

The *Oncor* Geodatabase for the Columbia Estuary Ecosystem Restoration Program: Handbook of Data Reduction Procedures, Workbooks, and Exchange Templates

NK Sather
AB Borde
HL Diefenderfer
JA Serkowski
AM Coleman
GE Johnson

May 2014

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Appendix C

Data Standardization

Appendix C

Data Standardization

Oncor will store data generated by multiple people from multiple agencies, each likely with its own set of data standards. To ensure proper integration of all data in *Oncor*, the maintenance of a single set of data standards is critical. Data quality will be ensured through the use of data and formatting standards.

A.1 Data Standards

Data standards help ensure that data from disparate sources have consistent meaning and can be properly compared. For *Oncor*, data standards are defined as the set of rules that applies to the contents of fields and records stored in a database. Standards include field type specifications (e.g., integer, date/time, Boolean, etc.), content constraints (e.g., times are local, etc.), and consistency in nomenclature.

Oncor implements data standards in several ways. At the lowest level, certain standards are controlled by the definitions of the table structures in the database. These definitions physically prevent certain improper data from entering the database. Every field in the table structure has a specific data type to which new values must adhere. For example, numeric measurement values are always placed in a field with a specific format—double precision floating point number—even if reported as an integer. Dates use must be in date/time format. Thus, string values in a numeric field or invalid dates in a date field would be automatically rejected by the database. Database tables may also have fields defined as *required*, so records cannot be saved unless values are present for these fields. In summary, incomplete records or records that do not conform to the required data type would be rejected from loading into such tables.

A higher level of data standardization concerns the prevention of loading values that are “legal” so far as the data structure is concerned, but undesirable for one reason or another. Standard value naming conventions fall under this category. Inconsistent naming of identical entities can obfuscate the data and significantly reduce its utility. When different data owners refer to the same standard value, e.g., a location, by different names, retrieving all data for that value becomes very difficult. To ensure consistent nomenclature, *Oncor* uses a coded-field scheme for certain standard values, such as location, instrument name, sampling method, person name, and others. An integer value, rather than a text string, uniquely defines these standardized entities. This allows users to maintain multiple string names (aliases) for identical entities. For example, one user may refer to a location as “BBM,” while another may use the term “Baker Bay Marsh” and a third just “Baker Bay.” *Oncor* assigns a single `Location_ID` integer value for this site, but allows each user to refer to it, both in their load files and in their data queries, using his or her own name or “alias.”

In addition to maintaining naming conventions, the *Oncor* loader performs various checks on incoming data to ensure its validity before being entered into the database. These checks include verifying that required fields are present. For example, most measurement values are meaningless if they lack a measurement date associated with them. Values are also checked for “reasonableness,” such as whether measurements dates are in the past or water temperatures are below boiling. In addition, the data exchange templates (DETs) provide a structure that enforces certain standards by supplying the fields and

record types required for each data category. The DRPs and data exchange template (DETs) document the overall requirements for data loaded into the database. These rules include standards that cannot be enforced through automated checks, but that must be adhered to in order to maintain consistency in the database. For example, in specifying the time standard used for measurement dates or the units used for specific measurements, software can detect if a measurement date erroneously occurs in the future, but it cannot determine if the date uses local daylight savings time, as it should. Data generators are responsible for providing their data in compliance with these rules.

Data that are not compatible the *Oncor* data standards will result in rejection of records by the data loader, leading to data-entry delays and possible frustration for the data generator. *Oncor* will use the following mechanisms to facilitate data standards:

1. **Data exchange template.** The DET is the primary data-standards enforcement mechanism. DETs allow users to use their own data standards (aliases) to the greatest degree possible. The DET enforces standards by specifying the fields required for each subject area, defining how they may be filled, and automatically checking selected fields using simple validity tests before being accepted for loading.
2. ***Oncor* data standards.** Data generators will be responsible for adhering to the *Oncor* Data Standards. While *Oncor* can facilitate many aspects of data standardization, it is impractical to guarantee accurate data entry without the cooperation of the data generator.
3. **Data custodian.** The data custodian is responsible for establishing data standards, ensuring that data are loaded into the database correctly, and assisting data generators with loading issues.

A.2 Standardized Values

Standard values are identified and defined in this appendix and on the *Oncor* website, and managed by the data custodian. The *Oncor* coordination process will manage the definitions of standard values and may help resolve issues relating to values. Current standard values are as follows:

- **Agency.** Organizations associated with lower Columbia and estuary (LCRE) data (e.g., Lower Columbia River and Estuary Partnership, Columbia Land Trust, etc.).
- **Person.** Full name and affiliation of people associated with LCRE projects.
- **Program.** Highest organizational level guiding a data collection effort; missions, research questions, and protocols are defined at this level (e.g., Columbia Estuary Ecosystem Restoration Program; CEERP)
- **Project.** Specific research activity that supports one or more of the research questions of the Program (e.g., Reference Site Study).
- **Document.** Reference document citation provided as supportive material for methods that may be referenced in various DRWs.
- **Instrument.** Specifications for equipment used to collect data (e.g., data logger, fish net, etc.).
- **Method.** Data collection method.
- **Species.** Specific information for a plant or animal (e.g., scientific name, species code, common name, etc.).

- **Unit.** Measurement units reported.

Standard-value records are stored in *Oncor* as “groups.” In the *Oncor* data model, a group record provides the hierarchical data structure that relates associated data together. For example, storage of a **Person** consists of the group name **Person_ID** having a unique long-integer code for each distinct person in the database. Associated with every **Person_ID** are the attributes **Person_LastName**, **Person_FirstName**, **Person_MI**, **Person_Agency**, and possibly others.

A.3 Alias Names

An important component of standard-value management is the concept of alias names. It is recognized that users will have preferred names for many of the standard values. So that users are not forced to adopt standard names, the *Oncor* data model allows access to standard-value records by multiple reference names. This is accomplished using an alias table, an example of which is shown below.

Table C.1. Example Alias Table

Group_Name	Code_Value	Alias_Group	Alias_Name
Instrument_ID	333	<i>Oncor</i>	HOBO #2
Instrument_ID	444	<i>Oncor</i>	HOBO #6
Instrument_ID	555	<i>Oncor</i>	HOBO #1
Instrument_ID	666	<i>Oncor</i>	HOBO #12
Instrument_ID	555	PNNL Default	HOBO #1
Instrument_ID	555	PNNL Short	H1
Instrument_ID	666	PNNL Default	H2
Instrument_ID	555	USGS	Old Logger
Instrument_ID	666	USGS	New Logger
Instrument_ID	444	USACE	Data Logger
Instrument_ID	333	USACE	H2
Instrument_ID	555	USACE	H1
Location_ID	10001	<i>Oncor</i>	Baker Bay Marsh
Location_ID	10001	PNNL Short	BBM
Location_ID	10001	USGS	Baker Bay
Location_ID	10001	USACE	Baker Bay Marsh

The columns in the alias table contain the following information:

- **Group_Name:** group name of the standard value (e.g., **Person_ID**, **Instrument_ID**, etc.).
- **Code_Value:** long-integer value of specific instance of the group. (A long-integer value is a computer variable data type that can take on values between -2^{31} and $+2^{31}$. Some data in *Oncor* use long-integer codes, rather than text-string names, to uniquely identify records.)

- **Alias_Group:** a group of related aliases belonging to a specific data user that may be requested for output in a query. Every standard value has at least one alias that is in the **Alias_Group** called “*Oncor*”, which is the default name used in query results. Associations between Alias_Groups and users
- **Alias_Name:** name associated with **Code_Value** for given **Alias_Group**.

Note that a single user may refer to a specific standard value by multiple names, which are distinguished by the Alias_Group value. For example, a single user at Pacific Northwest National Laboratory (PNNL) may be associated with both the “PNNL Default” and “PNNL Short” Alias_Groups and have the option of referring to instrument 555 by either “HOBO #1” or “H1”, respectively. Different users may call different entities by the same name. For example, users associated with “PNNL Default” and “USACE” both have instruments called “H2”, but they refer to different actual units: 666 and 333, respectively. *Oncor* will always maintain one default name under the Alias_Group “*Oncor*” for each standard value. This will be the preferred name and will appear in query results unless a user specifically requests a different Alias_Group.

Upon entry into the *Oncor* community, every data generator will provide information about his or her user-specific standard-value information to the data custodian. This will be done via the Metadatasheet of the DET using a Standard Values button. The data custodian will verify that the necessary information for each type of standard value is complete and will resolve duplicates, making sure that no two standard values point to the same object. The data custodian will be responsible for building and maintaining the alias table. The alias table will not only translate user-specific nomenclature at load time, but will also provide lookup lists for appropriate fields in the DET that will force data generators to enter only valid names. Data generators may continue providing the data custodian with new aliases and standard values via the same method.