Committee for the Coordination of Statistical Activities

Conference on Data Quality for International Organizations
Newport, Wales, United Kingdom, 27 - 28 April 2006

Session 5: Tools and practices for collecting quality metadata from data providers and informing users on the quality of statistics

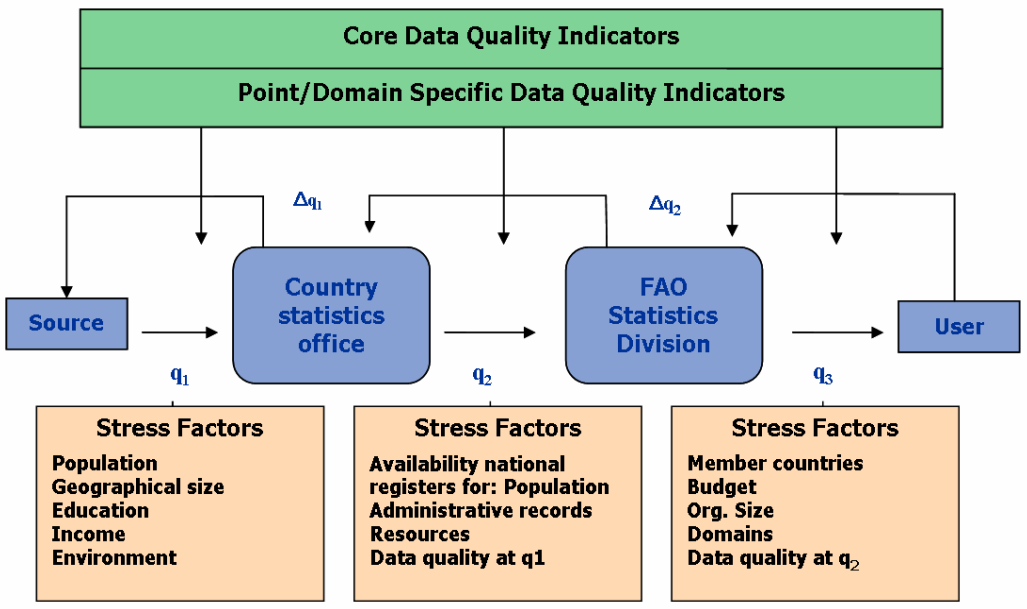
"Data Quality Measures and Related Stress Factors: A conceptual framework to account for differences in statistical environments at country and international levels"

Food and Agriculture Organization

Introduction and Background

This paper builds on the previous work of the United Nations Food and Agricultural Organization (FAO) on data quality frameworks: focusing on the essential components of data quality at key points of the statistical process and relating these essential quality components to the data quality framework. The paper also illustrates how summary data quality measures can be adjusted (deflated/inflated) depending upon the stress factors experienced by national or international statistical offices. At the previous Conference on Data Quality for International Organizations in Wiesbaden/Germany[1] in 2004, FAO presented a Data Quality Framework to monitor data quality in all points of the international data process (from country collection to FAO dissemination). Data quality evaluation and monitoring is the focus at three different points in the statistical process: when data enters the national office, when data leaves the national office, and when data is disseminated by FAO. This process takes full advantage of information on data quality available at these specific points in the statistical process (see *Figure 1*). The concept of Data Quality Stress Factors was introduced in the Data Quality Framework to recognize and make it possible to adjust data quality measures to specific situations where some national or international statistical offices find it easier to produce higher quality statistical outputs than others.

Figure 1. Data Quality monitoring and feedback system

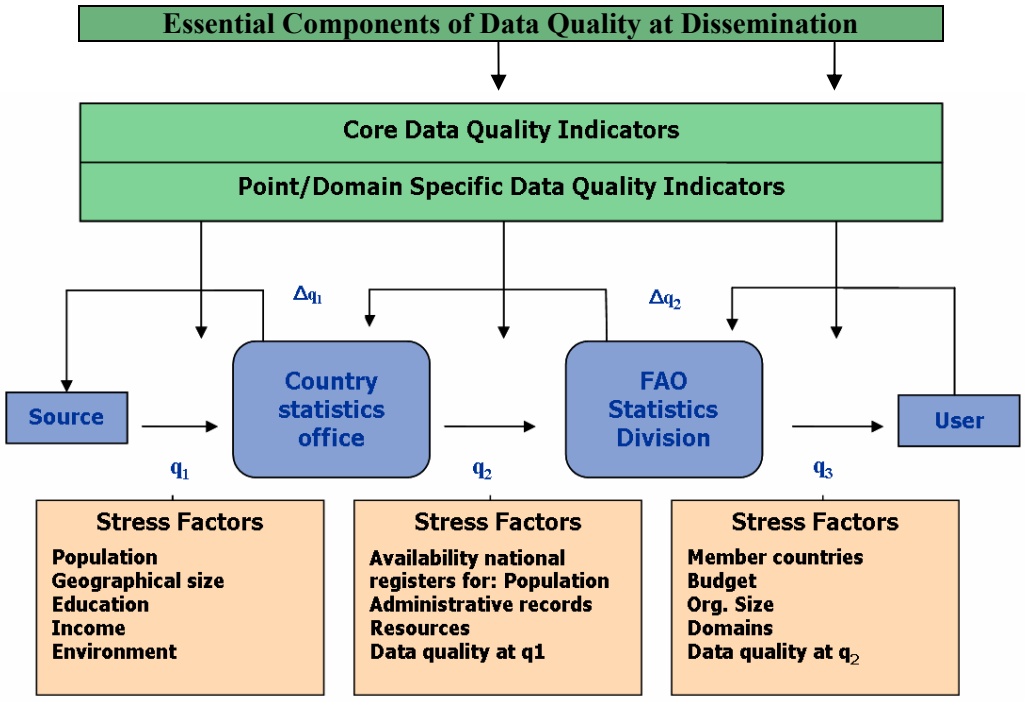


[1] http://unstats.un.org/unsd/accsub/2004docs-CDQIO/1-FAO.pldf

The Essential Components of Data Quality at Dissemination

The extensive number of data quality indicators proposed and used by various international statistical agencies (see *Annex 1*. for some examples) cover various dimensions of data quality. New data quality indicators have been introduced on a regular basis over the past few years. There development will continue especially as additional statistical domains review their need for data quality evaluation and monitoring and produce new indicators that meet their specific requirements. National and international statistics offices generally do not have the resources to monitor all their data on al the proposed data quality indicators for all aspects of the statistical process. It is therefore important to focus and rationalize what dimensions statistical offices need to concentrate on and for what particular aspect of the statistical process. This paper focuses on the dissemination aspects of the statistical processes and proposes data quality dimensions that are of "essential" importance at those points. A focus on the "essential" components of data quality is required in order to ensure that a minimum level of quality is assured in the statistical process at all stages. If data quality is not monitored for the appropriate dimensions of quality, then data that may be unfit for dissemination may subsequently be disseminated. The following discussion focuses on the proposed essential dimensions of data quality when disseminating data from national or international statistics office (i.e. points q_2 and q_3 in *Figure 2*).

Figure 2. Essential Components of Data Quality at Dissemination



Accuracy and reliability are the corner-stones of data quality. Without accuracy and **reliability** there is little further need to measure and monitor other dimensions of quality. Accuracy and **reliability** should be primarily assessed and monitored at the statistical element and item levels and summarized at the statistical domain level. Reporting on accuracy and **reliability** should be undertaken at both points q_2 and q_3 in the statistical dissemination process. The domain, element and item level accuracy assessments and reporting should then be included in the second essential component of data quality at dissemination – **Transparency and Metadata.** Without sufficient statistical metadata there is little understanding of the data item. Sufficient

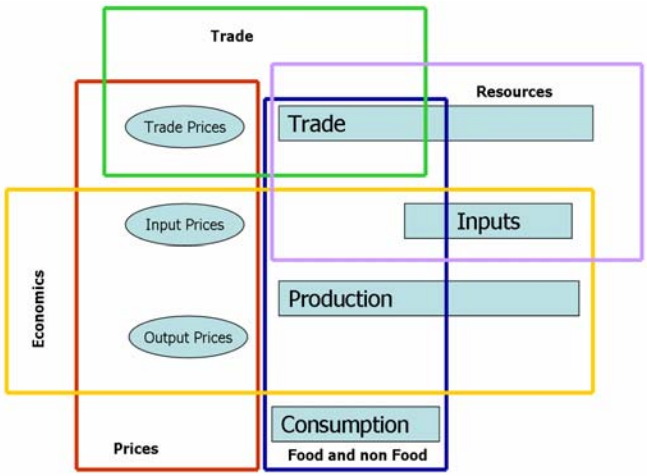
statistical metadata covering the standard items (concepts, definitions, source, etc) is required at both points q_2 and q_3 in the statistical dissemination process. Sufficient statistical metadata is also required at the domain and element levels for points q_2 and q_3 in the statistical dissemination process.

The third essential component is the use of **Standard Classifications**. When statistical data do not use or correspond to standard (common or international) classifications, they are isolated and of little use outside the specific data domains. Thus, the use of standard classifications is essential for points q_2 and q_3 in the statistical process. It is also essential for statistical domains (and by definition elements and items within those domains) to apply standard classifications.

The fourth essential data quality for the data is **Timeliness.** Data that does not reflect the reality of the situation that is being measured will likely have a negative impact on analysis and subsequent policy development that it is used for. Data does not need to come from the most recent data time-period, but it needs to be refreshed and available as appropriate. It is also essential for statistical domains (and by definition elements and items within those domains) to have data that is up-to-date. Having up-to-date data within a domain/element is as important as having current data at the dissemination point of national and international agencies (points q_2 and q_3) as at those data points it is used in a cross-sectional fashion and individual data that is out-of-date at the element level would negatively impact analysis and policy formulation.

The fifth essential component is the concept of **Comparable Data Domains**. Data can be isolated by not using common or international classifications, but it can also be isolated by the lack of integration. Integration can be in a physical form such as the data stored in formats which are not common or in databases that cannot communicate with other databases. Data can also be isolated in a conceptual context: that is data without an integrated framework. For example, over the past three years with the re-development of the FAOSTAT database and the development of CountrySTAT[2] the development of an integrated approach to food and agricultural statistical data has taken place at FAO. The following integrated conceptual framework was developed and is being adopted into the FAOSTAT and CountrySTAT systems (see *Figure 3*.). The sixth essential component of data quality at dissemination is the concept of **Accessibility.** Without suitable methods of dissemination, data looses substantial quality. By integrating data whether in a physical form or a virtual form data also gains quality by becoming more accessible.

Figure 3. FAOSTAT – CountrySTAT integrated conceptual framework



[2] http://www.fao.org/faostat

A statistical office, be it national or international, which wants to use and benefit from an integrated conceptual framework, needs to have data that conforms to the preceding essential components of data quality (accuracy, transparency - metadata, standard classifications and timeliness). The integrated framework provides additional value to the statistical data. In the agricultural domain, for example, data on the physical inputs to agriculture, the physical production resulting from agriculture, the input and out-put prices, will enable the calculation of: costs of production; value of production and the production of economic, not just physical accounts. The integrated conceptual framework dimension is important at the domain level and at the dissemination points of national and international agencies (points q_2 and q_3) where integrated conceptual frameworks add value to the individual data domains/elements.

An integrated approach also provides the statistical office with analytical tools to monitor quality, such as, having the possibility of preparing Supply and Resource Utilization Accounts in the agricultural sector. Using an integrated approach is dependent upon minimal levels of accuracy, metadata and standard classifications, otherwise integration is meaningless or impossible to achieve

Adding Additional Value to Data

International statistical offices can provide additional value (quality) for data beyond the essential components discussed in the previous section. The **scope and coverage** of data is one aspect for increasing the value (quality) of a data set/domain. At the domain/element level, coverage is important for the specific domain frameworks. Missing items or elements may negatively impact the domain. For example, when an international statistics office has 190 members and can produce datasets with data for all 190 members for an extensive time-series and within an integrated conceptual framework, it produces substantial value (quality) to the individual country data it collects. However, if there are breaks in the time-series or data is reported only for some member countries each year, then value (quality) is lost from the dataset/domain. The same applies to national statistics offices with respect to sub-national data or statistical domain coverage.

Providing a **data up-date schedule** for dissemination also adds value. Statistical users need up-to-date statistical data at the domain and element level, but they also information about the availability of data. Too often statistical offices overlook the importance of letting their users know when the next data will be released. This would enable the users to produce analysis and subsequent policy from the most appropriate data. For national and international statistical offices, the publishing of a regular dissemination schedule also establishes a target for providing timely data.

When data elements and domains are incomplete, **estimates** are often produced by national or international statistical offices. These estimations for missing observations add value to the data elements and domains and datasets. When consistency **calibrations** are performed on elements and domains, and data is adjusted, these adjustments can also be seen as adding value to the original data.
The production of **analytical indicators,** by putting together information from different sources and domains (see *Figure 3.*) also adds value to the data domain.

The essential and value adding components of data quality are not specific indicators of quality but the major dimensions that should be monitored as data is disseminated in the statistical process. Specific indicators of data quality can be general or specific, either for statistical domains, elements (statistical series) or for the points - q_1, q_2 and q_3 in the process. *Table 1*. shows nine essential and value adding components of data quality at dissemination for the element and domain levels for both point q_2 and q_3 in the statistical process. Once a statistical office identifies specific indicators for these dimensions and collects the necessary information, it can produce summary measures of statistical data quality performance at points q_2 and q_3 for statistical elements, domains and offices.

Table 1. Essential Components of Data Quality at the Key Points - q_2 and q_3

Essential and Value Adding Components of Data Quality at Dissemination				
	National office dissemination- q_2		**International office dissemination - q_3**	
	Element	**Domain**	**Element**	**Domain**
Accuracy	Yes	Yes	Yes	Yes
Metadata	Yes	Yes	Yes	Yes
Classifications	Yes	Yes	Yes	Yes
Timeliness	Yes	Yes	Yes	Yes
Comparable	Yes	Yes	Yes	Yes
Coverage	Yes	Yes	Yes	Yes
Up-date schedule	Yes	Yes	Yes	Yes
Estimates - calibration	Yes	Yes	Yes	Yes
New Analytical indicators	Yes	Yes	Yes	Yes

Measuring Overall Quality in the Statistical Process

It is important to monitor the overall quality of data disseminated by national and international statistical offices. However, problems arise when attempting to measure and compare quality between countries and analysing the quality of data from one country over time. The same problems exist when comparing quality of data produced by different international organisations and analysing the quality of data from one international organisation over time. As internal and external evaluations of the national and international statistical offices are becoming more frequent, a methodology that provides a "level playing field" for these evaluations is necessary.

Data quality can be considered in absolute and relative terms with reference to the performance of the statistical agency. The performance of a statistical agency operating under difficult conditions in the collection and production of statistical data should not be assessed in absolute terms and compared to other statistical agencies that function under less restrictive conditions. Absolute measures of data quality are more applicable for comparison within a statistical agency, whilst relative measures are more applicable for comparison between statistical agencies.

The FAO is developing a methodology to monitor and compare data quality primarily focussing on relative quality which could be applied at point q_2 in order to compare quality across countries or at point q_3 to compare of quality across international statistics offices. First, summary measures of statistical dissemination data quality performance at point q_2 for statistical components and/or domains need to be prepared. Second, a "benchmark statistics office" is selected which could be a virtual one from statistical offices working with FAO or a combination. The "benchmark statistics office" would be the office that scores highest on summary measures of statistical dissemination data quality performance at point q_2. Third, a data quality index would be prepared using the "benchmark statistics office" as the base for the index. The index levels would then be calculated for all other statistics offices on the country dissemination data quality index. The same methodology could also be applied to international statistics offices.

Table 2 provides an illustrative example of three countries where Country A achieves high quality and is the "benchmark office" thus being identified as having the "Reference Quality Level." The Data quality measurement is indexed to the base year 2002, thus Country A ranks 100 in 2002. Country data quality index levels are also calculated for countries B and C.

Table 2. Country Data Quality Index

Country Quality	Year 2002=100					
	2000	2001	2002	2003	2004	2005
A	93	95	100	103	103	105
B	70	75	75	80	80	80
C	70	77	75	70	65	65

This approach provides a clear way to communicate data quality levels to users as well as providing a framework for adjusting country or international level data quality.

Stress factors - subjective assessment and adjustment

When countries, regardless of their size, population or level of development provide statistical data and are evaluated according to various quality concepts, it is not possible to assess quality across countries for a specific domain (i.e. at point q_2). Some countries have greater difficulty in collecting and processing specific data series than others and this should be taken into consideration in the assessment and adjustments made. For example, a developing country that is large geographically and has a high population is less likely to be able to perform an agricultural census as often as a smaller more developed country. It would therefore be hard to justify a comparison of these two countries using the same standard data quality criteria.

Key stress factors have been identified at various data quality points (*see Figure 1*). The data quality indicators at points q_1, q_2 and q_3 are functions of the stress factors at those stages of the statistical process. For example, specific stress factors need to be considered when evaluating the output data quality indicators of an international statistical agency with over 200 member countries, a limited budget, a small staff and a large number of statistical domains. Countries also have similar stress factors related to their level of development and the environment in which they operate.

Adjusting Data Quality Indicators for Stress

When summary data quality measures are available and "benchmark quality levels" are established, it is possible to adjust for Stress. The Stress Factors for a national statistical office can be calculated as a function of: Population (number), Geographical size (sq kms), Education (completion rate), Income (gross national income), Environment (e.g. recent natural disasters, conflict) and a summary Stress Factor score produced. The Stress Factor score would then be applied in a similar fashion to a GDP deflator/inflator.

The following example illustrates the adjustment of the original Country Data Quality Performance Index from *Table 2* for Stress. If Country A has been identified as having the current benchmark level of data quality then all other countries will be compared to Country A.

Table 3. Country Stress Factors

Country Stress Factors	Year					
	2000	2001	2002	2003	2004	2005
A	1.00	1.00	1.00	1.00	1.00	1.00
B	1.10	1.15	1.15	1.15	1.20	1.20
C	1.10	1.10	1.05	1.10	1.30	1.35

In *Table 3,* three observations can be made. First, the Stress Factor for the benchmark Quality Country (Country A) did not change between 2000 and 2005. Second, the Stress Factor for Country B increased from 90 in 2000 to 105 in 2005. Third, the Stress Factor for Country C, increased from 90 in 2000 to 120 in 2005. When the Stress Factors is applied as deflators/inflators, "Adjusted Quality Performance of each Country" can be calculated (*Table 4*).

Table 4. Adjusted Country Data Quality Index

Country Quality	Year 2002=100					
	2000	2001	2002	2003	2004	2005
A	93	95	100	103	103	105
B	77	86	86	92	96	96
C	77	85	79	77	85	88

The Adjusted Country Quality Performance is calculated by amending Country Quality Performance of Countries B and C by their specific Stress Factors. Thus the Country Quality of Country B in 2000 (70) is multiplied by the Stress Factor for that Country for 2000 (1.1) with the adjusted value being 77. The Adjusted Country Quality Performance of Country B is closer to the Optimum Quality Country (Country A) than previously thought and the quality performance of Country C is improving rather than worsening. These examples illustrate the application of Quality Stress Factors and provide us with a method to better understand the Data Quality performance of countries and international organizations in producing statistics under conditions that are not ideal.

References

Eurostat, *Assessment of quality in statistics,* UN/ECE Work Session on Statistical Metadata. Washington DC, United States of America, 28-30 November 2000.

Food and Agricultural Organization of the United Nations, Statistics Division, *The Agricultural Bulletin Board on Data Collection, Dissemination and Quality of Statistics Project (ABCDQ),* http://faostat.fao.org/abcdq/abcdq.jspx

Food and Agricultural Organization of the United Nations, *FAO Statistical Data Quality Framework: A multi-layered approach to monitoring and assessment,* Conference on Data Quality for International Organizations, Committee for the Coordination of Statistical Activities Wiesbaden/Germany, May 2004.

Food and Agricultural Organization of the United Nations, *FAOSTAT- Redevelopment of the FAO Statistical Database System,* Asia and Pacific Commission on Agricultural Statistics (APCAS), New Delhi, Republic of India, September 2004.

Food and Agricultural Organization of the United Nations, *The Development of CountrySTAT as a Vehicle for Organizing National Agricultural Sector Data,* Asia and Pacific Commission on Agricultural Statistics (APCAS), New Delhi, Republic of India, September 2004.

Food and Agricultural Organization of the United Nations, *Integrating Data Quality Indicators Into the FAO Statistical System,* Quality in Statistics, EUROSTAT, Seventh meeting, Luxembourg, 23-24 May 2005.

International Monetary Fund and Eurostat, *Data Quality: A comparison of IMF's data quality assessment framework (DQAF) and Eurostat's quality definition,* January. 2004.

United Nations, Statistics Division, *Handbook of statistical organization*. (third edition). Studies in Methods. Series F. No. 88. ST/ESA/STAT/SER.F/88. New York. 2003.

Annex 1: Quality indicators and frameworks of selected International Organizations

Eurostat

Quality aspects	Indicators
Relevance (Completeness merged with relevance)	R1. User satisfaction index R2. Number of publications disseminated and/or accesses to databases CP1. Rate of available statistics
Accuracy	A1. Coefficient of variation (CV) A2. Unit response rates A3. Item response rates A4. Editing rates and ratios A5. Imputation rates and ratios A6. Over-coverage and misclassification error rates A7. Average size of revisions
Timeliness and Punctuality	T1. Punctuality of time schedule of effective publication T2. Average time between the end of reference period and the date of the first results T3. Average time between the end of reference period and the date of the final results
Accessibility and Clarity	AC1. Number and types of means used for disseminating statistics
Comparability (refers to cross country)	C1. Number and proportion of differences in concepts or/and measurement from the European norms C2. Number and length of comparable time series C3. Asymmetries for statistics mirror flows
Coherence (comparability over time or over datasets)	CH1. Number and proportion of sets of statistics that satisfies the requirements for the main secondary use

IMF (DQAF)

Source: http://dsbb.imf.org/vgn/images/pdfs/dqrs_Genframework.pdf

Quality dimensions & elements	Indicators
0 Prerequisites of quality	
0.1 Legal and institutional environment	0.1.1 The responsibility for collecting. processing. and disseminating statistics is clearly specified 0.1.2 Data sharing and coordination among data producing agencies are adequate 0.1.3 Respondents' data are to be kept confidential and used for statistical purposes only 0.1.4 Statistical reporting is ensured through legal mandate and/or measures to encourage response
0.2 Resources	0.2.1 Staff. financial. and computing resources are commensurate with statistical programs 0.2.2 Measures to ensure efficient use of resources are implemented
0.3 Relevance	0.3.1 The relevance and practical utility of existing statistics in meeting users' needs are monitored.
0.4 Other quality management	0.4.1 Processes are in place to focus on quality. 0.4.2 Processes are in place to monitor the quality of the collection processing and dissemination of statistics 0.4.3 Processes are in place to deal with quality considerations including tradeoffs within quality. and to guide planning for existing and emerging needs
1. Assurance of integrity	
1.1 Professionalism	1.1.1 Statistics are compiled on an impartial basis 1.1.2 Choices of sources and statistical techniques are informed solely by statistical considerations 1.1.3 The appropriate statistical entity is entitled to comment on erroneous interpretation and misuse of statistics
1.2 Transparency	1.2.1 The terms and conditions under which statistics are collected. processed. and disseminated are available to the public 1.2.2 Internal governmental access to statistics prior to their release is publicly identified 1.2.3 Products of statistical agencies/units are clearly identified as such 1.2.4 Advance notice is given of major changes in methodology. source data. and statistical techniques
1.3 Ethical standards	1.3.1 Guidelines for staff behaviour are in place and are well known

Quality dimensions & elements	Indicators
	to the staff.
2. Methodological soundness	
2.1 Concepts and definitions	2.1.1 The overall structure in terms of concepts and definitions follows internationally accepted standards. guidelines. or good practices
2.2 Scope	2.2.1 The scope is broadly consistent with internationally accepted standards. guidelines. or good practices
2.3 Classification / Sectorization	2.3.1 Classification / sectorization systems used are broadly consistent with internationally accepted standards. guidelines. or good practices
2.4 Basis for recording	2.4.1 Market prices are used to value flows and stocks 2.4.2 Recording is done on an accrual basis 2.4.3 Grossing/netting procedures are broadly consistent with internationally accepted standards. guidelines. or good practices
3. Accuracy and reliability	
3.1 Source data	3.1.1 Source data are obtained from comprehensive data collection programs that take into account country-specific conditions 3.1.2 Source data are consistent with the definitions. scope classifications. valuation. and time of recording required 3.1.3 Source data are timely
3.2 Assessment of source data	3.2.1 Source data – including censuses. sample surveys and administrative records – are routinely assessed. e.g.. for coverage sample error. response error. and non-sampling error; the results of the assessments are monitored and guide statistical processes
3.3 Statistical techniques	3.3.1 Data compilation employs sound statistical techniques to adjust data sources 3.3.2 Other statistical procedures (e.g.. data adjustments and transformations. and statistical analysis) employ sound statistical techniques
3.4 Assessment and validation of intermediate data and statistical outputs	3.4.1 Main intermediate data are validated against other information where applicable 3.4.2 Statistical discrepancies in intermediate data are assessed and investigated 3.4.3 Statistical discrepancies and other potential indicators of problems in statistical outputs are investigated
3.5 Revision studies	3.5.1 Studies and analyses of revisions are carried out routinely and used to inform statistical processes

Quality dimensions & elements	Indicators
4. Serviceability	
4.1 Periodicity and timeliness	4.1.1 Periodicity follows dissemination standards 4.1.2 Timeliness follows dissemination standards
4.2 Consistency	4.2.1 Statistics are consistent within the dataset (e.g.. accounting identities observed) 4.2.2 Statistics are consistent or reconcilable over a reasonable period of time 4.2.3 Statistics are consistent or reconcilable with those obtained through other data sources and/or statistical frameworks
4.3 Revision policy and practice	4.3.1 Revisions follow a regular. well-established and transparent schedule 4.3.2 Preliminary data are clearly identified 4.3.3 Studies and analyses of revisions are made public
5. Accessibility	
5.1 Data accessibility	5.1.1 Statistics are presented in a way that facilitates proper interpretation and meaningful comparisons (layout and clarity of text. tables. and charts) 5.1.2 Dissemination media and format are adequate 5.1.3 Statistics are released on a pre-announced schedule 5.1.4 Statistics are made available to all users at the same time 5.1.5 Statistics not routinely disseminated are made available upon request
5.2 Metadata accessibility	5.2.1 Documentation on concepts. scope. classifications. basis of recording. data sources. and statistical techniques is available. and differences from internationally accepted standards. guidelines or good practices are annotated 5.2.2 Levels of detail are adapted to the needs of the intended audience
5.3 Assistance to users	5.3.1 Contact person for each subject field is publicized 5.3.2 Catalogs of publications. documents. and other services including information on any charges. are widely available

OECD Quality Framework

Source: http://www.oecd.org/document/43/0.2340.en_2825_293564_21571947_1_1_1_1.00.html

Quality dimensions	Definition
Relevance	The relevance of data products is a qualitative assessment of the value contributed by these data. Value is characterized by the degree to which the data serves to address the purposes for which they are sought by users.
Accuracy	The accuracy of data products is the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure. OECD context: Accuracy of the data is largely determined by the accuracy of the data received from the contributing organizations. On the other hand. the activities carried out by OECD can influence the overall accuracy.
Credibility	The credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer. i.e.. the brand image. Credibility is determined in part by the integrity of the production process. OECD context: publishing bad quality data received from countries affects the overall credibility of the OECD. Furthermore. once agreement between the OECD and countries has been reached on collection of specified data. the data subsequently collected cannot be withdrawn in response to political pressure.
Timeliness	The timeliness of data products reflects the length of time between their availability and the event or phenomenon they describe. but considered in the context of the time period that permits the information to be of value and still acted upon.
Accessibility	The accessibility of data products reflects how readily the data can be located and accessed from within OECD data holdings. OECD context: internal and external users might have quite different perceptions of accessibility because of the differences in access methods.
Interpretability	The interpretability of data products reflects the ease with which the user may understand and properly use and analyse the data. The range of different users leads to such considerations as metadata presentation in layers of increasing detail. The adequacy of the definitions of concepts. variables. and terminology. information describing the limitations of the data etc.
Coherence	Reflects the degree to which data are logically connected and mutually consistent. Distinction can be made between coherence within a dataset. coherence across datasets. coherence over time and coherence across countries. Ensuring coherence across countries is one of the major sources of value added provided by the OECD.
(Cost-efficiency)	The cost-efficiency with which a product is produced is a measure of the costs and provider burden relative to the output. Whilst the OECD does not regard cost-efficiency as a dimension of quality. it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions.