Building Science Institute, Ltd. Co. Procedure F-2023 Data Quality Assessment

Data Quality Dimensions

1. **Reliability**: the completeness, relevance, accuracy, uniqueness, and consistency of the dataset for the intended purposes of use, and the ability to trace the data to a trustworthy source.

    a. BSI measures *Reliability* as a benchmark for Verification Organizations through annual conformity assessments.

    b. BSI reports *Reliability* scores to the US EPA, US DOE, Quality Council, and Verification Organizations.

    c. BSI publishes the Verification Organization's *Reliability* scores in the annual conformity assessment report.

    d. BSI publishes the aggregate scores of our clients on the BSI website.

    e. *US EPA/US DOE could measure Reliability as a benchmark for BSI.*

    f. Key steps to measure *Reliability*

        i. Assess the risk related to data reliability

            1. The data in the context of the final report and the extent of the role the data will play through these Risk Reliability criteria (RRC)

                a. will the data be used to influence legislation or policy with significant impact? "Yes" = 1, "No" = 0

                b. will the data be used for significant decisions by individuals or organizations? "Yes" = 1, "No" = 0

                c. will the data form the basis for numbers that are likely to be widely quoted or published? "Yes" = 1, "No" = 0

                d. will the data form the basis of a regulatory or compliance project? "Yes" = 1, "No" = 0

            2. Nature of engagement through these Engagement Nature criteria (ENC)

                a. is the engagement concerned with a sensitive or controversial subject? "Yes" = 1, "No" = 0

                b. does the engagement involve external stakeholders who have taken positions on the subject? "Yes" = 1, "No" = 0

                c. is the overall engagement risk medium or high? "Yes" = 1, "No" = 0

d. does the engagement have unique factors that strongly increase risk? "Yes" = 1, "No" = 0

3. *Reliability* risk is scored by (RRC1 + RRC2 + RRC3 + RRC4 + ENC1 + ENC2 + ENC3 +ENC4) / 8

    a. The closer your *Reliability* Risk score is to "1", the data reliability risk is significantly high and you need a high level of confidence in the data.

ii. Identify the attributes required for intended purposes

1. Identify whether data elements in the data set are relevant through these Relevance criteria (RC)

    a. Are the data elements in the dataset sufficient? "Yes" = 1, "No" = 0

    b. Are the identified data elements relevant? "Yes" = 1, "No" = 0

    c. Do the data stored in the data elements actually represent what you think you are measuring? "Yes" = 1, "No" = 0

    d. Are any data elements in the dataset missing that would affect the desired outcome? "Yes" = 1, "No" = 0

2. Relevance *Reliability* is scored by (RC1 + RC2 + RC3 + RC4) / 4

    a. The closer your Relevance *Reliability* score is to "1", the more reliably relevant the data is.

iii. Determine whether data are sourced from trustworthy sources

1. Define and apply data quality business rules for analyzing the data through these quality criteria (QC)

    a. Are any values of the key data elements missing? "Yes" = 1, "No" = 0

    b. Does the data demonstrate the relationship of one data element to another? "Yes" = 1, "No" = 0

    c. Is the data coverage sufficient? "Yes" = 1, "No" = 0

    d. Are the data values accurate? "Yes" = 1, "No" = 0

    e. Does the dataset have erroneous duplicates? "No" = 1, "Yes" = 0

2. Review system controls

    a. What checks are performed before data are stored in the system? (Control criteria, CC)

        i. Are completeness checks performed before data are stored in the system? "Yes" = 1, "No" = 0

            ii. Are limit and range checks performed before data are stored in the system? "Yes" = 1, "No" = 0

            iii. Are sign checks performed before data are stored in the system? "Yes" = 1, "No" = 0

            iv. Are validity checks performed before data are stored in the system? "Yes" = 1, "No" = 0

        3. Source *Reliability* is scored by (QC1 + QC2 + QC3 + QC4 + QC5 + CC1 + CC2 + CC3 + CC4) / 9

    iv. Two factors to consider whilst assessing errors in a data file

        1. Proportion of data that are erroneous

        2. Magnitude of the error

    v. Data is "Sufficiently Reliable" when:

        1. Reliability Risk $\geq$ 0.125 and Relevance Reliability + Source Reliability $\geq$ 0.82 and

        2. The likelihood of significant errors or incompleteness is minimal

        3. the use of the data would not lead to an incorrect or unintentional message or drive incorrect decisions

    vi. Data is "Not Sufficiently Reliable" when:

        1. Relevance Reliability + Source Reliability < 0.82 and

        2. Significant errors or incompleteness in some or all of the key data elements

        3. Use of the data would probably lead to an incorrect or unintentional message

    vii. Data is "Undetermined Reliability" when:

        1. Limited or no access to a data source

        2. Time limitations

        3. A wide range of data that cannot be examined with current resources

        4. The deletion of original computer files

        5. Insufficient documentation about the data

        6. Lack of access to needed documents

2. **Timeliness**: the degree to which the period between the time of creation of the real value and the time the dataset is available is appropriate

    a. BSI measures *Timeliness* as a benchmark for Verification Organizations and BSI.

   b. BSI reports *Timeliness* scores to US EPA, US DOE, Quality Council, and Verification Organizations.

   c. BSI publishes the Verification Organization's *Timeliness* score in the annual conformity assessment report.

   d. BSI publishes the aggregate *Timeliness* score of our clients on the BSI website.

   e. *Timeliness* is measured by data delivery time - occurrence time.

      i. HouseRater data export "Timeliness Scoring Check"

3. **Validity**: the extent to which data is present, in the correct format, within the accepted range, of the correct lookup type

   a. BSI measures *Validity* as a benchmark for Verification Organizations.

   b. BSI reports *Validity* scores to US EPA, US DOE, Quality Council, and Verification Organizations.

   c. BSI publishes the Verification Organization's *Validity* score in a monthly report to the Verification Organization.

   d. BSI publishes the aggregate *Validity* score of our clients on the BSI website.

   e. *Validity* is measured by number of valid records relative to total records.

      i. (number of valid records in dataset x 100) / number of records in dataset or

         1. 1 - ((number of records in data set - number of invalid records in data set) / number of records in data set)

      ii. HouseRater data export "Validity Scoring Check"

4. **Compliance/Conformance**: the degree to which data and composition of datasets is in accordance with laws, regulations, or standards

   a. BSI measures *Compliance/Conformance* during our file/field conformance assessments as a benchmark for Verifiers and Verification Organizations.

   b. BSI reports *Compliance/Conformance* scores to US EPA, US DOE, Quality Council, and the Verification Organization.

   c. BSI publishes the Verification Organization's *Compliance/Conformance* score in a monthly report to the Verification Organization.

   d. BSI publishes the aggregate score of our clients on the BSI website.

   e. *Compliance/Conformance* is measured by comparison of data to requirements of laws, regulations, or standards.

      i. (number of conforming values of a data element in dataset x 100) / number of values for the data element in dataset or

1. 1 - ((number of values for the data element in dataset - non-conforming values of a data element in dataset) / number of values for the data element in dataset)

5. **Accuracy**: the degree of correspondence between data values to real values.

    a. BSI measures *Accuracy* as a benchmark for Verifiers & Verification Organizations during file conformity assessments.

    b. BSI reports *Accuracy* scores to US EPA, US DOE, Quality Council, and Verification Organizations.

    c. BSI publishes the Verification's *Accuracy* score monthly to the Verification Organization.

    d. BSI publishes the aggregate *Accuracy* score of our clients on the BSI website.

    e. *Accuracy* is measured by comparison between data collected and original source,

    i. (number of accurate data elements x 100) / number of data elements) or

    1. 1 - ((number of data elements - number of inaccurate data elements) / number of data elements)

    ii. (number of accurate data records x 100) / number of data records or

    1. 1 - ((number of data records - number of inaccurate data records) / number of data records)

6. **Completeness**: the extent to which data is not missing and is of sufficient breadth and depth for the task at hand; all required data values are present, all required records in the dataset are present, all required attributes in the dataset are present, all required data files are present, metadata are fully described.

    a. BSI measures *Completeness* as a benchmark for Verification Organizations.

    b. *US EPA & US DOE could measure Completeness as a benchmark for BSI.*

    c. BSI publishes the Verification Organization's *Completeness* score monthly to the Verification Organization.

    d. BSI publishes the aggregate *Completeness* score of our clients on the BSI website.

    e. *Field Completeness* is measured by:

    i. (number of fields completed x 100) / number of total fields or

    1. 1 - ((number of total fields - number of incomplete fields) / number of total fields

    f. *Record Completeness* is measured by:

    i. (number of complete records x 100) / number of total records or

1. 1 - ((number of total records - number of incomplete records) / number of total records)

g. *Attribute Completeness* is measured by:

i. (number of complete attributes x 100) / number of total attributes or

1. 1 - ((number of total attributes - number of incomplete attributes) / number of total attributes)

h. *File Completeness* is measured by:

i. (number of complete files x 100) / number of total files or

1. 1 - ((number of total files - number of incomplete files) / number of total files)

7. **Consistency**: the degree to which data values of two sets of attributes within a record, within a data file, between data files, within a record at different points in time comply with a rule; the degree to which data values of two sets of attributes between records comply with a rule; the degree to which data values between two sets of attributes between datasets comply with a rule; the degree to which data values of a set of attributes of a dataset at different points in time comply with a rule.

a. BSI measures *Consistency* as a benchmark for Verification Organizations and HouseRater.

b. BSI publishes the Verification Organization's *Consistency* score in the Verification Organization's annual conformance assessment report.

c. BSI publishes HouseRater's *Consistency* score in a monthly report to HouseRater.

d. *Consistency* is measured by:

i. (number of consistent data element combination values in dataset x 100) / number of data element combination values in dataset or

1. 1 - ((number of data element combination values in dataset - number of inconsistent data element combination values in dataset) / number of data element combination values in dataset)

8. **Integrity**: the extent to which data is **A**ttributable, **L**egible, **C**ontemporaneously recorded, **O**riginal or true copy, **A**ccurate, with high levels of *Completeness* & *Consistency.*

a. BSI measures *Integrity* as a benchmark for Verification Organizations.

b. BSI publishes the *Integrity* scores to US EPA, US DOE, Quality Council, and Verification Organizations annually.

c. BSI publishes the aggregate *Integrity* score of our clients on the BSI website.

d. *Integrity* is the combination of the following integrity criteria (IC):

i. Is the data attributable to a known source? "Yes" = 1, "No" = 0

ii. Is the data legible? "Yes" = 1, "No" = 0

iii. Is the data contemporaneously recorded? "Yes" = 1, "No" = 0

iv. Is the data original or a true copy? "Yes" = 1, "No" = 0

v. Is the data accurate? fractional scale from *Accuracy* measurement, 0 to 1

vi. Is the data complete? fractional scale from *Completeness* measurement, 0 to 1

vii. Is the data consistent? fractional scale from *Consistency* measurement, 0 to 1

e. *Integrity* is measured by aggregate score of (A + L + C + O + A + C + C) / 7 using fractional scale of 0 to 1, 0 = less satisfactory and 1 = most satisfactory

9. **Data Pedigree**: the data pedigree is based on the following data pedigree criteria (DPC):

a. Does the data represent what it purports to represent? "Yes" = 1, "No" = 0

b. Was the data produced through a consistent & defined process? "Yes" = 1, "No" = 0

c. Did we get samples from that process? "Yes" = 1, "No" = 0

d. Is there a record of who created the data? "Yes" = 1, "No" = 0

e. Is there a record of when it was created? "Yes" = 1, "No" = 0

f. Are there controls over who has access to the data? "Yes" = 1, "No" = 0

g. Has the data been modified or deleted? "Yes" = 0, "No" = 1

h. Do we have access to the original? "Yes" = 1, "No" = 0

i. BSI measures *Data Pedigree* as a benchmark for Verification Organizations.

j. BSI publishes the *Data Pedigree* scores to the Quality Council annually.

k. BSI publishes the *Data Pedigree* scores to the Verification Organization in the annual conformity assessment report.

l. *Data Pedigree* is measured by (DPC1 + DPC2 + DPC3 + DPC4 + DPC5 + DPC6 + DPC7 + DPC8) / 8

10. **Objectivity**: the extent to which data values are created in unbiased, unprejudiced, and impartial manner.

a. BSI measures *Objectivity* as a benchmark for Verification Organizations.

b. *US EPA & US DOE could measure Objectivity as a benchmark for BSI.*

c. BSI publishes *Objectivity* scores to the US EPA, US DOE, Quality Council annually.

d. BSI publishes *Objectivity* scores for Verification Organizations in the annual conformity assessment report.

e. BSI publishes the aggregate *Objectivity* score of our clients on the BSI website.

f. *Objectivity* is measured by survey on these objectivity parameters (OP)

i. Is the information source authentic? "Yes" = 1, "No" = 0

ii. Does the publisher have a personal impact on the data provided? "Yes" = 1, "No" = 0

iii. Can the data be affected due to the organization sponsors or policy? "Yes" = 1, "No" = 0

iv. Is the accountability of information or data clearly defined? "Yes" = 1, "No" = 0

v. To what extent are the independent sources of data available for confirmation of facts? Values between 0 and 1, 0 = Lowest, 1 = Highest

vi. To what extent do processes for data collection, processing, and dissemination ensure objectivity? Values between 0 and 1, 0 = Lowest, 1 = Highest

vii. Per user objectivity measurement, (OP1 + OP2 + OP3 + OP4 + OP5 + OP6) / 6

viii.Overall objectivity measurement, sum of all surveys ((OP1 + OP2 +OP3 + OP4 + OP5 + OP6)/6) / number of users who participated in survey

11. **Uniqueness**: the degree to which there are no duplicate records for the same entity or event in the same dataset.

a. BSI measures *Uniqueness* as a benchmark for Verification Organizations.

b. *US EPA & US DOE could measure Uniqueness as a benchmark for BSI.*

c. BSI publishes the *Uniqueness* scores to the US EPA, US DOE, and Quality Council quarterly.

d. BSI publishes the *Uniqueness* score to the Verification Organization in a quarterly report.

e. BSI publishes the aggregate *Uniqueness* score of our clients on the BSI website.

f. *Uniqueness* is measured by a search for duplicate records:

i. (number of unique values of a data element in dataset x 100) / number of values for the data element in dataset or

1. 1 - ((number of values for the data element in dataset - duplicate values for the data element in dataset) / number of values for the data element in dataset)

12. **Credibility**: the degree to which data values are regarded as true and believable by data consumers.

a. *Credibility* is based on the following credibility criteria (CC):

i. The good faith of the data provider can be relied on to ensure the data represent what they are supposed to represent

ii. The good faith of the data provider can be relied on to ensure there has been no intent to misrepresent the data

iii. There is a guarantee that data are protected from unauthorized modification

b. BSI measures *Credibility* through a survey sent to US EPA, US DOE, and homebuilders.

c. BSI reports the *Credibility* score to the Quality Council in a quarterly report.

d. BSI publishes the *Credibility* score on the BSI website.

e. *Credibility* is measured with a fractional scale, 0 to 1, 0 = low credibility and 1 = high credibility

i. *Credibility* rating per user, (CC1 + CC2 + CC3) / 3

ii. Overall *Credibility* rating, sum of user ((CC1 + CC2 + CC3) / 3) / number of users who participated in survey

13. **Trustworthiness**: the extent to which the data originate from trustworthy sources.

a. BSI measures *Trustworthiness* through a survey sent to US EPA, US DOE, and homebuilders.

b. BSI publishes the *Trustworthiness* score to the US EPA, US DOE, and the Quality Council in a quarterly report.

c. BSI publishes the *Trustworthiness* score on the BSI website.

d. *Trustworthiness* is measured by a survey on the following trustworthiness parameters (TP) using a fractional scale 0 to 1, 0 = lower trustworthiness and 1 = higher trustworthiness:

i. Are the data sourced from an authoritative source or provider with a known control environment and track record? "Yes" = 1, "No" = 0, "Do Not Know" = 0

ii. Can data be traced to the source? "Yes" = 1, "No" = 0

iii. Number of complaints or issues reported on the data in the last six months? "0" = 1, "> 0 ≤ 3" = 0.75, "> 3 ≤ 5" = 0.50, "> 5 ≤ 10" = 0.25, "> 10" = 0

iv. Number of requests issued for data in the last six months? "0" = 0, "> 0 ≤ 3" = 0.25, "> 3 ≤ 5" = 0.50, "> 5 ≤ 10" = 0.75, "> 10" = 1

v. What is the degree to which reporting on data quality statistics is published? "Not at all reported" = 0, "Reported intermittently and incomplete to a greater degree" = 0.25, "Reported more or less periodically and incomplete to a greater degree, or reported intermittently and incomplete to a lesser degree" = 0.50, "Reported more or less periodically and incomplete to a lesser degree" = 0.75, "Reported periodically and complete" = 1

vi. Per user trustworthiness, (TP1 + TP2 + TP3 + TP4 + TP5) / 5

vii. Overall trustworthiness, sum of users ((TP1 + TP2 + TP3 + TP4 + TP5) / 5) / number of users who participated in the survey

14. **Believability**: the extent to which data is regarded as true and credible.

a. BSI measures the *Believability* score by:

        i. (*Credibility* + *Trustworthy*) / 2

    b. BSI publishes the *Believability* score to the US EPA, US DOE, and the Quality Council quarterly.

    c. BSI publishes the *Believability* score on our website.

15. **Reputation**: the extent to which data is highly regarded in terms of its source or content.

    a. BSI measures *Reputation* as a benchmark for BSI.

    b. BSI publishes the *Reputation* score to US EPA, US DOE, and the Quality Council quarterly.

    c. BSI publishes the *Reputation* score on the BSI website.

    d. BSI measures *Reputation* through a survey sent to US EPA, US DOE, and homebuilders on the following reputation parameters (RP) using a fractional scale 0 to 1, 0 = lower reputation satisfaction and 1 = highest reputation satisfaction:

        i. The data source provides accurate data consistently

        ii. Data source issues, when found, are resolved quickly

        iii. The data source is recommended by other reputed data sources

        iv. Per user reputation measurement, (RP1 + RP2 + RP3) / 3

        v. Overall reputation measurement, sum of users ((RP1 + RP2 + RP3)/3) / number of users participating in survey

16. **Accessibility**: the ease with which data can be consulted or retrieved.

    a. BSI measures *Accessibility* as a benchmark for HouseRater.

    b. BSI reports *Accessibility* scores to HouseRater in a quarterly report.

    c. *Accessibility* is measured through a Likert Scale survey sent to BSI, US EPA, US DOE, homebuilders, and Verification Organizations on a scale of 1 - 4:

        i. 1 = Completely unacceptable (Low),

        ii. 2 = Marginally unacceptable (Slightly Low),

        iii. 3 = Marginally Acceptable (Slightly High),

        iv. 4 = Completely Acceptable (High)

17. **Ease of Operation**: the extent to which data is easy to manipulate and apply to different tasks.

    a. BSI measures *Ease of Operation* as a benchmark for HouseRater.

    b. BSI publishes the *Ease of Operation* score to HouseRater in a quarterly report.

    c. *Ease of Operation* is measured by a Likert Scale survey sent to BSI, US EPA, US DOE, and homebuilders with a fractional scale of 0 - 1:

    i. 0 = very difficult to manipulate

    ii. 1 = extremely easy to manipulate

Adapted from Data Quality: Dimensions, Measurement, Strategy, Management, and Governance *by Rupa Mahanti, Ph.D. Published 2018 by American Society for Quality, Quality Press*

Data Quality Assessment Procedure

Step 1: Assemble last 100 records your group used or created

1. Focus on 10-15 critical data elements or attributes
2. Lay them out on a spreadsheet

Step 2: Call a meeting

Ask 2 or 3 people who have knowledge of the data to join a 2-hour meeting. This Data Quality Assessment Procedure is known as the FAM, Friday Afternoon Meeting.

Step 3: Mark obvious errors in noticeable color (red or orange)

Work record by record, and have your colleagues mark obvious errors in red or orange

Step 4: Summarize the results

1. Add a "record perfect or not" column to spreadsheet
2. Mark record with "yes" if there aren't any errors and "no" if red or orange appear on the record.
3. Total the number of perfect records.
4. Divide number of perfect records by number of total records to get data quality score
5. Perform cost analysis for cost of bad data
    a. Bad data has a cost of roughly 10X of perfect data
6. Visualize the attributes with high number of errors through Pareto chart analysis

Step 5: Make Improvements

1. Use Root Cause Analysis Process to identify the root cause
2. Identify improvements that will solve the challenge
3. Have the people responsible for the data creation problem make the improvements

Adapted from "Assess Whether You Have A Data Quality Problem" by Dr. Thomas C. Redman, Ph.D. Published by Harvard Business Review, July 28, 2016

Approved by the Building Science Institute, Ltd. Co. Quality Council on May 22, 2023.
Approve: Kevin Burk, Erik Straite, Brian Christensen, Amber Wood
Reject: None
Not Voting: Wes Davis, Brett Dillon (Chair)