# Purpose

The purpose of this document is to provide data users and stakeholders a standardized, structured/step-wise protocol for performing a fitness-for-use (i.e., data quality) assessment of datasets included in the EDTF data inventory spreadsheet. Ultimately, this document is intended to assist the EDTF in assessing whether a given dataset should be considered “preferred” (i.e., data that meets quality standards and is potentially useful to consider in regional transmission planning). Data users and stakeholders include EDTF members, other subject matter experts (SMEs), and geographic information system (GIS) personnel. The data inventory spreadsheet consists of a Microsoft© Excel spreadsheet that is updated as new datasets become available. The current version of the data inventory spreadsheet is available in the [documents section](https://www.wecc.org/committees/BOD/TEPPC/SPSG/EDTF/Shared%20Documents/Forms/AllItems.aspx) of the EDTF webpage.

# Data Quality Assessment Process Overview

The current process for determining which datasets are preferred for consideration in regional transmission planning consists of the following three steps:

1. *Technical Dataset Quality Assessment*: A technical GIS-based assessment of the quality or fitness-for-use of individual datasets, to the extent that such quality or fitness-for-use can be reasonably ascertained by reviewing a dataset’s metadata and data features in GIS. A GIS analyst performs this step.
2. *Subject-Matter Dataset Quality Assessment*: A subject-matter assessment of fitness-for-use of the scientific content of individual datasets.[[1]](#footnote-1) For example, an assessment of a dataset that contains the results of a wildlife habitat model based on vegetation, water features, terrain conditions, etc. A SME in the field of wildlife biology would evaluate the modeling methodology, date of data compilation, and other factors to determine the dataset’s fitness-for-use. In this example, an appropriate SME might include a stakeholder from a state fish and wildlife department or a wildlife conservation representative for the EDTF.
3. *EDTF Relevance Review*: Beyond data quality, the EDTF would meet to discuss the relevance (i.e., usefulness) of the datasets for consideration in regional transmission expansion planning. This step would involve a review of the findings from steps 1 and 2. Ultimately, the EDTF makes the decision about which datasets are considered preferred data.

**Figure 1** highlights the relationship of these processes, while detailed descriptions of implementation steps begin on page 2.

**Figure 1.** Data Quality Assessment Process

# Detailed Data Quality Assessment Process Steps

### Technical Dataset Quality Assessment

This step involves a review of fitness-for-use based on quality components derived from the United States Geological Survey (USGS) Spatial Data Transfer Standards (SDTS)[[2]](#footnote-2) criteria set. This SDTS criteria set is a national spatial data transfer mechanism for the United States that includes the following distinct criteria: lineage, positional accuracy, attribute accuracy, logical consistency, and completeness. The SDTS criteria were adapted to derive the quality components and metrics identified below.

Because it involves review of data covering much of western North America, this step is not intended to serve as a Quality Assurance/Quality Control (QA/QC) of datasets (e.g., whether a particular feature’s geometric shape accurately represents on-the-ground conditions), nor is it intended to edit or repair datasets. Instead, this step is intended to assess quality components of readily available and observable information and characteristics of the datasets. See *Step 2 Subject Matter Expert Data Quality Assessment* for information about how this independent testing and data validation would occur.

### Spreadsheet Entries

The EDTF data inventory spreadsheet contains the following quality assessment metrics and quality components, stored as data columns within the spreadsheet tabs for Federal, State, Non-governmental Organization (NGO), Decision Support System (DSS), and Tribal data categories:

* Reviewer
* Review Date
* Metadata
* Lineage
* Compilation Scale
* Positional Accuracy
* Seamless
* Completeness
* Currency
* Attribution
* Measurement Scale
* Overall Usability
* Quality Comments

The following is a description of the methods ICF GIS analysts will employ to assess and record observations about each quality component across all identified datasets. For any component, the following abbreviations may be used:

TBD: Measurement is to be determined (to be used instead of blanks for entries requiring completion)

UNK: Measurement is unknown or cannot be determined from available information

Reviewer – The initials of the analyst who performed the quality assessment for a given dataset.

Spreadsheet entry:

Initials of reviewer (DM; JW; EP; KA)

Review Date – The date the analyst performed the quality assessment for a given dataset.

Spreadsheet entry:

Date (Year\_MonthDay)

Metadata – A metadata record is a file of information, usually presented as an XML document, that captures the basic characteristics of a data or information resource. Metadata records enable data users to ascertain the other quality components (those listed above) of the data. The metadata entry will indicate whether there is electronic metadata for the dataset and, if so, its level of completion. The reviewer will read the entire metadata using ArcCatalog’s metadata viewer.

Spreadsheet entry:

C = Complete
SC = Substantially complete (80% or more complete)
PC = Partially Complete (10 to 80% complete)
A = Absent (0 to 10% complete)

Add comment if:
There is something in the metadata that bears on the reliability or quality of the dataset beyond the quality components described herein

Lineage – This entry indicates whether the lineage (history of processing) can be determined through examination of the metadata.

Spreadsheet entry:

Y = Yes, the lineage information is substantially complete
N = No, the lineage information is absent or incomplete

Compilation Scale – This entry stores the map scale, or aerial scale, at which the dataset was compiled. This information can be found under “Data Quality Information” or in the content descriptions of the metadata, or on the data source’s website.

Spreadsheet entry:

Scale denominator (e.g., 24,000 for a 1:24,000 scale map)

Positional Accuracy – This entry stores the stated or inferred horizontal accuracy of the features. If positional accuracy is not explicitly provided in the metadata, it may be estimated by the map compilation scale (if known) using the following guide from the National Map Accuracy Standards:

*For maps on publication scales larger than 1:20,000, no more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50 inch.[[3]](#footnote-3)*

Using this rule, a 1:100,000 map would calculate to 100,000/12/30 = 278 feet positional accuracy.

Spreadsheet entry:

Distance in U.S. feet

Add comment if:
Entry is calculated from map scale

Seamless – This entry states whether the dataset is geographically continuous within either the U.S. portion, the Canada portion, or the Mexico portion of the WECC area (as seamless as can be expected by the particular content of the dataset). For example, is there an absence of abrupt discontinuities at state or provincial lines? The analyst will enter a comment if data do not cover the entire WECC area.

Although the WECC area covers portions of three countries, few available environmental datasets cross international boundaries. Analysts will add comments to the data inventory spreadsheet as applicable. It should be noted that just because a dataset is not available for the entire WECC region, this does not mean that it fails the fitness-for-use test. For example, some datasets, such as wildlife agency data, might be available only at the state or province level; however, because these data might represent relevant data for environmental and cultural features not available elsewhere, the EDTF could decide to consider it as a potential preferred dataset.

Spreadsheet entry:

Y = Yes or N = No

Add comment if:
The entry is N, and if that is due to discontinuities at administrative boundaries
The dataset seamlessly crosses the Canada and/or Mexico – U.S. boundary

Completeness – This entry states whether the dataset has complete representation of features. A dataset is not considered complete if there are some areas for which data features are missing (e.g., the features are still in development).

Spreadsheet entry:

Y = Yes or N = No

Add comment if:
The entry is N, explain what is missing and why.

Currency – This entry states the date of data compilation and provides an indication of the currency of the dataset. The date can be obtained from the metadata (Federal Geographic Data Committee [FGDC] ESRI style) “Publication Information.”

Spreadsheet entry:

Date of compilation or publication, if provided

Add comment if:
The metadata does not provide a publication date but states that the dataset is updated as needed, or provides other qualitative information

Attribution – This entry states the quality and usability of the feature’s tabular attributes, based on a qualitative assessment of the usability of attributes to assess environmental sensitivity.

Spreadsheet entry:

Adequate
Inadequate

Add comment if:
Elaboration on attribute completeness or quality is needed

Measurement Scale – This entry states the type of measurement scale on which the primary attribute value is based.[[4]](#footnote-4)

Spreadsheet entry:

Nominal or categorical (the value is a name or text, e.g., a species name)
Ordinal (an ordered set of numbers, e.g., a ranking of best to worst)
Numerical (values on a numeric scale, either Ratio or Interval under the Stevens classification)
Binary (a numeric or text value to indicate the presence or absence of a feature; an implied binary measurement scale is used in the absence of an attribute value, where the inclusion of a geometric feature implies that the feature exists)

Overall Usability – This entry provides an overall qualitative assessment of the usability of the dataset to support regional environmental assessments, based on the quality components listed above.

Spreadsheet entry:

Good
Fair
Poor

Quality Comments – This field stores comments to clarify any entries for the quality components or any information provided by SMEs and/or EDTF members.

### Assessment Steps

Operationally, the GIS analyst will perform the fitness-for-use assessment through the following steps:

1. Open the current version of the data inventory spreadsheet for editing[[5]](#footnote-5).
2. Begin assessment of a particular dataset.
3. Download and, if necessary, unzip the data to the established project folders.
4. Open the dataset (or a representative example) in ArcCatalog.
5. Determine whether metadata is present.
6. If metadata is present, thoroughly review it using the different viewing styles in ArcCatalog, depending on what item is being investigated.
7. Inspect the metadata to try to assess each of the other quality components (as listed above), and complete the spreadsheet entries for those components.
8. After reviewing the metadata, peruse the data itself, or a representative sample, in ArcCatalog or ArcMap, for:
	1. Geometry integrity
	2. Attribution integrity
	3. General usability
9. If there are any remaining unknown values for any quality components in the assessment, attempt to ascertain quality through the data source website or other published sources.
10. Save the spreadsheet with a name indicating the version date.
11. Transmit the spreadsheet to relevant SMEs (e.g., WGA) for assessment, along with the dataset and any metadata (see step 2).

### Subject Matter Expert Dataset Quality Assessment

This step involves a SME review of the content of the dataset. This step in the data quality assessment process involves the use of a set of additional assessment criteria: potential applicability, collection/compilation methods, collection/compilation date, modeling approach, and methodology.

It is anticipated that membership of the EDTF and other WECC bodies could serve as the SMEs for most data sets. For example, discussion occurred at the December 09, 2011 meeting of the EDTF about the WGA serving as a SME reviewer for state wildlife data from the forthcoming crucial habitat assessment tools. The dates or review periods for SME reviews to occur can be documented in the EDTF Engagement and Outreach Plan.

### Spreadsheet Entries

The data inventory spreadsheet contains the following subject-matter quality assessment metrics and quality components, stored as data columns:

* SME Reviewer
* SME Review Date
* Potential Applicability
* Data Collection/Compilation Methods
* Data Collection/Compilation Date
* Modeling Approach/Methodology (if applicable)

The following describes the methods SMEs will employ to assess and record observations for each quality component across all identified datasets. For any component, the following abbreviations may be used:

TBD: Measurement is to be determined (to be used instead of blanks for entries requiring completion)

UNK: Measurement is unknown or cannot be determined from available information

SME Reviewer – The initials of the SME who performed the quality assessment for a given dataset.

Spreadsheet entry:

Initials of reviewer (e.g., PO)

SME Review Date – The date the SME performed the quality assessment for a given dataset.

Spreadsheet entry:

Date (Year\_MonthDay)

Potential Applicability – A preliminary assessment of whether the areas or features mapped in a given dataset have potential relevance to transmission planning. Applicability can be a function of the nature or legal protection of the mapped resource, the potential magnitude and type of impacts transmission development could present to the feature or area, or the potential risk to transmission planning the feature or area represents. NOTE: While this step can provide a valuable perspective, the EDTF would ultimately decide the relevance of a given data set to transmission planning and whether that dataset should be considered preferred data (see step 3).

Spreadsheet entry:

H = High relevance

M = Moderate relevance

L = Low relevance

Data Collection/Compilation Methods – An assessment of whether the methods used to collect and compile the data are reliable, meet industry standards, and would be expected to lead to producing a high-quality dataset. To assess methods, the SME should rely on the dataset metadata and other information provided by the data source (e.g., through websites and telephone consultations) in addition to his/her own experience in data collection, statistics, and information management.

Spreadsheet entry:

E = Excellent methods employed
S = Standard satisfactory methods employed

U = Unsatisfactory methods employed

Data Collection/Compilation Date – An assessment of whether the date of data compilation is acceptable for producing a high-quality dataset (e.g., if a dataset has been at least partially derived from vegetation or land cover features that are changeable over time).

Spreadsheet entry:

E = Excellent; the age of data exceeds all assessment requirements
S = Satisfactory; the age of data meets minimum assessment requirements

U = Unsatisfactory; the age of data does not meet minimum assessment requirements

Modeling Approach/Methodology – An assessment of the acceptability of methods used to process, derive, or model data to create the dataset. For example, a dataset might be derived from a wildlife habitat model based on vegetation, water features, terrain conditions, etc. The SME would examine the modeling approach and assess its validity and conformance with industry standards, which would affect the reliability and usability of the results. To assess the modeling approach and methodology, the SME should rely on the dataset metadata and other information provided by the data source (e.g., through websites and phone consultations) in addition to his/her own experience in data modeling.

Spreadsheet entry:

E = Excellent; the modeling approach exceeds requirements
S = Satisfactory; the modeling approach meets minimum assessment requirements

U = Unsatisfactory; the modeling approach does not meet minimum assessment requirements

### Assessment Steps

Operationally, the SME will perform the fitness-for-use assessment as follows:

1. Receive a copy of the data inventory spreadsheet and dataset.
2. Begin assessing a particular dataset. Review the entries from the Technical Dataset Quality Assessment.
3. Open the dataset (or a representative example) in ArcCatalog.
4. If metadata is present, thoroughly review it using the different viewing styles in ArcCatalog, depending on the item being investigated.
5. Inspect the metadata to try to assess each of the subject-matter quality components (as listed in Section 2.1) and complete the spreadsheet entries for those components.
6. After reviewing the metadata, peruse the data itself, or a representative sample, in ArcCatalog or ArcMap, for quality and completeness of content.
7. If there are any remaining unknown values for any quality components in the assessment, attempt to ascertain quality through the data source website or other published sources. Conduct telephone or email interviews with data providers as needed.
8. Complete the relevant sections of the data inventory spreadsheet and transmit the spreadsheet for compilation and completion to the party responsible for maintaining the master inventory version.

### EDTF Relevance Review

Upon completion of steps 1 and 2, the EDTF would meet to discuss the relevance and potential usefulness of the data sets for consideration in regional transmission expansion planning. Such meetings will likely occur under a schedule agreed to in the Stakeholder Engagement and Outreach Plan. At such meetings, the EDTF would likely need to review the determinations made in steps 1 and 2, and then determine whether the inclusion of the dataset as preferred would add value to transmission planning. For datasets the EDTF determines to be preferred, the task force would also likely need to decide the effect of the features/areas in the dataset on transmission planning (e.g., to what Risk Classification Category should it be assigned?).

1. As data collection and review commenced for the EDTF’s [Environmental Recommendations for Transmission Planning report](https://www.wecc.org/committees/BOD/TEPPC/SPSG/EDTF/Shared%20Documents/Report%20to%20SPSG/Environmental%20Recommendations%20for%20Transmission%20Planning%20-%20Synopsis%20Revised%2005-27-2011.pdf) (May 2011), EDTF members provided SME reviews. Following the December 09, 2011 meeting of the EDTF, the SME assessment process was formally added as a step in the Data Quality Protocol. [↑](#footnote-ref-1)
2. United States Geological Survey, National Mapping Division. Spatial Data Transfer Standard (SDTS) – Part 1, Logical Specifications, pp. 15-17. 1997, available at: <http://www.ncrs.fs.fed.us/pubs/databases/archive_info/metadata/SDTS_part1.pdf> [↑](#footnote-ref-2)
3. United States Geological Survey. Map Accuracy Standards. Fact Sheet FS-171-99 (November 1999). <http://egsc.usgs.gov/isb/pubs/factsheets/fs17199.html> [↑](#footnote-ref-3)
4. Measurement scale is an adaptation of the Stevens theory of scale types, as explained in the following document: http://en.wikipedia.org/wiki/Level\_of\_measurement. [↑](#footnote-ref-4)
5. As of the publication of this version of the Data Quality Assessment Protocol, ICF maintains the master version of the data inventory spreadsheet and posts updates to the EDTF webpage at milestone dates. [↑](#footnote-ref-5)