Washington State
Department of Ecology

Exchange Network Node Enhancement Project

Data Flow Design

Water Quality Data Exchange (WQX)

Version 0.6
Draft

August 20, 2008

WINDSOR
SOLUTIONS, INC.
Environmental + Health
Information Systems

Version Control

Date	Author	Changes	Version
8/20/2008	Windsor	Updated to clarify and further detail the proposed approach.	0.6
7/15/2008	Windsor	Draft version	0.5

Contents

INTRODUCTION.......... 1
DATA EXTRACTION.......... 2
Overview.......... 2
Insert/Update Management.......... 2
Delete Management.......... 6
Data Extract Approach.......... 6
DATA SERVICES.......... 8
Overview.......... 8
WQXGetInsertUpdateSubmission.......... 8
WQXGetDeleteSubmission.......... 9
Submission History Table.......... 9
DATA SUBMISSION.......... 11
Overview.......... 11
Flow Prerequisites.......... 11
Header File Creation.......... 12
Submission Processing, Feedback, and Error Correction.......... 13

THIS PAGE INTENTIONALLY LEFT BLANK

Introduction

The Washington Department of Ecology (Ecology) was the first Exchange Network partner to implement a fully functional production Network Node. Following this implementation in 2003, Ecology has continued to lead the implementation of the Exchange Network by implementing a significant number of data flows.

Ecology is currently engaged in a project to upgrade the existing Network Node and data flows to the latest version, as well as to develop several new data flows. The objective of this upgrade is to allow Ecology to take advantage of important functional and technical improvements to the current Windsor Node, as well as to meet the requirements of the latest version of the specifications for operation of Exchange Network Nodes, version 2.0.

The purpose of this document is to describe the design and implementation of a new data flow to be supported by the upgraded Ecology Node to enable data submissions to the EPA Water Quality Exchange (WQX). At this time it is expected that this data flow will only support submissions to the EPA WQX system. No data publishing services will be provided at this time.

It should be noted that the WQX data flow will initially be implemented using the new Ecology Node's 1.1 endpoint. This will be changed in the future to operate using the Ecology Node 2.0 endpoint once the EPA CDX Node has been upgraded to support this.

The remainder of this document is organized as follows:

Data Extraction details the process by which data will be extracted from the source database and made available in a staging database for submission to EPA.

Data Services describes the data services that will be provided on the Ecology Node to establish the data flow to WQX..

Data Submission specifies the individual workflow steps involved in submitting data to the EPA WQX system and validating the results.

Data Extraction

Overview

The source data for submissions to WQX is managed by Ecology in the Environmental Information Management (EIM) system. This system manages all of the information required by the WQX data flow and XML schema.

To support the WQX data flow to EPA, selected data will be extracted from the EIM database and will be copied to a staging environment to be used by the Ecology Node to generate XML documents for submission to the EPA CDX Node. During this extraction, the source data will be reformatted to more closely align with the data structure required by the WQX schema. In general terms, the WQX data structure is consistent with the EIM data structure although reference values used throughout the WQX schema, for example, to indicate the analyzed characteristic, are different to those used by EIM so some data translation will be required. Once available in the staging environment, the data will be available to the Ecology Node for submission processing.

The Flow Configuration Document (FCD) for the WQX data flow requires that organizations first, submit data for insert/update into the WQX database, and second, submit data for delete from the WQX database. Two separate XML schemas are available for these purposes

Insert/Update Management

The generalized structure of the WQX 2.0 XML schema is represented in the figure on the following page. It should be noted that this latest version of the XML schema has only recently been finalized and it includes a number of significant additions to the prior version, chiefly related to the inclusion of support for information related to habitat and tissue analysis.

The schema is built around six key logical components: “Organization”, “Project”, “Monitoring Location”, “Activity”, “Activity Group”, and “Biological Habitat Index”. Each of these primary components includes various data elements and additional subordinate complex types. Only a single Organization may provided in each XML document.

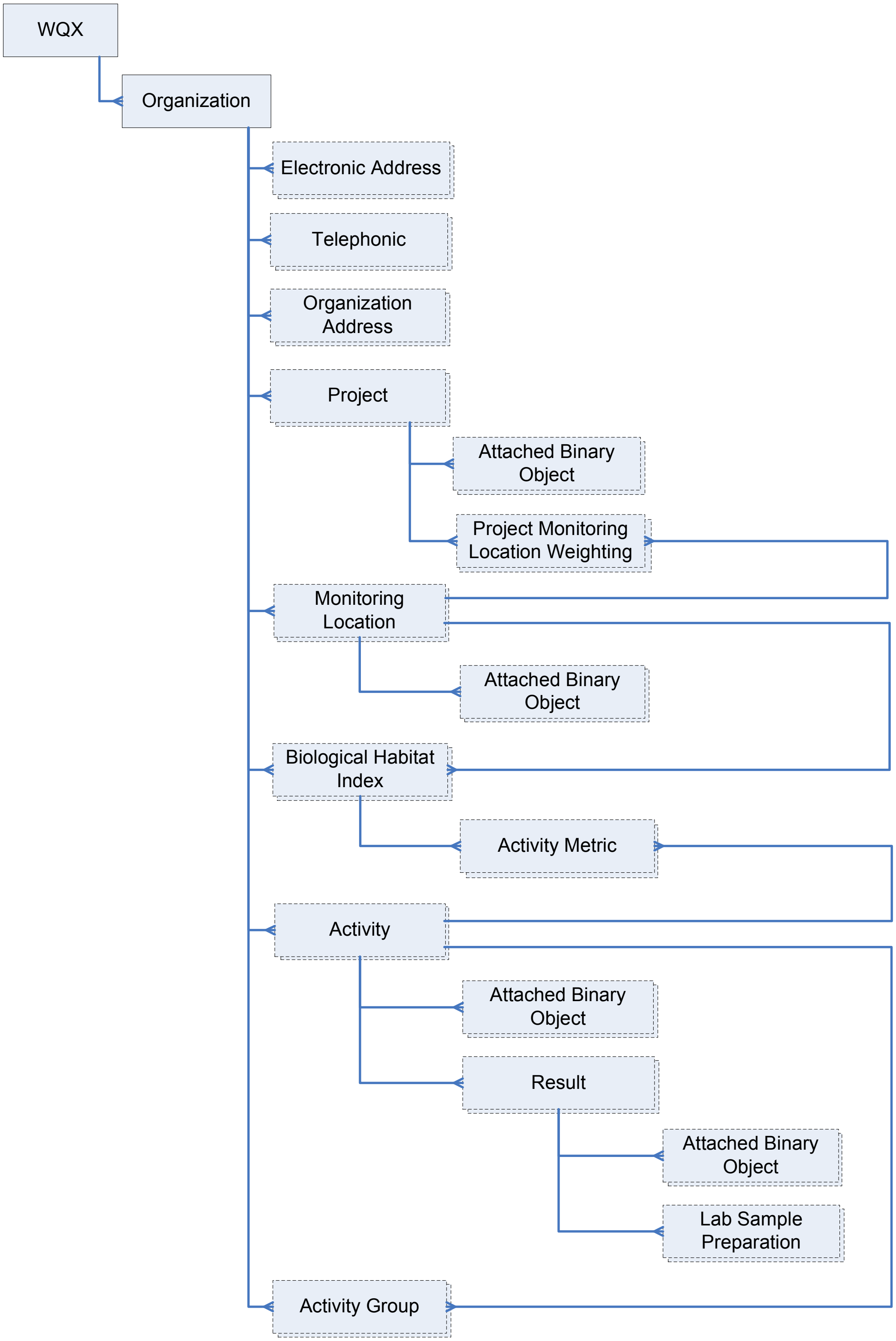
A detailed description of the processing rules for inserts and updates of records to the WQX database is provided in the FCD, but at a high level, other than the Organization components, the process uses the data provided at the level of each of the other components to determine, based on unique identifiers provided by the submitting organization, whether to insert new records if they do not already exist, or update existing records if they do already exist. In each case, subordinate data to these five primary components is replaced in each case.

Ecology has already begun the process of mapping complex types and data elements from the source EIM database to the WQX XML schema. Ecology will use this mapping to develop a stored procedure that will extract selected data from EIM, transform reference values and other elements as necessary, and load the resulting data into a set of staging tables that will be structured to mirror the structure of the XML schema.

All existing data will first be purged from the staging tables before being reloaded from the source EIM database tables. A three year history of information will be included in the staging database. This approach will:

- Enable support for future data publishing services if required by EPA in the FCD
- Allow Node to manage submissions and resubmissions in the case of errors.

- Standardize the Node plug in to enable future reuse.
- Keeps source data extracts as straightforward as possible.



WQX 2.0 XML Schema Components

As noted above, the data structures broadly correspond between EIM and WQX. The following general processing rules will be implemented by the stored procedure to populate the staging tables:

Organization

WQX requires that each data submission include information for only one organization. At this time, it is expected that Ecology will only be submitting data for one organization, itself, so only a single organization record will be established in the staging environment. This information will be created and updated manually and the extraction stored procedure will not need to modify this table.

Project

- This schema component generally maps to the EIM Project table and related subordinate tables.
- A new date field will be added to the EIM Project table, “WQX_Updated_Date”. The purpose of this field will be to track the date on which the Project record and any ancillary data was updated within the EIM database, for the specific purpose of identifying changes to be submitted to EPA WQX. This same field will also be included in the relevant staging database table.
- Database triggers will be established on the Project table in the EIM database and any related tables which store data elements that are mapped to the WQX XML schema. These triggers will set this new date field to the current date whenever a change is made to a record in one of those tables. This will allow relevant modifications to the Project records to be identified during later submission processing steps. Insert and update triggers will be established on the EIM Project table itself, and insert, update and delete triggers on the following related tables:
    - Study Area
    - Agency Role
- Project data will be purged and reloaded in the staging environment at each execution.
- A subset of Project records will be extracted for submission to WQX. Not all EIM Project records will be submitted. A new field will be added to the Project table to flag whether data from the entire project should be excluded from WQX upload. This will be used for data submitted to EIM from other “Providing Organizations” which may have an independent feed to WQX.
- Relevant reference code values will be translated to WQX equivalent values using alias values that will be established and maintained by program users in the EIM database.
- Data types from the EIM tables will be converted to the appropriate simple data types in the staging tables.
- While the WQX XML schema supports the inclusion of binary objects along with project information, binary objects will not be extracted or provided at this time.

Monitoring Location

- This schema component generally maps to the EIM Station table and related subordinate tables.
- A new date field will be added to the EIM Station table, “WQX_Updated_Date”, and this field will also be included in the relevant staging database table.
- Database triggers will be established on the Station table in the EIM database and any related tables which store data elements that are mapped to the WQX XML schema. These triggers will set this new date field to the current date whenever a change is made to a record in one of those tables. This will allow relevant modifications to the Station records to be identified during later submission

processing steps. Insert and update triggers will be established on the EIM Station table itself, and insert, update and delete triggers on the following related tables:

- Geographic_Loc
- Well_Station
- Well_Interval
- Station_Alias
- Project_Station

– Station data will be fully refreshed in the staging environment at each execution.

– Station records will only be extracted where they are associated with a Project that is selected for submission to WQX. Not all EIM Station records will be submitted.

– Relevant reference code values will be translated to WQX equivalent values using alias values that will be established and maintained by program users in the EIM database.

– Data types from the EIM tables will be converted to the appropriate simple data types in the staging tables.

– While the WQX XML schema supports the inclusion of binary objects along with monitoring location information, binary objects will not be extracted or provided at this time.

Activity

– This schema component generally maps to the EIM Field Activity table and related subordinate tables.

– A new date field will be added to the EIM Field Activity table, “WQX_Updated_Date”, and this field will also be included in the relevant staging database table.

– Database triggers will be established on the Field Activity table in the EIM database and any related tables which store data elements that are mapped to the WQX XML schema. These triggers will set this new date field to the current date whenever a change is made to a record in one of those tables. This will allow relevant modifications to the Field Activity records to be identified during later submission processing steps. Insert and update triggers will be established on the EIM Field Activity table itself, and insert, update and delete triggers on the following related tables:

- Sample
- Laboratory
- Result
- Sample_Composite

– Field Activity data will be fully refreshed in the staging environment at each execution.

– All Result information related to the Activity will be extracted and refreshed in the staging environment.

– Field Activity records will only be extracted where they are associated with a Project that is selected for submission to WQX. Not all EIM Field Activity records will be submitted.

– Relevant reference code values will be translated to WQX equivalent values using alias values that will be established and maintained by program users in the EIM database.

- Data types from the EIM tables will be converted to the appropriate simple data types in the staging tables.
- While the WQX XML schema supports the inclusion of binary objects along with activity information, binary objects will not be extracted or provided at this time.

Activity Group

- The WQX Activity Group logical component broadly corresponds to the legacy STORET "trip" and "station visit" constructs. A possible equivalent in the EIM database is the Field_Acty_Group which is an optional grouping of Field Activities. However, EIM does not capture the data elements indicated in the WQX schema, and only has a reason description.
- No data will be created for this logical component in the staging tables.

Biological Habitat Index

- No data will be created for this logical component in the staging tables.

Delete Management

A detailed description of the processing rules for deletes of records from the WQX database is provided in the FCD, but at a high level, the process simply requires that the submitting organization provide a list of unique identifiers for each of the key components detailed earlier. These records and any subordinate data will then be deleted from WQX based on matching the provided identifiers. For Ecology data submissions, only Project, Monitoring Location, and Activity components will be deleted.

Ecology will develop a stored procedure that will identify deleted records for each of these key components using existing audit table functionality for the EIM Project, Station, and Field Activity tables. The data extract stored procedure will then insert the EIM identifiers for the relevant records into a table in the staging environment. Again, a WQX_Updated_Date field will be added to this table to enable data submission management. This field will be set to reflect the date the relevant component was deleted from the EIM database.

Data Extract Approach

A set of new database tables will be established in the existing NODE_FLOW database to support the WQX data flow. These staging tables will conform to a structure that mirrors the structure of the XML schemas for insert/update and delete. Table and column names will propagate the terminology used by the WQX schema and only simple data types will be used to minimize XML serialization issues. Additional WQX_Updated_Date columns will be added to the relevant key component tables.

These staging tables will be populated by the new stored procedure described above using the following data extraction, transformation and load process:

1. The stored procedure will be executed on a weekly basis, with the schedule established to minimize performance impacts to EIM users. The schedule will also be coordinated with the schedule established for the data submission process discussed below.
2. Data will be extracted from the EIM production database[1].

[1] The EIM Reporting database cannot be used since it does not include the necessary audit tables and WQX_Update_Date fields that will be required for transaction management.

3. The stored procedure will first purge all existing data in the staging environment. Data will be fully refreshed at each execution[2].
4. Data will be extracted from the source database. Only three years worth of data will be made available in the staging environment. The stored procedure will select only records where the WQX_Updated_Date for the key EIM table (Project, Station, Field Activity) is more recent than the current date less three years. Including three years of historical EIM data in the staging database will allow the Node to support WQX data publishing requirements if required in the future by the FCD.
5. Data will be transformed to the target staging table structures following the data processing rules described in the previous section, and the detailed data element mappings prepared by Ecology.
6. Reference values will be translated to relevant WQX required values using the approach described in the previous section.
7. Data will be inserted into the staging database tables and will become available for submission to WQX according to the workflow described in the following section.

[2] This purge/reload approach will allow the data flow to support resubmissions in the event of delayed processing failures, as well as possible future data publishing services.

Data Services

Overview

Two new data service plugins will be implemented on the Ecology Node to support the WQX data submissions, one to manage insert/update processing and one to manage delete processing.

> **WQXGetInsertUpdateSubmission**
>
> This data service will retrieve new and updated project, monitoring location, activity, and biological habitat index information, as well as related data, from the staging environment for a specific data range. The resulting data will be serialized into the WQX XML 2.0 Schema format and submitted to EPA CDX.
>
> **WQXGetDeleteSubmission**
>
> This data service will retrieve identifying information for any project, monitoring location, activity, and biological habitat index records that have been deleted from the EIM database in order to synchronize these record deletions with the EPA WQX database.

Using these services it will be possible for Ecology to provide all information needed by the EPA WQX data flow. These data services will only be available internally to the Ecology Node and will not be made publicly available.

WQXGetInsertUpdateSubmission

This data service plugin will compare the WQX_Updated_Date field on the key staging tables to the date of the last data submission to determine what data should be extracted and composed into an XML document for submission to EPA. A “Submission History” table will be created in the staging database to track the various data submissions over time. Each time the data service executes, it will insert a new record into this table to reflect the type of submission, the date of execution, and the WQX_Updated_Date value that was used at that time. At each execution, it will use the prior WQX_Updated_Date value to compare against the WQX_Updated_Date on the key staging tables to determine what data to extract to the XML document.

The Submission History table will also track the EPA CDX Node transaction id that was received upon submitting the generated file, and the CDX Node processing status for the transaction. This status will be determined later in the processing flow.

The data service will accept the following input parameters:

Parameter	Required	Business Rules
1. OrganizationIdentifier	Required	Used to provide the identifying number by which Ecology is known as an organization in the EPA WQX database. Any one data submission to EPA must include data for only one Organization.

Parameter	**Required**	**Business Rules**
2. WQXUpdateDate	Optional	Allows the provision of a date from which new and updated records are to be selected from the staging database. In normal operation, the data service will use the Submission History table to determine the starting date for a given data extract, but this parameter is provided to allow the user to override the data in the Submission History table when errors are encountered.

WQXGetDeleteSubmission

This data service plugin will compare the WQX_Updated_Date field on the staging table containing the identifiers of deleted records to the date of the last data submission to determine what data should be extracted and composed into an XML document for submission to EPA. The same Submission History table will be used for this purpose.

The data service will accept the following input parameters:

Parameter	**Required**	**Business Rules**
1. OrganizationIdentifier	Required	Used to provide the identifying number by which Ecology is known as an organization in the EPA WQX database. Any one data submission to EPA must include data for only one Organization.
2. WQXUpdateDate	Optional	Allows the provision of a date from which deleted records are to be selected from the staging database. In normal operation, the data service will use the Submission History table to determine the starting date for a given data extract, but this parameter is provided to allow the user to override the data in the Submission History table when errors are encountered.

Submission History Table

As described above, the operation of the two data services will depend upon the information managed in a submission tracking table in the staging database. This “Submission History” table will include the following columns:

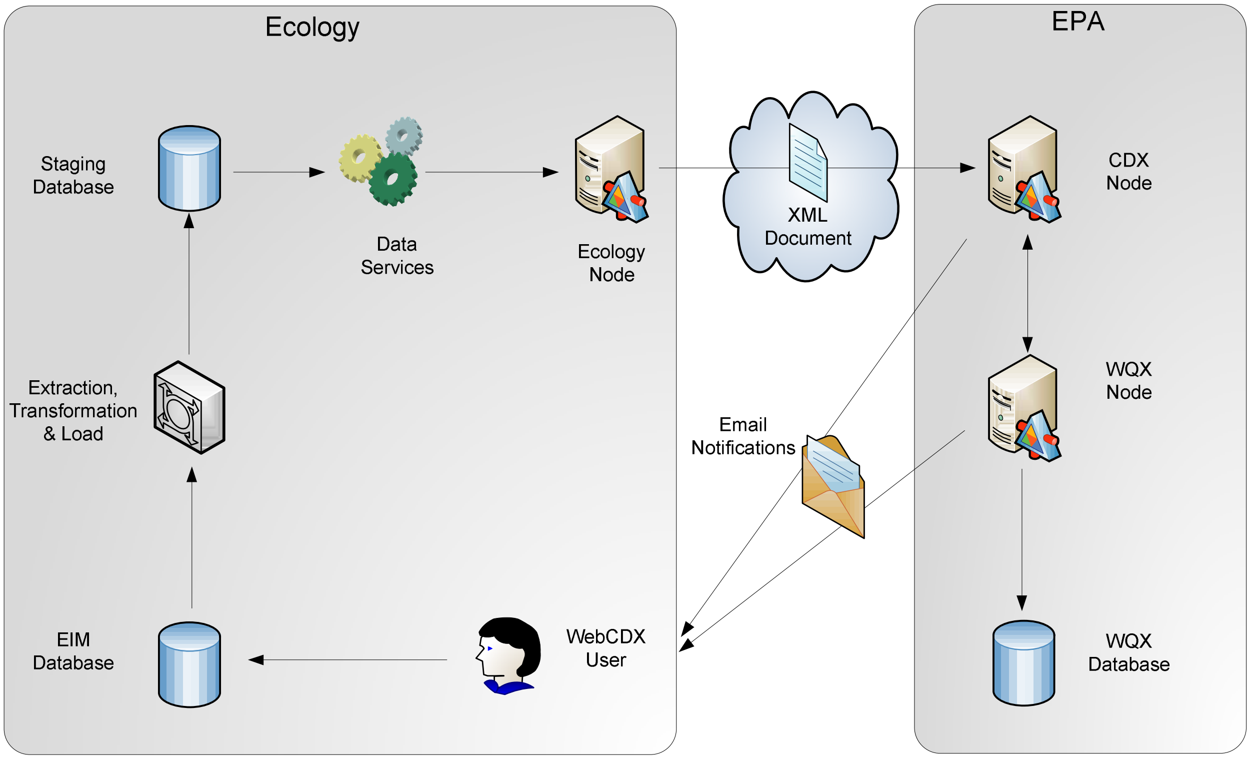
Column	**Data Type**	**Description**
Schedule_Run_Date	Date	The date on which the Node Schedule was executed.

Column	**Data Type**	**Description**
WQX_Update_Date	Date	The date from which records are to be selected from the staging environment for submission to WQX. Will equal the Schedule_Run_Date of the last successful submission.
Submission_Type	String	Indicates whether the submission was for new and updated records or for deleted records. Values are: “Insert/Update” and “Delete”
Local_Transaction_Id	String	The transaction id stored in the local Ecology Node Administration database.
CDX_Processing_Status	String	The status of processing of the data submission at the CDX Node and WQX system. Values are: “Pending”, “Failed”, and “Completed”

Data Submission

Overview

The following figure illustrates the data submission process.



Flow Prerequisites

There are two essential prerequisites for any organization to begin submitting data to WQX.

The first is that a unique Organization Identifier must have been created in WQX by the EPA WQX Administrator. This will be used by WQX to properly assign data submitted to WQX to the appropriate entity. If Ecology does not already have an Organization Identifier, the responsible program manager should request that this be created. Once the Organization Identifier is known, a record should be manually created in the relevant Organization table in the Ecology staging environment to provide Ecology’s details associated with this Organization Identifier. All data submissions to WQX will then include this information.

The second is that the NAAS account used by the Ecology Node to perform WQX submissions must be associated with a WebCDX user account that has permission to modify data for the Organization. If Ecology does not already have a WebCDX account with authority to update Ecology’s WQX data, the responsible program manager should request that this be created. Once the WebCDX user account has been established, the Ecology Node Administrator should request that the Ecology Node NAAS account be linked to this WebCDX account. It should be noted that the individual that owns the WebCDX user account that is linked to the Node NAAS account will be the recipient of all processing messages generated by WQX.

Header File Creation

The WQX FCD specifies that all submissions to WQX through the EPA CDX Node must use the Exchange Network Header Document structure.

The Header Document includes two sections: “Header” and “Payload”. The following values will be used to compose the Header File for Ecology’s WQX submissions:

Header

Element	Description	Ecology Value	Required
Author	First and Last Name of individual generating the XML document	Bill Kellum	Yes
Organization	Name of company, environmental agency or individual generating the XML document	Washington Department of Ecology	Yes
Title	Type of Submission	WQX	Yes
Creation Time	Date/Time when the document was generated	Derived by the plugin at runtime in a valid XML date format string	Yes
Comment	Free text description of the message contents.	<blank>	No
Data Service	Unused	<blank>	No
Contact Info	Name, mailing address, city, state, zip, telephone number, and email address of person who may be contacted with questions concerning the submission.	Bill Kellum Application Data Services Washington Department of Ecology <address> <phone> <email>	Yes
Notification	This element is used only by the CDX Node (internally).	<blank>	No
Sensitivity	Unused	<blank>	No
Property	Unused	<blank>	No

Payload

Element	Description	Ecology Value	Required
Operation (attribute)	This describes the operation to be performed on the payload. Multiple values are not allowed	“Update-Insert” or “Delete”	Yes

Element	Description	Ecology Value	Required
Schema Reference	WQX approved schema for submission	WQX_WQX_v2.0.xsd	Yes

Submission Processing, Feedback, and Error Correction

The following process will be employed.

Data Extraction

1) Data will be managed in the EIM database normally. As data is created, updated, or deleted from the database, triggers on specific tables will update the WQX_Updated_Date field on the Project, Station, and Field Activity tables to the date of the change.
2) On a weekly basis, a stored procedure will be executed to extract and transform selected data from the EIM database according to the business rules and mappings discussed earlier in this document. Existing data will be purged and then new data will be loaded into the WQX data flow staging tables in the NODE_FLOW database.

Insert/Update Data Submission

3) A schedule will be established in the Node Administration Tool to execute the *WQXGetInsertUpdateSubmission* data service plugin which will compose an XML document and will submit the results of the execution to the EPA CDX Node. The WQX Organization Identifier will be specified as an input parameter to support the data service execution. This schedule will be executed monthly.
4) The *WQXGetInsertUpdateSubmission* data service will first confirm that there are no current submission records in the Submission History table in the staging database which have a Submission_Type of "Insert/Update", and a CDX_Processing_Status of "Pending". If any such records do exist, the data service will terminate with a null return value.
5) The *WQXGetInsertUpdateSubmission* data service will retrieve the Schedule_Run_Date from the Submission History table in the staging database for the most recent submission record where the type of submission is "Insert/Update", and where the CDX_Processing_Status is "Processed".
6) The *WQXGetInsertUpdateSubmission* data service will then query the staging tables corresponding to the five key XML schema components to retrieve any records where the WQX_Updated_Date value is later than the Schedule_Run_Date value obtained from the Submission History table. Relevant subordinate records will also be retrieved.
7) The data service will create a new record in the Submission History table corresponding to the execution, setting the WQX_Updated_Date to the prior Schedule_Run_Date value, setting the Schedule_Run_Date for this new record to the current date, setting the type of submission to "Insert/Update", setting the Local_Transaction_Id to the local Ecology Node transaction id, and setting the CDX_Processing_Status to "Pending".
8) The data service will then serialize the retrieved data into an XML document conforming to the WQX XML schema.
9) The data service will then submit the resulting file to the specified partner endpoint which will typically be the EPA CDX Node.

10) A second schedule will be established in the Node Administration Tool to execute the "GetStatus" primitive method against the CDX Node using the recorded Ecology Node transaction id stored in the Submission History table to determine whether the submission has succeeded or failed. The CDX_Processing_Status value will be updated for the relevant submission record. This schedule will be established to execute on a daily basis. Note that no Node based notifications will be initiated by this schedule.

Delete Data Submission

11) A schedule will be established in the Node Administration Tool to execute the *WQXGetDeleteSubmission* data service plugin which will compose an XML document and will submit the results of the execution to the EPA CDX Node. The WQX Organization Identifier will be specified as an input parameter to support the data service execution. This schedule will be executed monthly immediately following the execution of the schedule to process insert/updates.

12) The *WQXGetInsertUpdateSubmission* data service will first confirm that there are no current submission records in the Submission History table in the staging database which have a Submission_Type of "Delete", and a CDX_Processing_Status of "Pending". If any such records do exist, the data service will terminate with a null return value.

13) The *WQXGetDeleteSubmission* data service will retrieve the Schedule_Run_Date from the Submission History table in the staging database for the most recent submission record where the type of submission is "Delete", and where the CDX_Processing_Status is "Processed".

14) The *WQXGetDeleteSubmission* data service will then query the deleted record staging table to retrieve any records where the WQX_Updated_Date value is later than the Schedule_Run_Date value obtained from the Submission History table.

15) The data service will create a new record in the Submission History table corresponding to the execution, setting the WQX_Updated_Date to the prior Schedule_Run_Date value, setting the Schedule_Run_Date for this new record to the current date, setting the type of submission to "Delete", setting the Local_Transaction_Id to the local Ecology Node transaction id, and setting the CDX_Processing_Status to "Pending".

16) The data service will then serialize the retrieved data into an XML document conforming to the WQX XML schema.

17) The data service will then submit the resulting file to the specified partner endpoint which will typically be the EPA CDX Node.

18) A second schedule will be established in the Node Administration Tool to execute the "GetStatus" primitive method against the CDX Node using the recorded transaction id stored in the Submission History table to determine whether the submission has succeeded or failed. The CDX_Processing_Status value will be updated for the relevant submission record. This schedule will be established to execute on a daily basis. Note that no Node based notifications will be initiated by this schedule.

Feedback and Error Handling

19) When either an insert/update or delete submission file is received by the EPA CDX Node, the XML document is validated against the XML schema. If errors are encountered, the processing status is set to "Failed" and a notification email is sent to the email address of the WebCDX user account associated with the NAAS Account used by the submitting Node. If no errors are encountered, the processing status is set to "Pending" and the file is submitted to the WQX Node.

20) The WQX Node then processes the received file into the WQX database. If errors are encountered the WQX Node notifies the CDX Node which updates the processing status from "Pending" to "Failed". If no errors are encountered, the WQX Node notifies the CDX Node which updates the processing status from "Pending" to "Completed". The WQX Node provides the CDX Node with a processing report which describes the result of the processing activity, including any errors encountered. The CDX Node stores this report against the original CDX transaction id in its administration database.

21) Whether errors are encountered or not, WebCDX is configured to send a notification email to the email address of the WebCDX user account associated with the NAAS Account used by the submitting Node notifying them of the final status of the submission.

22) Upon receiving a notification email indicating a failed submission, the responsible user will manually download the processing report from the CDX Node using a NodeClient tool and specifying the CDX transaction id provided in the notification email.

23) Steps will then be taken to manually resolve any errors identified in the processing report by modifying data in the EIM database or other actions.

24) Once errors have been resolved, the data submission process will be manually initiated to resubmit the data in error. This will be accomplished using the Node Administration Tool to request that the Schedule be run one-time and immediately. For insert/update processing, this will result in steps 4) through 10) being executed, and for delete processing, steps 11) through 18).