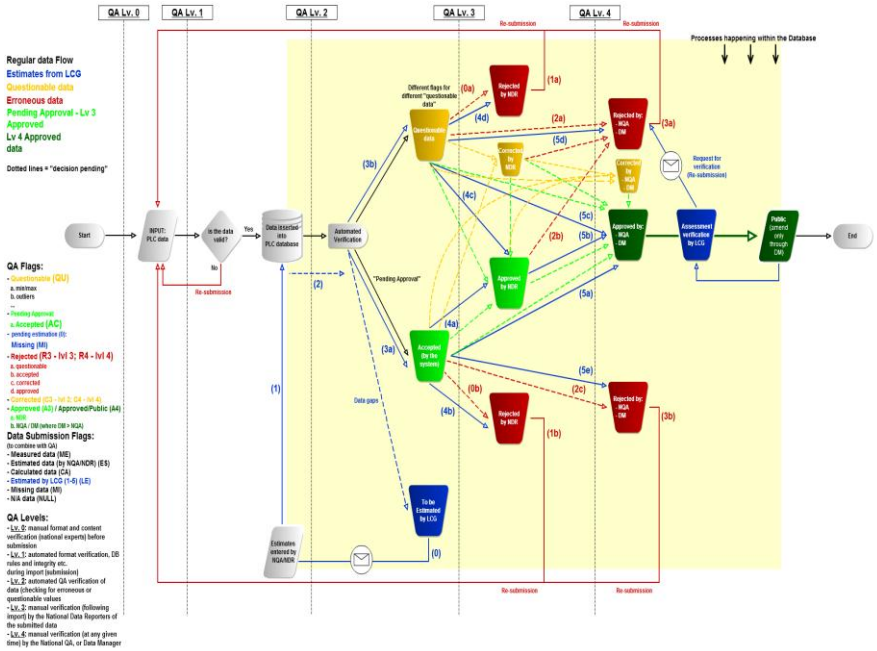

Figure 2. QA Process

2.2 Database Software Utilities

Identify any utility software that will be used to support the use or maintenance of the database.

Vendor	Product	Version	Comments
Microsoft	MS-SQL Server	Standard Edition 2012	Database Management System

Table 2: Database Software Utilities

3. Database Specifications

3.1. Physical Design

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

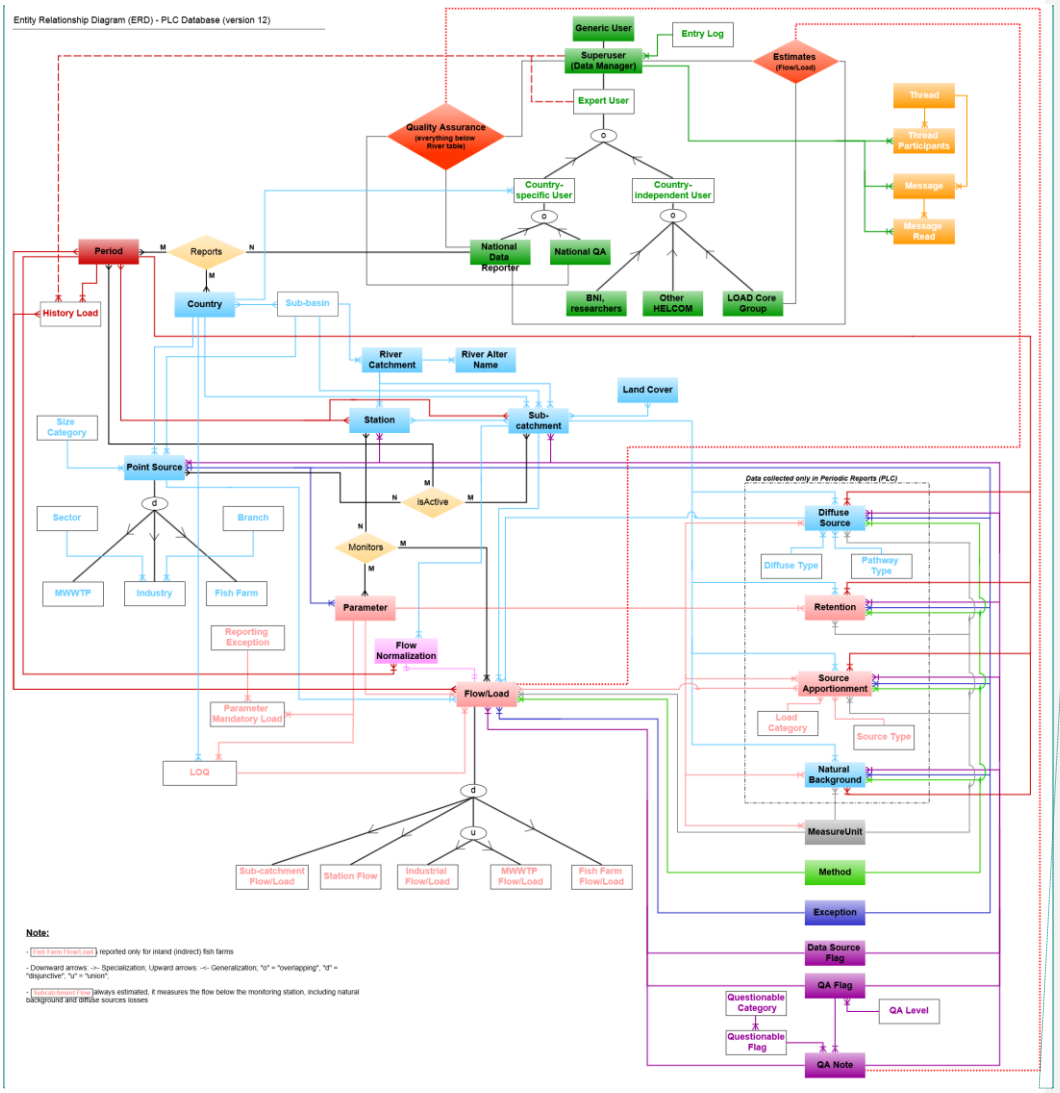

Figure 3. Entity Relationship Diagram

Comment [MM2]: to be updated

3.2. Physical Structure

The excel attached here provides the exact physical structure of the database, with the tables, relationships and description of the tables and fields.


Data definitions Feb 19 2014

4. Database Design and Functionalities

4.1. Design & Functional Support

The database has been designed meet the below listed functional requirements.

A separate Use Case document is available under in the meeting portal the below location for all the Functional requirements listed below.

Data retrieval

Requirement Id # 2 – Priority - High

Easy retrieval of information from new PLC Database about catchments, stations, point sources (in order to check with my national information)

Information regarding catchments is stored in the tables TBL_RIVER_CATCHMENT and TBL_SUBCATCHMENT. These include border and transboundary rivers.
Information regarding monitoring stations is found in the table TBL_STATION. A subcatchment can be linked to 0 (when sea or coastal area) or more stations, but from 0 (unmonitored area) to 1 station can be active in a subcatchment during a period of time.
Information about point sources is contained in TBL_POINT_SOURCE. Point sources, during a reporting period, can be either located in a monitored subcatchment (in which case they are linked in a many-to-many relationship with TBL_SUBCATCHMENT and TBL_PERIOD), or they can be direct, in this case belonging to a sub-basin for a specific country.

The information and related metadata can be easily obtained by querying the database. For more details see Section 3.3 on the list of information pertaining to structural implementation.

The river catchment table contains the name of the river, type of river (country, boundar or transboundary) and the coordinates of the river mouth. These are necessary to identify the country where the lowest (i.e. closest to the sea) monitoring station is established.

Requirement Id # 3 – Priority -High

Comment [MM3]: Due to the new HELCOM sub-basin division this might have to be revised (more info after the PLC workshop Feb. 24-26)

Nomenclature consistency over time for the established sub-basins (e.g. Baltic Proper, Kattegat, Gulf of Finland)

The naming of the sub-basins are specified according to the definitions contained in the PLC-6 Guidelines, as shown below

Sub-basins		Abbreviation
1. GULF of BOTHNIA		**GUB**
1.1	Bothnian Bay The Quark	BOB
1.2	Bothnian Sea	BOS
1.3	Archipelago Sea	ARC
2. GULF of FINLAND		**GUF**
3. GULF of RIGA		**GUR**
4. BALTIC PROPER		**BAP**
4.1	Northern Baltic Proper Western Gotland Basin Eastern Gotland Basin	BPN
4.2	Southern Baltic Proper Gulf of Gdansk Bornholm Basin Arkona Basin	BPS
5. BELT SEA and KATTEGAT		**BSK**
5.1	Belt Sea	BES
5.1.1	Western Baltic	WEB
5.1.2	The Sound	SOU
5.2	The Kattegat	KAT

The database stores the codes (abbreviations) for each sub basin as shown in the above table.

Requirement Id # 4 – Priority- High

Easy retrieval of information on a point source from PLC Database, even though its name has changed

The database stores relevant information with regard to point sources in the TBL_POINT_SOURCE and in the tables TBL_INDUSTRY, TBL_MWWTP and TBL_FISH_FARM.
A point source is primarily identified using the PLANT_CODE - which is a combination of the Point source type (Fish Farm (F), Municipal Waste water (M), Industrial Waste (I)) + country code + unique id number - as well as the PERIOD_ID, in order to identify changes in relevant data. These data include, among others, the name of the point source.
As such, it is possible to retrieve information on the point source even if the name of the plant has changed.

Requirement Id # 5 –Priority- Medium

Ablility to check historical (previous) point sources from PLC Database

The table TBL_POINT_SOURCE contains the following fields:
ACTIVITY_START_DATE (Start date for the monitoring activity related to a point source) and ACTIVITY_END_DATE (End date for the monitoring activity related to a point source).

When a point source activity is not relevant for monitoring purposes (low emissions), or the outlet is closed, the old data related to a previously entered point source will still be available, as it is stored in the database.
When a point source is "reopened" (or parameter-specific loads need to be monitored again) on a certain date, this date becomes the new ACTIVITY_START_DATE value, and the ACTIVITY_END_DATE is reset to NULL. This information, together with the PERIOD_ID, allows to track the relevant activities of a point source in time.

Requirement Id # 7-Priority-Medium

Able to calculate the normalized flow and loads based on aggregated data

The PLC database provides the load and flow data required in order to perform the flow normalization calculations. Such Data is stored in the tables collecting load, flow and concentration values

VAL_SUBCATCHMENT_LOAD
VAL_STATION_FLOW_CONCENTRATION

For more detailed information, please refer to the PLC Database Structure defined in 3.2

The normalized flows are calculated using the techniques provided in the PLC guidelines.
For more information, please see Section 5.3 of the below document
http://helcom.fi/Lists/Publications/BSEP128.pdf

Requirement Id # 9 – Priority - Medium

Possibility to modify the content of the database

It will be possible to modify the data in the database using the Edit option available in the new web application. However, modifications can be carried out according to the respective user rights and the QA assurance process which is described in requirement Id #30.

National Data Reporters and National Quality Assurers will have the credentials to edit their respective national data, with the latter having higher privileges. As such, a Data Reporter cannot modify data which has been previously approved by the Quality Assurer. The PLC Data Manager will have unrestricted access to the database, with the possibility to modify all the content as need arise.

For the detailed description of tables and fields, please refer to the database specification document on section 3.2.2

Requirement Id # 10 –Priority - Low

Possibility to modify the structure of the database

Modification of database structure will be analyzed for costs and impact to the PLUS application, and if approved, handled as a PLUS Change request or in a separate maintenance release. This is due to the fact that structural changes to the database will affect also the web application.

Requirement Id # 11-Priority - High

Contracting Party to be able to upload data (partial data, or complete annual set) directly into the database

Data upload (partial and/or a complete dataset) to the PLC database is possible via upload operation. The upload operation will be performed using the web application, resulting in a predefined sequence of INSERT, UPDATE or DELETE operations to the PLC Database. Aannual reports can be uploaded in Excel format via the web application user interface by the national data reporter. The specifications and detailed instructions on how to fill in the Excel reporting file correctly will be available on the updated PLC-6 Guidelines.

Data which have been reported in the correct format and within the constraints described in the quality assurance process, will be either inserted anew in the PLC database, update existing records, or otherwise marked as rejected, to be deleted manually from the web application at later stage by the Data Reporter him/herself, or the Quality Assurer, after verification.

Please refer Req Id 11 in the Use case document.

Requirement Id# 12 – Priority - High

Able to add comments on data, when the data is missing or questionable.

The PLC database includes a Quality Assurance mechanism (Requirement #30) to ensure that the data entered has at least a certain level of quality and reliability.
The definition of missing data is related to data which is considered mandatory from the PLC Guidelines, but for some reason hasn't been yet provided by the Contracting Party(ies). This is in opposition to "Not Available" data (N/A), which is instead coded as NULL in the database. It is possible for the national experts, upon consultation with the Data Manager, to modify the flag of missing data as "Not Available", in case these cannot be provided, together with the reason why. In this way, national reporters won't be requested to fill in missing data, which they are not able to provide for some particular reason.

For missing data, the information is stored in the field DATA_STATUS_FLAG_ID. This is available in the following tables:

VAL_SUBCATCHMENT_LOAD
VAL_STATION_FLOW_CONCENTRATION
VAL_INDUSTRIAL_FLOW_LOAD
VAL_FISH_FARM_LOAD
VAL_MUNICIPAL_FLOW_LOAD

TBL_SOURCE_APPORTIONMENT
TBL_DIFFUSE_SOURCE
TBL_RETENTION
TBL_NATURAL_BACKGROUND

In addition to this, the user is able to add comments to the data when it is questionable. This is managed in the PLC database using the following QA tables:

QA_LEVEL
QA_FLAG
QA_QUESTIONABLE_CATEGORY
QA_QUESTIONABLE_FLAG
QA_NOTE

The QA_FLAG and QA_NOTE tables are related to the different load, concentration and flow tables using a QA_FLAG_ID and a QA_NOTE_ID. Data reporters and Quality Assurers, as well as the Data Manager, are then able to add comments for questionable and missing data in the load, flow and concentration tables.

For more details on the table relationships, please refer the data model in Section 3.2

Requirement Id# 13 – Priority - High

Able to modify previously entered data in the database

The users shall be able to edit the previously entered data using the edit option in the web application. This would translate to an update query on the database.

The users shall be allowed to edit only the data that he or she is authorized to, depending on his or her nationality, role, and/or respective user rights. For more details on user privileges, please see section 4.2 (User Management).

Requirement Id# 15 – Priority - Medium

Able to report different unmonitored areas or monitoring stations according to different parameters (e.g. total N, NO23-N, Ni, discharge etc.)

NOTE: Varying unmonitored areas of subbasin have been added to the structure in order to allow reporting of varying areas by parameter. Testing of the structure is going on.

Requirement Id# 19 – Priority - Medium

Allow reporting of data based on individual point sources including their coordinates.

Point sources are reported via the table TBL_POINT_SOURCE, which includes a field PLANT_TYPE to identify the type of point source. The point sources can be in this way categorized into one of the following:

- I = Industry,
- M = Municipal wastewater treatment plant
- F = Fish farm.

Specific information related to these plant types are collected in TBL_INDUSTRY, TBL_MWWTP and TBL_FISH_FARM, respectively.

The coordinates of a point source can be provided by the data reporters through the fields PS_LAT (Point source Latitude) and PS_LON (Point source Longitude), when the point sources are reported as individual (and when it is allowed under national legislation). Normally the coordinates would indicate the location of the outlet, except in the Russian case, where these indicate the city (or municipality) where the point source is located.

Comment [MM4]: to be verified with Natasha

Requirement Id# 21 – Priority - Low

Able to report point sources in an aggregated way.

Typically small point sources are reported not as individual sources, but as aggregated. This is implemented in the database via the SIZE_CATEGORY_ID field in the TBL_POINT_SOURCE, which refers to the following codes, grouping the point sources by their size (clear definition on how to categorize by size the point sources will be included in the PLC-6 Guidelines):

BI – Big Industry
SI – Small Industry
AI – Aggregated Industry
BM – Big Municipal Waste Water treatment plants
SM – Small Municipal Waste Water treatment plants
AM – Aggregated Municipal Waste Water treatment plants
BF – Big Fish Farms
SF – Small Fish Farms
AF – Aggregated Fish Farms

Thus, Industries, WWTPs, and Fish Farms, can all be reported as aggregated if needed.

Requirement Id# 22 – Priority - Low

Able to report monitored and unmonitored areas, point sources, aggregated point sources, and combination of these for different catchments in each year

The annual reporting form allows to submit data by defining monitored sub-catchments and unmonitored areas Country-wise and per sub-basin. These are listed under the table TBL_SUBCATCHMENT, and identified primarily by a unique combination of SUBCATCHMENT_CODE and PERIOD_ID. The field IS_MONITORED specifies whether the area is a monitored catchment (in which case an active station – via table AGG_STATION_SUBCATCHMENT - should be linked to it) or an unmonitored area.

Point sources, individual and aggregated, can be reported annually when considered as directly discharging to the sea, while point sources located in a monitored catchment are included in the monitored loads, concentrations and flow measurements (in annual reporting). Point sources are stored in the TBL_POINT_SOURCE, and specific information related to different types of point sources can be found in TBL_INDUSTRY, TBL_MWWTP and TBL_FISH_FARM.

Please see Requirement 22 in the Use case document.

Requirement Id# 23 – Priority - Low

Able to include the calculation method for aggregated data when submitting them into the database

The type of calculation method used is stored in the DEF_METHOD table, including a method Id, name of the method used, and type of method (calculation, estimation, retention, etc.). A list of the recommended methods, including their description, can be found in chapters 2 & 3 of the PLC-6 guidelines.

Each of the tables collecting flow, load and concentration measurements (starting with the prefix "VAL") include a field called METHOD_ID, identifying the method of calculation (or estimation) used.

Requirement Id# 24 – Priority - Medium

To be able to include the calculation method for estimated data (e.g. to fill in data gaps) when submitting them into the database

The user shall be able to specify the method used for the data, irrespective of whether these have been physically measured or estimated. See requirement #23 for details on how this information is collected. If the methodology differs from those recommended in the PLC-6 Guidelines, it should be described in detail and provided via a written document (preferably in MS Word format) by the end of the reporting year. This is necessary also to provide further information when performing manual Quality Assurance (verification) on the data.

Requirement Id# 26 – Priority - Medium

Able to administer user rights (granting/denying actions to other users)

The Data Manager has full priviledges on the PLC database, and as such isable to administer user rights for all the other users of the system. National experts should also be able to delegate priviledges within their own national system or limited parts of it for national users (experts). For instance a Swedish National Quality Assurer should be able to allow a reporter who is in charge of submitting fish farm data, to perform different operations on such specific data, without the need to ask to the data manager to grant the required permissions.

Please see section 4.2 for additional information.

Requirement Id# 28 – Priority - Medium

Retrieve a report about which data has been modified, and indicating by whom

The system could store information on changes related to the data performed by users, if needed, via log files. Such information would include whether the values have been inserted, deleted, or updated, a timestamp, and the person who performed the specified action.
The PLUS system will provide the possibility to retrieve the information regarding which data has been modified and when using a report, when the Application is developed in Work Package 5. The information regarding this can be stored in the system tables provided by SQL Server.

Requirement Id# 29 – Priority - Medium

Able to retrieve a report about which data is missing from the report in the database

As a result of the import process (data reporting), among the possible warnings displayed to the user after uploading the reporting form, information about missing (mandatory) values is also included, and coded accordingly.
Missing data is thus flagged and indicated as missing. This information is contained in the DATA_SOURCE_FLAG_ID field, specifically in the value "MI" (missing mandatory data).

Once the data is inside the database, this can be queried for retrieving Country-specific information (or on a finer level of detail, even sub-basin or source-specific), showing what data is missing and should be provided in accordance with the PLC-6 Guidelines.
Missing values can be provided for the following entities (tables):

VAL_SUBCATCHMENT_LOAD
VAL_STATION_FLOW_CONCENTRATION
VAL_INDUSTRIAL_FLOW_LOAD
VAL_FISH_FARM_LOAD
VAL_MUNICIPAL_FLOW_LOAD
TBL_SOURCE_APPORTIONMENT
TBL_DIFFUSE_SOURCE
TBL_RETENTION
TBL_NATURAL_BACKGROUND

The tables above are used for reporting loads, flows and concentrations from monitored rivers, unmonitored areas and point sources.

Requirement Id# 30 – Priority - High

New PLC Database to have a built-in automated quality control mechanism warning about possible questionable data

A Quality Assurance system will be implemented to provide a certain minimum level of quality on the reported datasets. The general flow of the quality assurance for the PLC database is described in the figure 2 (Section 2.1).
Specifically, the automated part of the quality assurance is implemented in QA level 1 and QA level 2. The preliminary step of Quality Assurance (QA level 0, see HELCOM PLUS 4/2013 document 4/2 http://meeting.helcom.fi/c/document_library/get_file?p_l_id=80219&folderId=2467787&name=DLFE-55030.pdf) and the following levels (QA level 3 and 4), are to be performed by national experts, either prior to the data submission (level 0), or via the web application which will be created to utilize the new system (3, 4).

Comment [PK5]: QA level 0 should be listed at least as examples.
-ATTRIBUTE FORMAT and LENGTH
-PRE-DEFINED CODES
-DATA TYPE

In QA level 1, the focus will be on verifying that the data is in the correct format, that the data integrity, constraints and other logical rules to ensure the correct functioning of the database are all preserved, and the data as such will be either accepted and stored into the database, or rejected altogether (in which case it will generate an error-specific message, identified by an error code and a human-readable message, including the data that has generated it) and will thus need to be re-submitted after being corrected.
Examples of data that would not be entered in the database are:

- Codes not following the correct format, or inexistent.
- Country, sub-basin, and other area sizes when provided must be >=0
- Wrong data type (e.g. text string inserted where integer number is expected)
- Negative data values (except for retention)
- Violation of primary keys constraints (double entries)
- Violation of foreign key constraints (referential integrity)
- Missing mandatory fields (indicated as "NOT NULL")
- Codes provided are out of range of available codes

- Violation of other database constraints

Passing level 1, the QA level 2 is performed also during data submission, but only on the data that has been deemed correct in respect to the specific format and limitations of the database structure. These data will be verified against some specific conditions (or rules), indicating whether the data could be considered scientifically correct, or deemed suspicious. The verification procedures for this kind of quality control still needs to be defined upon input from the national experts and Data Manager.

However, some examples are:

- Missing mandatory (data) values (as in, e.g. mandatory load values)
- Other essential but non-mandatory information missing (e.g. coordinates)
- 0 <= Number of measurements below LOQ (LOD) <= total number of measurements
- Total number of measurements always >= 0 (zero only if value is not measured but estimated/calculated)
- Values are falling above/below predefined ranges, or "outliers" (ranges to be provided by national experts)
- Values reported with wrong units / parameters / parameter types combinations (e.g. flow reported as mg/l, loads as m3/s)
- checking on combination of values / respective LOQ (LOD), number of measurements and number of measurements below LOQ (LOD)
- illogical combination of methods, treatments, and/or other metadata (e.g. AVG value < MIN value or > MAX value)

Comment [PK6]: Should here be added the examples from Jytte's document

As a result of this step, the data will be flagged into the database (via the field QA_FLAG_ID) as "Questionable" (with different types of questionable data), or "Accepted". Of course, being an automated procedure, it is possible that some values could generate false positives (questionable data which is, in fact, correct) or false negatives (accepted data which is instead erroneous). For this reason, further steps are provided for human checking. These are performed in QA level 3 and 4.

Requirement Id# 31 – Priority - Medium

Able to receive a report warning about possible suspicious data as a result of the automated quality control

As a result of the import process, similarly to the process described for missing data (Requirement 29), warnings concerning suspicious data will be displayed to the user in order to be able to perform already a preliminary verification.

These suspicious data can be retrieved from the database separately from the other data, according to the mechanism specified in Requirement 30. Identification of data as suspicious is possible through the use of the QA_QUESTIONABLE_FLAG_ID field, in combination with the QA_FLAG_ID (when the latter is set as "QU" (= questionable). QA_QUESTIONABLE_FLAG_ID should be NULL, when QA_FLAG_ID indicates a flag other than "QU".

Tables containing such quality control are all those tables storing load, flow and concentration values (both annual and periodic, including natural background, retention, and source apportionment), as well as TBL_SUBCATCHMENT, TBL_STATION, and TBL_POINT_SOURCE.

Requirement Id# 33 – Priority - Medium

New PLC Database to generate a report about which data have been rejected during the submission, and why (type of error)

Comment [MM7]: This could be eventually saved e.g. in a txt file to be downloaded, if needed?

As for Requirements 29 and 31, errors preventing data from being inserted into the database are returned to the users as a result of the data submission procedure. These errors will be categorized according to the reason why they occurred, in a human-readable form, and will have an associated code, as well as indication of what data has generated the error.Eventually such kind of report resulting from the data submission (listing possible errors and warnings), could be downloaded for later use.

Requirement Id# 34 – Priority - Medium

Able to flag data as suspicious when necessary

The data in the database comes already flagged as "questionable" or "accepted" as a result of the quality controls performed during data import. In case after the automated QA process, the data needs further verification from the national experts (reporters and quality assurers), these have the possibility to verify manually their (national) data through the web application, and where necessary to modify the quality flag of one or more values to questionable, selecting a specific reason why, and optionally adding a descriptive comment about the possible issue.

This is achieved by changing the QA_FLAG_ID field to "QU", and selecting a category of possible errors from the QA_QUESTIONABLE_CATEGORY / QA_QUESTIONABLE_FLAG_ID combination. The reason can be also described in textual form in the field QA_NOTE_TEXT.

Requirement Id# 36 – Priority - Medium

Possibility to specify the criteria according to which the data have been marked as suspicious

Users are able to provide additional comments on the questionable data by creating a QA_NOTE linked to the data. This link will be identified by the QA_NOTE_ID field.

This "note" is where the user can specify quality-related information on the data, and in case it is deemed suspicious, possibly specify further the reason why, by associating the note with a particular kind of error, via the QA_QUESTIONABLE_FLAG_ID. The latter refers to the QA_QUESTIONABLE_FLAG table, listing the categories of possible errors on the values (e.g. logical rules, scientific rules, outliers detection, data definition, station rules…), and the specific kind of error related to the data.

< elaborate possibly a bit further…>

Requirement Id# 41 – Priority - Medium

Able to fill in possible data gaps/missing data

Missing mandatory data (identified by DATA_SOURCE_FLAG_ID = "MI") should be replaced (when possible) either with measured data, with estimates coming from the national experts, or, in the worst case (i.e. if no such data can be provided Country-wise), with estimates suggested by the LOAD CORE Group. Ultimately the responsibility of filling these "data gaps" is left to the national experts, or under special circumstances to the Data Manager.

Once the data is filled-in, they lose the flag "MI", and acquire one of the other possible flags:
ME = measured
CA = calculated
ES = national estimate
LE = LOAD CORE group estimate

The selection of the correct one among the above flags is responsibility of the person inserting the value(s).

In cases where no real data nor a reliable estimate can be provided at all, to avoid that the missing data is repeatedly requested, the DATA_SOURCE_FLAG_ID can be also reset to "NULL", indication of data which is "Not Available" (N/A).

Requirement Id# 42 – Priority - Medium

Able to mark these estimated data as estimated values filling in data gaps in the database

Data which has been estimated as a result of filling the data gaps can be categorized in two different ways:
ES = estimates provided from sources within the Country responsible for reporting the data
LE = estimates provided from the LOAD CORE Group experts, which are suggested to the national experts, and can be inserted into the database by the national reporters or quality assurers only after being accepted on a Country-wise basis.

National experts who have the rights to insert or update their national data, will be able to fill in these data gaps via the web application developed for the new PLC database.

For further details about filling data gaps, see Requirement 41.

Requirement Id# 44– Priority – High

Contracting Party should be able to approve estimated data added by LOAD core group to fill eventual data gaps

Entering the data suggested by the LOAD CORE group in order to fill-in data gaps is the responsibility of the Contracting Party. These data will then follow the QA process described in Requirement 30, and as such can be flagged as "Approved" once they are deemed correct. This operation can be done by National Data Reporters, National Quality Assurers, or in exceptional cases by the Data Manager.

It is important to note that in terms of approval, what is approved by the data manager cannot be edited without his consent (or without "unlocking" these data), and similarly what has been approved by the QA personnel cannot be edited by the Data reporter. In practice:

Data Manager > National Quality Assurer > National Data Reporter

Requirement Id# 47– Priority – High

Able to verify a newly uploaded dataset, before it is made officially available in the database

Any new dataset will go through the QA process, before it is officially available in the database. If data is entered to fill-in data gaps, then these values will automatically move to QA level 2, as they have been already verified during the previous step when inserted into the database.
The National Data Reporters will be able to verify the data uploaded as a part of the QALevel 3 and flag it for further approval to the National Quality Assurers. The final approval is done in QA level 4 by the National Quality Assurers.

Requirement Id# 48– Priority – High

Able to approve the reported data before it is made officially (publicly) available in the database

All newly reported data should be approved as a part of the QA process, before these are made available in the database. For more details, see Requirement Id #47.

Requirement Id# 49– Priority – Medium

New PLC Database to allow retrieval of individually reported point sources including their coordinates

General information regarding each point source is available in the table TBL_POINT_SOURCE. In addition, data specific to the type of source (Industry, WWTP, Fish Farm) is available through the tables TBL_INDUSTRY, TBL_MWWTP, TBL_FISH_FARM, respectively. For each point source, the coordinates are expressed by the fields:

PS_LAT (Latitude)

PS_LON (Longitude)

Point sources which are reported individually, can be retrieved with the respective coordinates (these should be reported as much as possible by the Contracting Parties, and ideally it should be mandatory or at least common practice to report point sources and stations, including their coordinates). Without coordinates it won't be possible to identify the location of a point source (or better, of its outlet). The only case where it is considered acceptable to avoid reporting the coordinates, is when the point sources are reported as aggregated.

Requirement Id# 51– Priority – High

To be able to download aggregated data (e.g. by Baltic Sea sub-basin per Contracting Party) based on approved data (including estimated data)

From the database it will be possible to browse for Country-wise data and data divided per Sub-basin, or a combination of these. Some examples are:

- Total nitrogen load from Poland;
- Heavy metal loads in Gulf of Finland (as a whole, or divided per Country);
- Swedish loads into the Bothnian Sea sub-basin;

These data are available through the tables collecting load, flows, and concentration values, and the structure of the database (division in sub-basins per country, sub-catchments, etc.) allows for aggregating (compiling) the data by querying the database to request the desired information per Contracting Party, per sub-basin, or combinations of these.

Requirement Id# 54– Priority – High

New PLC Database to allow me to retrieve total loads apportioned by source for periodic assessments

The TBL_SOURCE_APPORTIONMENT contains Anthropogenic, Natural and Transboundary load categories. Each category is further divided to sources of
NL = Natural load
NBL = Natural background load

DL = Diffuse load
DIL = Diffuse load (if it cnnot be further specified).

Otherwise:
ATL = Atmospheric deposition
SCL = Scattered dwellings
AGL = Agriculture and managed forestry
SWL = Stormwater, over flow or urban area load

PL = Point source load
INL = Industrial load
MWL = Municipal waste water load
FIL = Fish farm load
OTL = Other point source load; and

TL = Transboundary load
TBL = Transboundary load

Sum of the loads of different sources should be equival with the reported total for every periodic reporting.

Requirement Id# 55– Priority – Medium

Missing sources from total load apportionment to be estimated

The TBL_SOURCE_APPORTIONMENT contains information on the total load apportioned from individual sources. Missing sources can be identified using the DATA_SOURCE_FLAG_ID field (when the field is set to “MI”).

For further details on the estimation of missing data, refer to Requirements 41, 42, 44.

Requirement Id# 56– Priority – Medium

New PLC Database to allow me to access data which I can use for modelling and research purposes

Open, read-only access to approved data (or with a certain level of quality, to be decided by Contracting Parties) is provided via web application to the PLC database underneath.

For more details, see Requirement Id #2.

Requirement Id# 57– Priority – Medium

To be able to access the validated data at all times as soon as they are approved in the database

Access to the data is provided to the national experts and Data Manager via a password-protected application, allowing them to perform different actions on the data, according to their role credentials and user rights.

The public, and all other end-users who don’t have restricted access, can access the data made available through the web application. These data need to have a minimum level of quality in order to be approved, and thus made openly available. The QA process ensures the possibility to the national reporters and QA assurers to approve the datasets (see Requirements 44, 47, 48). The availability of the approved data (and in specific occasions and with an appropriate disclaimer, of “lower quality” data if required) is guaranteed via the user interface developed in the PLUS web application, and as long as the system is kept operational and regular back-ups and maintenance are taken care of, these should be available at most times, except during

preannounced maintenance downtime or updates or in case of accident (such as e.g. power shortage).

Requirement Id# 58 – Priority – Medium

To be able to access metadata on sub-catchments and rivers (size of catchment, size of river, land use)

Subcatchment metadata is stored in the table TBL_SUBCATCHMENT and information linked to subcatchments such as links to a river catchment, station(s), point sources, etc. can be found in the related tables. Regarding the size of the subcatchment, TBL_SUBCATCHMENT provides the following information on an annual basis:

TOTAL_DRAINAGE_AREA: total surface area of the river catchment to which the subcatchment is linked to, or total size of the unmonitored area (per sub-basin) in km2 (the established monitored area size is specified in the TBL_STATION, as there is a link between stations and subcatchments via the table AGG_SUBCATCHMENT_STATION).
COUNTRY_DRAINAGE_AREA: in case the river is a border river, or a transboundary one, this field is used to indicate the total drainage area for one country. This helps in dividing the loads country-wise.
TRANSBOUNDARY_AREA: Complementing the field above, this indicates the total drainage area which falls instead outside of the country borders. If present, the sum of COUNTRY_DRAINAGE_AREA + TRANSBOUNDARY_AREA should be equal to TOTAL_DRAINAGE_AREA.

Requirement Id# 59 – Priority – Medium

To be able to access metadata on stations

The following metadata regarding stations is stored in the TBL_STATION:

STATION_NAME: Name of a station (where available)
NATIONAL_STATION_CODE: National code for a station (where available)
STATION_TYPE: whether a station is chemical (measuring loads and concentrations), hydrological (measuring flow), or a combination of both (hydrochemical)
ST_LAT: Station latitude in decimal degrees
ST_LON: Station longitude in decimal degrees
MONITORED_AREA :Total size of the monitored part of a monitored subcatchment controlled by a chemical station in km2, the station should be the closest to the river mouth
IS_ACTIVE : Indicates whether a monitoring station is in use or not in relation to monitoring during a given time period (specified through the field PERIOD_ID)
IS_WFD :- Indicates if the station is also utilized under Water Framework Directive (WFD) obligations.
WFD_CODE: If the station is reported under WFD obligations

Further additional information or exceptions can be reported for each station/period combination under the field REMARKS.

Requirement Id# 60 – Priority – Future

Access metadata on number of measurements

The PLC database provides access to metadata information on the number of measurements, when this is reported by the Contracting Parties. This information is available from the NR_MEASUREMENTS field of the following tables

VAL_SUBCATCHMENT_LOAD
VAL_STATION_FLOW_CONCENTRATION
VAL_INDUSTRIAL_FLOW_LOAD
VAL_FISH_FARM_LOAD
VAL_MUNICIPAL_FLOW_LOAD

Requirement Id# 61 – Priority – Low

To be able to access metadata on background data of point sources (treatment method, number PE, number of people connected to treatment plant)

The metadata information for point sources is available the TBL_POINT_SOURCE on an annual basis. This includes the number of PE, as well as the number of people connected to treatment plant.
The treatment method, in case waste water has been treated before being released from the plant's outlet, is available in the TREATMENT_METHOD field of the VAL_MUNICIPAL_FLOW_LOAD table, when the TREATMENT_STATUS is different from "UNTREATED" or "NULL".

Comment [MM8]: Clarifications needed from the Contracting Parties.

Requirement Id# 64 – Priority – Medium

New PLC Database to be able to produce graphics (stack bars, pie charts, lines, etc.) and maps (containing stations, point sources, graphs, etc.), e.g. for HELCOM assessments and reports

The Database contains the information and data necessary for the generation of the reports (in form of tables, charts, graphs etc. when needed). The development of graphic visualization will happen as a part the Web Application development in Work Package 5.
Graphics related to maps will be handled in WP6 (Release 2 of PLUS)

Requirement Id# 65 – Priority – Medium

New PLC Database to allow uploading aggregated data once a year, and to use these data as a basis for producing graphs, tables and maps

Data can be reported annually as individual or as aggregated (small rivers or small point sources).
The data can be uploaded using the web application and further used for producing graphs. The development of these functionalities will happen in WP4and WP5, respectively.
The integration of PLC data with maps will happen in Release 2 of the PLUS project.

Requirement Id# 66 – Priority – Medium

New PLC Database to easily generate data products (e.g. tables, figures and maps) on total riverine and direct loads into the sea by country (including monitored and estimated inputs)

The information regarding total riverine and direct loads into the sea by country is obtained from the database by compiling the total loads by country and/or subbasin (or even by subcatchment) of
val subcatchment loads of monitored rivers and unmonitored areas;

and direct loads of:
val_municipal_load_flow
val Indutrial_load_flow
val_fish_farm_load

Graphical presentation can be defined as bar graphs or lines and as time series.

Requirement Id# 67 – Priority – Medium

New PLC Database to easily generate data products (e.g. tables, figures and maps) on total loads to different sub-basins

Total loads into the sea, from different subbasins, are available in the following tables:

VAL_SUBCATCHMENT_LOAD: for the data related to unmonitored subcatchments, coastal areas, and unmonitored parts of monitored rivers
VAL_STATION_FLOW_CONCENTRATION: data from monitored rivers
VAL_MUNICIPAL_FLOW_LOAD, VAL_INDUSTRIAL_FLOW_LOAD, VAL_FISH_FARM_LOAD: point source data
TBL_DIFFUSE_SOURCE: diffuse losses (agriculture, managed forestry, scattered dwellings, etc.)
TBL_NATURAL_BACKGROUND: natural background losses
TBL_RETENTION: data indicating nitrogen and phosphorus retention to be subtracted to the loads in order to obtain more correct figures on the totals loads to the sea.

Requirement Id# 68 – Priority – Medium

New PLC Database to easily generate data products (e.g. tables, figures and maps) on total loads for individual point sources and rivers

Total load of different parameters for individual point sources and rivers is obtained from the tables listed under Requirement 67.
For a list of load parameters, please refer to the PLC Guidelines.

Requirement Id# 69 – Priority – Medium

New PLC Database to easily generate data products (e.g. tables, figures and maps) on aggregated data on pollution loads by country, sub-basin, sub-catchment, point sources, source apportionment

See Requirement 67.

.

Requirement Id# 70 – Priority – Medium

New PLC Database to easily generate data products (e.g. tables, figures and maps) on trends in loads (rivers, point sources, sub-basin, country)

The web application will provide tables and figures to indicate the trends in load for rivers, point sources, sub-basins etc. See Requirement 67 for the tables listing such information.

Requirement Id# 71 – Priority – High

The new PLC Database to easily evaluate the progress of each country towards the BSAP (Baltic Sea Action Plan) nutrient reduction targets

Pollution loads of nitrogen (N) and phosphorus (P) from each Country, divided per sub-basin, are stored in the database. These can be compared with the reduction targets defined for each Country in the Baltic Sea Action Plan (BSAP)

The reduction targets are not stored in the database.

Requirement Id# 72 – Priority – High

The new PLC Database to include a disclaimer in all figures, graphs and maps where suspicious or estimated data has been used, and/or where data is missing, indicating that the results could differ by those obtained with more up-to-date data extracted from the database directly

The database provides the possibility to store missing, questionable and estimated data. Disclaimers can be added to the reports, maps, graphs and figures, when the implementation of web application is carried out in WP5.

Requirement Id# 74 – Priority – Low

The new PLC Database to include the date and copyright (HELCOM PLUS database) in all data products

Copyright information will be included as a part of the application during the implementation of WP5.

Requirement Id# 80 – Priority – High

The new PLC database user interface to have the following characteristics:

a) easy to use
b) intuitive and user-friendly
c) secure
d) available at all times
e) give quick access to the most useful functionalities (specific to my needs)
f) customizable according to my needs (e.g. add custom search link
g) easy to download big amounts of data
h) other, please specify:

The user interface will be implemented in WP3, WP4 and WP5.
A detailed User Functional Specification document, containing screen shots for all the functionalities linked to the requirements above will be provided at a later stage.

List of Requirements to be evaluated for Future Releases

Priority
F – Future
L – Low
M – Medium
H- High

Business Requirement ID	Short Description	Priority (F,L,M,H)
6	Retrieval of data on trans-boundary loads separately from national total inputs to the sea (after retention)	**M**
25	Include an automated notification system to inform Contracting Parties about the beginning of the reporting period	**M**
27	Retrieve a report about which data has been reported, and when	**M**
40	Retrieve a report about the current status delivery of submitted data, including data gaps, suspicious data, etc. (e.g. status could be: submitted, quality checked, approved)	**M**
71	View and download the PLC data from the HELCOM map and data service (http://www.helcom.fi/GIS/en_GB/HelcomGIS/), with the possibility of overlaying it with other datasets (e.g. land use, livestock, rivers, etc.)	**M**
78	Contribute to the implementation of EU INSPIRE directive (website: http://inspire.jrc.ec.europa.eu/)	**M**
35	Possibility to notify the Contracting Parties about flagging of data as suspicious	**F**
37	Include an automated reminder system to inform Contracting Parties about suspicious, or not yet quality controlled data	**F**
39	Generate a report about all data marked as suspicious (selecting them e.g. by country, year, sub-basin, stations, etc.)	**F**

45	Check eventual discrepancies between reported data and expected data (e.g. sum of monitored, aggregated and estimated data)	F
50	Enable data aggregation at bigger sub-catchment levels	F
75	Provide better access to the HELCOM PLC data for pan-European assessments	F
76	Provide better access to the HELCOM PLC data for global level assessments	F
88	Ability to save normalized data used in different assessments (including the history of the calculations)Ability to save normalized data used in different assessments (including the history of the calculations). ***Only to possible to store data and not history of calculations***	FF
46	Generate a map indicating the status of reporting in different catchment areas and be able to download this map	F

Comment [MM9]: Clarify what it means

Comment [MM10]: Clarify what it means

Formatted: Font: Bold, Italic, Font color: Red

List of Requirements proposed to be rejected

Business Requirement ID	Short Description	Priority (F,L,M,H)	Comments
52	Trace back to the single sources of aggregated data (e.g. single point sources, stations, etc.)	M	**This information is not stored in the database**
77	Enable displaying of the data on the EEA data and/or map portal (e.g. EU WISE MARINE system - https://webgate.ec.europa.eu/maritimeforum/category/554), other regional marine conventions' portals or other international organizations' (UN) portals, by using a suitable technology	M	**This might require a major study and integration options will need to be evaluated**
79	The new PLC Database (and its interface) to allow for the development of new types of graphs and maps in the future	M	**It should be possible to generate reports and graphs, as long as the data is available. Requirement is too generic.**
14	Submit the data using own national system for coding stations and point sources (the conversion is handled by the database)	L	**Requires all the countries to provide national codes and PLC codes**
16	Modify the information about whether a point source currently belongs to a monitored or unmonitored area	L	**Agreed with PLC group that itcannot be done automatically at this stage.**

Formatted: English (U.S.)

20	Report point sources data only individually and retrieve aggregated data from the database, at different aggregation levels	**L**	**This is not possible from the Contracting Parties**
17	Enter start and end date for a station	**F**	**Not relevant as they are included in the period**
38	Maintain a history of which data have been previously considered suspicious	**F**	**Excessive historical information**
43	Estimates including estimates of data gaps/missing data to be kept separately from officially reported data	**F**	
62	Access metadata on links to information sources (e.g. UBC WWTP database, E-PRTR database)	**F**	**Information is currently not provided**
63	Access other kind of metadata, please specify: Population density, WWTP performance, sludge energy production, energy consumption, benchmarking, eutrophication, all relevant data necessary for PLC assessments (e.g. fertilizer consumption, livestock, precipitation, temperature, rainfall); Note: information such as estimation methods, land use, soil type, are currently included into the existent database (when available).	**F**	**Information is currently not provided**

Formatted: English (U.S.)

Formatted: English (U.S.)

4.2.User Management

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 Data Manager is the user with the highest privileges (administrative). He shall be responsible for creating and managing the other user profiles.

Following are the other list of other user profiles available in the database

- National Expert
- Load Core Group
- National Quality Assurer
- National Data Reporter

The TBL_USR shall contain the following details regading the user

- 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

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 3.2.2

The Historical activities of the user related to the correction of data is stored in the Table TBL_HISTORY_LOAD. The User Id is used to identify a user's activities in the historical table.

In order to support the Quality Assurance of data, the users shall be able to enter notes regarding the Questionable data . The user entering the note on the Questionable data is identified by his unique User Id . The Table QA_NOTE contains the information regarding the note on the questionable data.
For more details, please refer to the Data model in Section 3.2.2

The Country of the user is helpful in identifying which data a user (National data reporter) would be authorized to modify.
The Users shall be able to exchange messages in order to support the Quality Assurance process. This information is stored in message table (TBL_MESSAGE). The user id's of the sender and receiver are available in the table.
For details, please refer to the data model in Section 3.2.2.

User Profile	Tables	Privelages (C-Create, M-Modify, D- Delete,R-Read)	Comments
Data Manager	All	CRUD	Administrative Privelages. Can create or Modify all tables. Is responsible for creating other users.
National Quality Assurer			Create , Read and Update access to country specific Data and Read only access for Non country specific data. Has the highest level of QA access . Able to modify the QA status of the data entered by the National Data Reporter . Final authority providing approval of country specific data
National Data Reporter			Create Read & Update rights to the data. Rights to Initial QA check .

			Cannot modify data if approved or if the Quality assurance responsibility is transferred to Quality Assurer.
Load Core Group			Read Access to all the data. Able to add comments (COMMENT column) to specific data as a part of Quality assurance.