A Suggested Framework for the Quality of Big Data

Deliverables of the UNECE Big Data Quality Task Team

December, 2014

Contents

1. Executive Summary

This report is a deliverable from the UNECE/HLG project*, The Role of Big Data in the Modernisation of Statistical Production*, and specifically describes the findings of the Big Data Quality Task Team. The team, which comprised representatives from several national statistical offices, was asked to investigate the implications of Big Data for the quality of official statistics, and to develop a preliminary framework for national statistical offices to conceptualise Big Data quality.

The team concluded that extensions to existing statistical data quality frameworks were needed in order to encompass the quality of Big Data. A preliminary framework was developed building on dimensions and concepts from existing statistical data quality frameworks. The Big Data Quality framework developed here provides a structured view of quality at three phases of the business process:

- Input – acquisition, or pre-acquisition analysis of the data;
- Throughput –transformation, manipulation and analysis of the data;
- Output – the reporting of quality with statistical outputs derived from big data sources.

The framework is using a hierarchical structure composed of three hyperdimensions with quality dimensions nested within each hyperdimension. The three hyperdimensions are the *source*, the *metadata* and the *data*. The concept of hyperdimensions has been borrowed from the administrative data quality framework developed by Statistics Netherlands.

The hyperdimension *source* relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated. The hyperdimension *metadata* refers to information available to describe the concepts, the contents of the file data set, and the processes applied to it. The hyperdimension *data* relates to the quality of the data itself.

In addition, three general principles are proposed when evaluating Big Data quality:

- Fitness for use (is the data source appropriate for the purpose)
- Generic and flexible (a quality framework such as the one proposed here should be broad and applicable over a wide variety of situations)
- Effort versus gain (is the effort involved in obtaining and analysing the data source worth the benefits gained from the data source)

At the input phase of the business process, a National Statistical Office (NSO) should engage in a detailed quality evaluation of a Big Data source both before acquiring the data (this known as the 'discovery' component of the input phase), and after (this is the 'acquisition component).

In addition to dimensions commonly used to assess statistical output, the task team recommended the use of new dimensions for an NSO to employ, including privacy and confidentiality (a thorough assessment of whether the data meets privacy requirements of the NSO), complexity (the degree to which the data is hierarchical, nested, and comprises multiple standards), completeness (of metadata) and linkability (the ease with which the data can be linked with other data).

For the throughput phase of the business process, three principles of processing are proposed:

1. System Independence: The result of processing the data should be independent of the hardware and software systems used to process it;
2. Steady states: that the data be processed through a series of stable versions that can be referenced by future processes and by multiple parts of the organisation ;
3. Application of Quality gates: that the NSO employ quality gates as a quality control business process.

For the output phase of the business process, the Team used the Australian Bureau of Statistics Data Quality Framework as a starting point.

The following quality dimensions were proposed: *Institutional/Business Environment, Privacy and Security, Complexity, Completeness, Usability, Time Factors, Accuracy, Coherence,* and *Validity.* Factors and possible indicators for each of these dimensions are presented for the input and output phases of the business process.

Important sub-dimensions of some of these dimensions were identified. For accuracy, the identified sub-dimension was *selectivity*. While this issue is not unique to Big Data, it was felt that problems surrounding selectivity and representativeness are more common when dealing with Big Data than when dealing with more traditional sources of NSO data such as surveys, and required special attention.

Similarly, in the dimension of coherence, the sub-dimension of *linkability* was identified, which involves an evaluation of the ease with which the data source can be integrated with other data sources. Another subdimension of coherence was *consistency* – the extent to which the data adheres to internal and external standards.

Rather than focusing quality efforts only on statistical outputs of Big Data, NSOs need a series of quality framework and quality principles that apply across the business process. The UNECE Quality Task Team has recommended some principles as well as dimensions that would be useful for an NSO to evaluate Big Data sources and products.

2. Background

The April 2013 meeting of the UNECE Expert Group on the Management of Statistical Information Systems (MSIS) identified Big Data as a key challenge for official statistics, and called for the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to focus on the topic in its plans for future work.[1] As a consequence, this project, The Role of Big Data in the Modernisation of Statistical Production, was undertaken in 2014.

The goals of the project were as follows:

- To identify, examine and provide guidance for statistical organizations to identify the main possibilities offered by Big Data and to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry
- To demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts
- To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

The project comprised four 'task teams', addressing different aspects of Big Data issues relevant for official statistics: the Privacy Task Team, the Partnerships Task Team, the Sandbox Task Team and the Quality Task Team.

This report is the outcome of the Big Data Quality Task Team, comprised of representatives from several national statistical offices to investigate the implications of Big Data for the quality of official statistics, and to develop a preliminary framework for national statistical offices to conceptualise Big Data quality.

The participants and their National Statistical Office affiliations were:

- David Dufty, Australia
- Hélène Bérard and Laurie Reedman, Canada
- Sylvie Lefranc, France
- Marina Signore, Italy
- Juan Munoz and Enrique Ordaz, Mexico
- Jacek Maślankowski and Dominik Rozkrut, Poland
- Boro Nikic, Slovenia
- Ronald Jansen and Karoly Kovacs, UNSD

The UNECE provided the secretariat function. This role was performed by Peter Struijs and Matjaz Jug. In addition to the secretariat role, Peter Struijs also participated as a team member and made significant contributions to the structure and content of the deliverables. David Dufty not only

[1] Final project proposal: The Role of Big Data in the Modernisation of Statistical Production. UNECE, November 2013. http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production

contributed significantly to the contents of the Quality Framework, but also chaired the meetings and did the crucial work of editing subsequent versions of this document. There was also some collaboration with the Statistical Network (SN) and material developed by the SN and this group were shared.

The deliverables for the Quality Task Team were specified by the UNECE as follows:

> **1. Quality framework(s) for Big Data:** This work should start with an assessment of current quality frameworks for official statistics from survey and administrative sources. These frameworks typically identify around 6 or 7 dimensions of quality. Some of these may be relevant for Big Data, whilst others may not. Similarly, additional dimensions may be needed.
>
> The typology of Big Data sources identifies 3 broad categories - Social networks (human sourced information), Traditional business systems (process-mediated data) and Automated systems (machine-generated data). It is possible that a separate quality framework, with different dimensions, could be needed for each category.
>
> A separate quality framework may also be needed for outputs based on Big Data sources. However, in many cases, outputs will be derived from multiple sources, including Big Data, so some synthesis of source-related quality frameworks will also be needed.
>
> **2. Testing the framework(s):** The quality framework(s) produced under point 1 above should be tested on Big Data sources and outputs, preferably based on previous experiences of using Big Data in statistical organisations. The outcome will be a validation of the framework(s) and/or proposals for enhancements.
>
> **3. Indicators and associated metadata requirements:** Applying quality frameworks in practice will require a number of quality indicators. This task identifies the relevant indicators for the framework(s) above, as well as the metadata required to populate these indicators.

The primary audience for this document is composed of statistical organisations producing official statistics; however this framework should also be useful for any Big Data user from the private sector, academia and the public in general. This document need not be read linearly: for instance a reader who is familiar with data quality dimensions and is more interested in practical suggestions for quality evaluation might skip the descriptions of the dimensions and move straight to the sections relating to input, throughput or output.

The work presented here is not the final answer on Big Data Quality. While we believe that it will be of use to National Statistical Offices, we envisage that these concepts and principles will be a useful starting point for future developments in the area of statistical data quality in relation to Big Data.

3. Introduction

What the Big Data phenomenon means for national statistical offices (NSOs) is that there now is an expanding range of data sources that have the potential to be used for official statistics. One of the challenges for NSOs is assessing the quality of such data sources, and the quality of statistics produced from them. A description of the different Big Data sources is provided in Annex 1 "Secondary Sources".

In assessing or describing the quality of statistical data or outputs, NSOs often make use of data quality frameworks. There are several frameworks currently in use. Examples include the IMF data quality framework, the European Statistical System quality framework, the Statistics Canada data quality framework and the Australian Bureau of Statistics data quality framework. These frameworks have many commonalities: they all have a dimension approach to quality, and many of the dimensions either overlap or correspond across different frameworks.

Some data quality frameworks were developed primarily with survey data in mind and may not be suitable in the wider context of Big Data. However, some NSOs make use of administrative data rather than surveys and their frameworks and protocols reflect this (see for example the Statistics Netherlands administrative data quality framework[2], and the Australian Bureau of Statistics paper on administrative data quality[3]). There is also currently an international effort to develop a quality framework for administrative data led by the Statistical Network[4] that is mapped to the following business processes namely: the input or acquisition phase, throughput or processing phase, and output or dissemination phase.

The Task Team reviewed several existing quality frameworks for official statistics with respect to their applicability to Big Data. This review included frameworks such as those produced by Statistics Sweden, Statistics Canada, the Australian Bureau of Statistics, the EU Code of Practice, the checklists produced as work package 2 (input phase) and work package 6 (output phase) of the ESSNet project on the Use of Administrative and Accounts Data for Business Statistics, and the work done by the Statistical Network.

It was noted that some existing frameworks tend to be output focused. This reflects the high degree of control that official statistics agencies have previously had over the creation and initial processing of data used in statistical products. Frameworks that deal with administrative data have a broader scope and are able to cope with a wider variety of data sources and data types. However Big Data goes beyond even the scope of administrative data, and the team concluded that the application of either traditional data quality frameworks or those designed for administrative data would be an inadequate response to Big Data.

[22]Daas, P., Ossen, S., Vis-Visschers, R., &Arends-Toth, J. (2009),*Checklist for the Quality evaluation of Administrative Data Source*s. Statistics Netherlands, The Hague/Heerlen

[3]ABS 2011, Information Paper: Quality Management of Statistical Outputs Produced from Administrative Data, March 2011, cat. no. 1522.0, ABS, Canberra.

[4] For more information on the Statistical Network see **http://www1.unece.org/stat/platform/display/statnet/The+Statistical+Network**

The team concluded that, due to the complex nature of Big Data, extensions to existing statistical data quality frameworks were needed. In order to encompass the quality of Big Data, a preliminary framework was developed building on dimensions and concepts from existing statistical data quality frameworks.

4. Principles

The considerations in developing this framework were:

- to keep it as consistent as possible with existing quality frameworks
- to capture as much of the diverse range of possible quality issues associated with Big Data; and
- to keep it as simple as possible.

Aligned with these considerations the following principles are proposed:

Fitness for use

The concept of 'fitness for use' is a central principle: the quality of any particular data source or product can only be evaluated in light of its intended use. This principle, used in the application of existing statistical data quality frameworks, is just as relevant in the evaluation of the quality of big data sources and the statistical products derived from them.

Generic and flexible

The intent is to produce a generic and flexible quality framework that can be applied at each phase (input, throughput and output) using the 3 hyperdimensions with a set of relevant quality dimensions. When a given quality dimension is relevant for different hyperdimensions and phases, the associated quality indicators are developed to reflect the different quality assessment done under each context.

Efforts versus gain

An overall assessment of the fitness for use of the data can only be performed once all quality dimensions and **relevant indicators** have been assessed.

In order to balance the effort involved in assessing data quality and the added value of using the data, a set of minimum requirements to be met should be identified.

5. Structure of a Quality Framework for Big Data

The Big Data Quality Framework (BDQF) provides a structured view of quality assessment for the three phases of the business process that lead to the production of statistical outputs. The three phases of the BDQF align closely with the stages of the General Statistical Business Process Model (GSBPM). They are:

- *Input* – when the data is acquired, or in the process of being acquired (collect stage);
- *Throughput* – any point in the business process in which data is transformed, analysed or manipulated. This might also be referred to as 'process quality' (process and analyse stages);
- *Output* – the assessment and reporting of quality with statistical outputs derived from big data sources (evaluate and disseminate stage).

The Big Data Quality Framework (BDQF) uses a hierarchical structure composed of three hyperdimensions, with quality dimensions nested within each hyperdimension. The concept of hyperdimensions has been borrowed from the administrative data quality framework developed by Statistics Netherlands.[5] Note that it is also being incorporated into the administrative data framework currently under development by the Statistical Network.

The three hyperdimensions are:

- *Source:* relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated.
- *Metadata:* relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated.
- *Data;* relates to the quality of the data itself.

For a given quality dimension, different quality indicators may be developed depending of the phases (input, throughout and output). For example, the complexity of the file structure can affect potentially its readability, the integration with other data, and the level of data that can be disseminated which will be evaluated respectively at the input, throughput and output phases (using different quality indicators).

An overall description of the different quality dimensions is given below. However, the factors to consider and potential quality indicators for each dimension are given separately under the input, throughput and output sections.

[5]Daas, P., Ossen, S., Vis-Visschers, R., &Arends-Toth, J. (2009), *Checklist for the Quality evaluation of Administrative Data Source*s. Statistics Netherlands, The Hague/Heerlen

6. Big Data Quality Dimensions

Institutional/Business Environment

This dimension refers to the institutional and organisational factors which may have a significant influence on the effectiveness and credibility of the agency producing the data. Consideration of the institutional environment associated with a statistical product is important as it enables an assessment of the surrounding context, which may influence the validity, reliability or appropriateness of the product.

In the process of data acquisition or data discovery, it is the institution that is providing the data that should be scrutinised. At later parts of the business process the quality of the institutional environment of the NSO itself is more relevant.

It may be the case that the organisation or entity providing the data is not considered to be likely to be able to provide data over an extended period of time. In this case, rather than focusing on the longevity and stability of organisation itself, the question becomes that of whether comparable data will be available in the future, from similar organisations or sources.

If a third party is providing data to an NSO, the transparency around data collection and analysis is a factor to be considered in a quality evaluation. If the provenance of the data is not well understood then the NSO needs to question whether they can guarantee the quality of statistics derived from that data.

Privacy and Security

This dimension refers to the institutional and organisational factors, for both the data provider Organisation and the NSO, which may have a significant influence on the intended use of the data given legal limitations, organisational restrictions, and confidentiality and privacy concerns.

In many standard data quality frameworks, for traditional survey data, issues surrounding privacy and security are sometimes included within the dimension of “institutional environment.” For big data sources, –privacy and security becomes a more prominent and complex issue. The Task Team therefore felt that privacy and security should be given more prominence in quality evaluations by allocating it to its own dimension, rather than being considered as part of the broader picture of the quality of the institutional environment.

Metadata should include enough information to infer that privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes are absolutely guaranteed.

Physical, technological and organisational provisions should be in place to protect the security and integrity of statistical databases.

A key issue in data acquisition from third parties is that of consent, and whether that consent is active or passive. Consent might be obtained via an “agreement to terms” or in more explicit ways. An NSO acquiring data should pay careful attention to whether consent is given and whether this

accords with the NSO's guidelines, policies, and regulatory environment. Additionally, perceived lack of consent due to data acquisition may undermine public trust.

Complexity

In short, complexity refers to the lack of simplicity and uniformity in the data.

Complexity of the data source can be assessed in four different aspects: data structure, data format, data itself and hierarchies used in the data. The way in which the data is received, read, validated, processed and stored by the NSO depends on the characteristics of the data.

Complexity of the data structure means that there can exist various relations between data, including complex keys in tables, that makes it difficult to integrate various data tables. Depending on the type of the data source it may be related to the number and size of unstructured files that must be integrated to create the unified data source.

Complexity of the data format, complexity can be measured by checking what kind of data standards were used to store the data (e.g., spatial data mapped in various formats).

Complexity of the data source can be regarded as complex when there is lack of information on the code lists used in the data or the code lists used in the data are not integrated in one data source (e.g., to code gender different labelling are used).

Hierarchical complexity; this reflects the extent of hierarchies and nested structures in the data. Depending on the requirements of statisticians it can be difficult to drill-down to a specific level of the data.

For output, complexity is only relevant if the output is in unit record form that may reflect the complexity of the input, or if there is a need to report on how complexity in input data has been dealt with in the previous stages and if it caused any limitations to the outputs.

Completeness

Completeness is the extent to which metadata are available for a proper understanding and use of data. It refers to the exhaustiveness of the descriptions available for the input data (i.e. covering all the required aspects mentioned in the hyperdimensions Source and Data as well as the level of detail of descriptions. It includes descriptions of objects (populations, units, and events), variables and reference times as well as applied procedures for data treatment and quality measures or qualitative assessment of input data quality.

Access to the data file record layout can be considered a minimal requirement for use. For complex data files, the usefulness of the data will depend heavily on the type of information available about the file structure, coding and classification variables. In some cases, the absence of relevant information may drastically limit the potential use of the data if this information cannot be deducted or evaluated from the data itself.

The completeness and interpretability assessment of the documentation should cover the steps necessary for the evaluation at the input stage but also for the subsequent stages (throughput and output). Essentially, evaluation of this dimension will help determine if quality information is available about the data that may be critical to the NSO at any subsequent stage.

It should be noted that if the data is being acquired from third parties, a thorough assessment of completeness is dependent on information provided by those third parties.

Usability

The usability of a dataset is the extent to which the NSO will be able to work with and use the data without the employment of specialised resources or place significant burden on existing resources; and the ease with which it can be integrated with existing systems and standards.

One of the features of Big Data is an increasing diversity in data types, structures and formats. A typical NSO is structured around the receipt and processing of long-standing data sources, most notably survey and census data, and more recently some kinds of administrative data. An NSO will be able to more easily make use of an incoming dataset that is compatible with existing systems and expertise.

For new, varied sources of data and new methods of processing and analysing those data sources, new systems and infrastructure will need to be developed. While there is a strong incentive for NSO's to develop new capabilities, the extent to which this will need to be done for any particular data source is an important consideration in considering the quality of the data source.

If new expertise or infrastructure is needed, the question then arises about the extent to which the improved capability will be transferrable to other data sources. Development of new capabilities provides benefits to the NSO but also comes at a cost; this trade-off is something that should be taken into account when considering a data source.

Time Factors

This dimension captures the timeliness of the data and its periodicity. Timeliness and frequency are the two important quality aspects of Big Data, and in fact in many cases are the added value provided by Big Data, that will be traded against quality in other areas. At this time, the promise of Big Data consists of timeliness, frequency, granularity and geo-spatial coverage. Therefore, in order to have a business case for the use of Big Data, these two quality aspects have to be high and deliver value above that provided by existing data sources.

Furthermore, data from external sources may have additional time-related problems such as delays between the reference period (the point in time that the data refers to) and the time of collection; in some situations, either the reference period or the time of collection may not be known with certainty.

Furthermore, data that is collected periodically has the additional benefit of allowing the option of benchmarking and time series, although again, this is very much dependent on the purpose.

Accuracy

The *accuracy* of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (*e.g.*, coverage, sampling, nonresponse, response)[6].A total survey error approach is desirable when analyzing the accuracy of a potential dataset in regard to statistical analysis.[7]

Selectivity

A key concern with many Big Data sources is the selectivity, (or conversely, the representativeness) of the dataset. A dataset that is highly unrepresentative may nonetheless be useable for some purposes but inadequate for others. Related to this issue is the whether there exists the ability to calibrate the dataset or perform external validity checks using reference datasets.

As explained by Buelens et al (2014) : "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective."

Selectivity indicators developed for survey data can usually be used to measure how the information available on the Big Data Source differs from the information for the in-scope population.

For example, we can compare how in-scope units included in the Big Data differ from in-scope units missing from the Big Data. To assess the difference it is useful to consider the use of covariates, or variables that contain information that allows to determine the "profile" of the units (for example, geographic location, size, age, etc.) to create domains of interest. It is within these domains that comparisons should be made for "outcome" or study variables of interest (for example, energy consumption, hours worked, etc.). Note that the covariates chosen to create the domains should be related to the study variables being compared.

Coherence

Coherence is the extent to which the dataset follows standard conventions, is internally consistent, is consistent over time, and is consistent with other data sources.

We considered two subdimensions to be particularly important in regard to Big Data: linkability and consistency.

Linkability

Linkability is the ease with which the data can be linked or merged with other relevant datasets.

A dataset may be acquired for use in conjunction with an existing dataset, or multiple datasets may be linked with each other. In these situations, the ability of the data to be accurately linked with

[6] Statistics Canada Quality Assurance Framework (2002)

[7] Weisberg, H. F. (2005*). The Total Survey Error Approach: A Guide to the New Science of Survey Research,* University of Chicago Press

other datasets of interest is of critical importance. As the practice of data linking increases in frequency and importance it is necessary to include a quality analysis of linkability in situations where linking is likely.

Linkage with other data sources is a planned activity in order to fully exploit the Big Data. For these cases special consideration should be given to the quality of the variables used to perform the linkage. In the data hyper dimension, we are looking at the success of record linkage in terms of percentages of linked and unlinked records.

The linkage process is a time consuming process that often involves pre-processing of the files to standardize the linking variables. Where possible, exiting methods already in used in the NSO to carry out this processes should be considered.

Consistency

This refers to the extent to which the dataset complies with standard definitions and is consistent over time.

A dataset that has changing internal standards is one that has the potential to create quality concerns over time, as outputs that seem comparable at a superficial level may have changing underlying definitions or structures.

Data that uses idiosyncratic or non-standard definitions for standard terms and concepts may also create potential quality issues.

Validity

The validity of a dataset is the extent to which it measures what the user is attempting to measure. The concept of validity is a long-standing one in methodology. In terms of data quality it has previously been subsumed under the concept of 'coherence.' However as NSOs move out of design based surveys and towards a greater range of outputs and products, the notion of validity may require more attention.

Aggregate statistics have historically been derived through survey methods, which are well understood, using sample statistics. With the introduction of new, diverse sources of data sampling theory may not be an appropriate metric for evaluating the utility of derived metrics. Furthermore, in design-based sample surveys there is a straightforward link between the measurement and the underlying concept; the sample is intended to represent the population, with (for the most part) well understood properties of error and variability.

Big Data provides the opportunity for a more diverse range of statistical products, that may estimate population parameters or characteristics from data sources that are not at face value representative of those characteristics (such as the spectra in a satellite image, or the volume of phone data at a particular point in time). In contrast to traditional surveys, the direct subject of measurement (e.g., the spectra of a satellite image pixel) may be of secondary interest in its own right, while its ability to assist in the inference of a population characteristic is of interest. To put this another way, in a standard social survey the population parameters of interest are measured directly, whereas a

statistical output derived from big data may be of interest only to the extent that it can predict a population parameter.

Consider a hypothetical example in which spectral frequencies of satellite data may be indicative of crop yields or agricultural activity. While accurately measuring the spectra is certainly a necessary part of the process, it is not analogous to a traditional measure of accuracy such as relative standard error, for example. That concept is better captured by the relationship between the metric derived from spectral data and the variable of interest such as agricultural activity.

In such cases it is important that the quality of the outputs be ensured even where traditional sample survey methods may be either inapplicable or insufficient for a complete quality assessment. There needs to be some kind of assessment of the validity of the output: an assurance that it is measuring what it is claimed to be measuring.

NSO's already engage in this kind of quality assessment for some non-survey outputs such as national indexes (e.g., consumer price indexes). In this case the validity of the index might be described through a transparent, rigorous methodology, including the logic that underlies the methodology and the conceptual reasons why it is believed to capture the concept in question. This might be thought of as 'model validity.' Other ways of validating an output include correlation with similar population metrics or related population characteristics.

Accessibility and Clarity

Accessibility refers to how easy is to access information (metadata and data) by the users.

Clarity refers to the availability of clear, unambiguous descriptions accompanying data, ranging from definitions (e.g. definitions of units, variables) to descriptions of data treatment (e.g. procedures, techniques,…) to provision of quality measures (e.g. number of item "corrected",…).

Relevance

This dimension refers to how well the statistical product or release meets the needs of users in terms of the concept(s) measured, and the population(s) represented. Consideration of the relevance associated with a statistical product is important as it enables an assessment of whether the product addresses the issues most important to policy-makers, researchers and to the broader users' community.

There are extensive resources available on the topic of evaluating relevance from a statistical data quality perspective. These materials are equally applicable in the Big Data context.

7. Input Quality

The input stage includes activities related to the initial acquisition of the data. In some cases, the data will be already available whereas in other cases only information about the data will be available. It is envisaged that this framework can be used for both assessing the suitability of acquiring a dataset, and assessing the quality of the dataset once acquired.

Within the input stage, the *source* and *metadata* dimensions deal with aspects of the data source that may be discoverable prior to actually obtaining the data. For this reason, *source* and *metadata* hyperdimensions provide the opportunity for assessment before the data is obtained. Such an assessment is sometimes referred to as the 'discovery' phase, and can be undertaken, for example, to decide the fitness of the data for its intended use, determine what uses the data might be put to, or how much effort should be expended in acquiring it. Quality dimensions in the *data* hyperdimension, on the other hand, can only be assessed once the data is actually acquired.

In some cases, the NSO requirements and the intended use of the data will be known prior to the start of the data quality evaluation. In other cases, potential use of the data will be discovered as the data is explored further. For both cases, at the onset, or as the evaluation progresses, the intended use should be clearly documented.

The table below gives an overview of the framework for the input phase of the business process. A more complete description of the factors to consider and potential quality indicators follows. Note that some quality dimensions are evaluated both under the hyperdimensions *metadata* and *data* at the input phase.

Table 1. Dimensional Structure of the Input Phase of the Big Data Quality Framework

Hyperdimension	Quality Dimension	Factors to consider
Source	Institutional/Business Environment	Sustainability of the entity-data provider Reliability status Transparency and interpretability
	Privacy and Security	Legislation Data Keeper vs. Data provider Restrictions Perception
Metadata	Complexity	Technical constraints Whether structured or unstructured Readability Presence of hierarchies and nesting
	Completeness	Whether the metadata is available, interpretable and complete
	Usability	Resources required to import and analyse Risk analysis
	Time-related factors	Timeliness Periodicity Changes through time
	Linkability	Presence and quality of linking variables Linking level
	Coherence - consistency	standardisation Metadata available for key variables (classification variables, construct being measured)
	Validity	Transparency of methods and processes Soundness of methods and processes
Data	Accuracy and selectivity	Total survey error approach Reference datasets Selectivity
	Linkability	Quality of linking variables
	Coherence - consistency	Coherence between metadata description and observed data values
	Validity	Coherence between processes and methods and observed data values

Institutional/Business Environment

Factors to consider:

1. Sustainability through time: factors (internal and external) which could affect the sustainability of the data provider's data in relation to the NSO requirements. If the data provider will not be available, will similar data providers or comparable data sources be available in the future?
2. Reliability status: status of the data provided in terms of overall reliability of the data

1. Transparency and Interpretability: Availability of relevant information about the data provider; transparency about data collection and processing.

Possible indicators

1. What is your estimate of the overall risk that the data provider will not meet the quality requirements of the NSO?
2. What is the risk that the BDS will not be available from the data provider in the future? If it will not, will there be comparable data sources in the future?
3. How relevant are the data, if they would be available for only a short period of time?
4. How long do these data need to be available to be relevant?
5. Is it likely that we can replace these data with similar (or next generation) data, once the data source or technology becomes obsolete?

Privacy and Security

Factors to consider

2. Legislation: Identify the various acts or laws related to the production of the data, its maintenance, its access to the data, and potential secondary use or the data.
3. Restrictions: Identify potential privacy, security and confidentiality restrictions that would limit the use of the data.
4. Perception: The intended use of the data may be perceived negatively from the various stakeholders. Specific actions which may require substantial funding may be needed to mitigate these risks.

Possible Indicators

1. Does the NSO have clear legal authority to obtain the data?
2. Are there legal limitations or restrictions on the use to which the data can be put?
3. Are the data provider and the NSO willing to enter negotiations to solve any legal issues, if necessary?
4. Was the data collected in accordance with relevant privacy laws?
5. Do the NSO's own confidentiality policies limit the utility of data?
6. Are stakeholders (private sector, public, others) likely to react negatively given the intended use of the data by the NSO?

7. Will there be a need to carry out privacy assessment exercises and public consultations in relation to using this data and its potential impact on the NSO reputation and credibility?

Complexity

Factors to consider

1. Technical constraints: identify the tools and technical requirements to receive, read, process and store the file.
2. Structure: How structured the data is and how easy it will be to work with that structure
3. Readability of the data: depending of the file structure, some data may not be accessible
4. Hierarchies and nesting: Whether the data is characterised by hierarchies and nestedness.

Possible indicators

1. Structure: how easy would it be to render the data source into a useable structure (i.e... one record per unit of observation,)?
2. Format: Is the data source in standard format (e.g., XLS, XML)? How many different formats were used in the data source? How easy would it be to render the data source variables into a useable format (i.e... parsing, grooming, coding, treatment of outliers or missing values)?
3. Data: how many different standards were used in the data source (e.g., ISO-3166 to describe countries)? Is there any non-standard code lists used in the data source that are not unified? How many different code lists were used in the data source?
4. Hierarchies: is there a hierarchical relationship between records or variables?
5. Structure: How many different files or tables are in the data source?

Completeness

Factors to consider

1. Information quality: Whether the metadata is available, interpretable and complete for the following:
    a. Processes that led to the collection of the data
    b. Processes related to the treatment of the data
    c. Description of the data itself

Possible Indicators

1. Qualitative assessment (e.g. score for completeness of metadata for input phase: 0 description missing, 1 description unsufficient, 2 description complete)
2. In case of missing/incomplete descriptions what are the consequences/drawbacks for data usability?
3. Are the population units defined clearly?
4. Are the variables defined clearly?
5. Qualitative assessment of completeness and clarity of metadata

6. In case of unclear/ambiguous descriptions what are the consequences/drawbacks for data usability?

Usability

Factors to consider:

1. Additional resources: What would be the skills set required to process and store this data? Would additional investments be required for training?
2. Risk analysis: consider the potential pitfalls and gains for the NSO if considerable investments are required in order to use the data

Possible Indicators

1. Will the NSO need to acquire new skills to use and analyse the data?
2. How much resourcing will cleaning and processing the dataset require?
3. How big is the data set?
4. Data transmission: Are special arrangements for data transmission required, and if so, can the NSO meet those requirements?
5. IT requirements: What would be the hardware and software requirements to process and store this data? Will there be in a need to develop a specific IT infrastructure?

Time Factors

Factors to consider

1. Timeliness: More recent data is, in many cases, higher quality data, although the value of recency can vary wildly depending on the data and the use to which it is put.
1. Reference period: Time between collection of data and reference period to which the data refers. If at the time of data capture, the data was referring to past events (for example, upload or manual entry into a database of historical transactions), then there is a reduction in the quality of the data, for two reasons: first, additional delays make the data further out of date; and second, delays in capture introduce additional possibilities for error.
2. Changes through time: Data that is collected periodically has the additional benefit of allowing the option of benchmarking and time series, although again, this is very much dependent on the purpose. Coherence in concepts and methods must be considered when using historical data.

Possible Indicators

2. Time between receipt of data and when the data was collected;
3. When was the data collected? What is the reference period of the data?
4. Whether data is collected and available periodically. Recurring data provides the opportunity for time series.
5. Could changes in concepts or methods limit the potential use of historical data?

Coherence - Linkability

Factors to consider

1. Linking variables: in many cases, linkage with other data is a planned activity in order to fully exploit the Big Data source. For these cases, special considerations should be given to the quality of the variables used to perform the linkage.
2. Level of linking: depending on the intended use, the level at which linkage must take place can be more or less precise. For example, when linkage is needed at the geographical level, the quality of the linkage may vary depending on the geographical detail but may be more strict for specific geographical administrative boundaries

Possible indicators

1. Are potential linking variables present on the file that could be used for data integration with other data files?
2. Calculate the percentage of units linked and not linked in both the Big Data (BD) and other data sources. The indicator is the percentage of units linked unambiguously (strong link) / percentage of units linked with a soft link (linking requirements were relaxed in order to link more units)

Coherence - Consistency

Factors to consider

1. Standardized concepts: The use of standards for key variables.
2. Coherence with metadata: range of values find in the data can help determine the coherence in the metadata and the actual data

Possible indicators

1. How do you rate the variables capturing the constructs that are of interest?
2. Are the definitions used aligned with NSO standards?
3. Do the anomalies in the data indicate important errors that would limit the potential use?

Validity

Factors to consider

1. Transparency of methods and processes: Availability and soundness of information about methods and processes to produce the statistical outputs. The methods and processes should include all major steps that led to the production of the data including adjustments that are made to the data.
2. Soundness of methods: determine if the data supports the methodologies described for its production by producing descriptive data analysis.

Possible Indicators

1. Is the metadata available sufficient to assess the soundness of the methods used?
2. Are there critical flaws in the processes that would limit potential use of the data?

Accuracy and Selectivity

Factors to consider

1. Total Survey Error approach to analysing accuracy; including in particular (but not restricted to), over-coverage, under-coverage, selectivity, missing data (non-observation and non-response), adjustments made to the data and the presence of anomalies.
2. Reference datasets: Many analyses require the use of reference data sets due to respondent related error or instrument generated error.
3. Selectivity: Imperfections in coverage.

Possible indicators

1. If a reference data set is available, assess coverage error. For example, measures of distance between Big Data population and the target population (e.g. Kolmogorov-Smirnov Index, Index of dissimilarity)
2. Does the file contain duplicates?
3. Are the data values within the acceptable range?
4. Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by Big Data source.
5. Assessment of spatial distribution of measurement instrument and of periodicity of observations
6. Selectivity: Derive R-index for unit composition[8]

[8] R-index: Representative Index, an indicator that estimates the selectivity of the data missing by using information available in other sources (Schouten and Cobben 2007, Cobben and Schouten 2008).

7. Throughput Quality

"Throughput" refers to all the intermediate stages between acquisition of the data and dissemination. In GSBPM terms, it encapsulates the process and analyse stages of the business process.

Given the enormous range of possible types of data and analyses on various types of data, it is beyond the scope of this document to provide a taxonomy of processes or the kinds of quality indicators that might be used with them.

Instead, we can describe some general principles for the quality of data in the throughput stage.

1. System independence

Transformations and analysis should proceed according to theoretical principles and not be dependent on the system that is performing them. For example, the residuals of a regression should be the same regardless of the analytical system performing the regression.

2. Steady States

A steady state is a version of a dataset that has met certain quality criteria. This dataset can then be further processed, analysed, transferred, and merged with other production lines. The use of steady states is preferable to the common practice of continuous improvement of quality within the business cycle, for several reasons:

- It means that people and teams working on a product have a clear understanding of the quality of the data at various points in time;
- It provides a common reference point for all those who use the data; this is particularly important with more complex production networks;
- It brings with it all the benefits of good versioning practices.

Steady state datasets should be internally accessible and clearly identified as a steady state with appropriate metadata such as the nature of the steady state dataset (e.g., where in the business process it is created, the level of quality it has, and what it is to be used for, and when it was created.

(See Struijs et al, 2013 for a discussion of steady states in the context of statistical production systems)[9]

3. Quality Gates

A quality gate is a checkpoint in the business process at which the quality of the data is explicitly assessed. Important features of quality gates are that the measures used to assess quality are decided in advance, and the location of the gate is decided in advance.

Quality gates are a supplement to standard 'quality checks' that are typically undertaken throughout processing, as they create structure and rigour around the quality assurance process, and provide a

[9]Struijs, Peter, et al. "Redesign of Statistics Production within an Architectural Framework: The Dutch Experience." *Journal of Official Statistics* 29.1 (2013): 49-71.

clear output as to whether the data has achieved acceptable levels of quality at a particular point in time.

Features of a quality gate are the following:

1. Placement: where in the business process the quality gate occurs;
2. Measures: the measurements used to asses quality are decided in advance;
3. Roles: who is responsible for the quality gate;
4. Tolerance: the thresholds for acceptable quality (according to the measures) are also decided in advance;
5. Actions: what to do if the quality gate fails;
6. Evaluation: monitoring and assessment of the quality gate itself.

(Further information on quality gates can be found in the article*, Quality Management of Statistical Processes Using Quality Gates,* ABS, 2010). [10]

Example: An administrative dataset

As an example of how these principles can be applied, consider a large administrative dataset of medical records that is received by an NSO.

Upon receipt of the dataset, the NSO converts the data to an internal standard, removes unwanted fields and runs an algorithm for duplicate records.

Quality gates: the NSO puts into place a quality gate at the end of the transformation process; the quality gate includes the following measures: The number of records corresponds to the number of unique records in the original dataset; that no fields have been corrupted; and that all fields conform to the metadata specifications for the file.

The quality measures are designed with dimensional system in mind, paying attention to the relevant dimensions of coherence and accuracy.

The transformed dataset undergoes a cleaning process at the end of which it is saved as a steady state. This steady state dataset is then made available internally for several production lines.

The cleaning process occurs according to pre-defined, transparent rules and does not occur according to an opaque algorithm embodied and implemented by unreadable source code.

Example: Social media

Big Data supports collecting data from large unstructured data sources. Social media is one of the most promising unstructured data source for Big Data. Based on the social media portals such as Twitter or Facebook, new statistical information may be retrieved, especially in terms of ICT skills. However there is a risk that data gathered from social media may be: wrong, noisy, irrelevant, inadequate or redundant.

[10]Quality Management of Statistical Processes Using Quality Gates, Dec 2010, cat.no. 1540.0, ABS, Canberra. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1540.0

The following presumptions may be identified, when working with Big Data, concerning the social media portals as the data source:

- there is a noise in the data;
- data is not clean;
- data may be ambiguous.

To assess the quality of social media during processing phase, traditional quality dimension may be used such as: accessibility, accuracy, comprehensiveness and coherence. These dimensions may be named as core dimensions. There are also contextual dimensions, such as value added, representatives, timelines and completeness.

The first step is to ensure that access to the social media source is legal and the data may be retrieved from the website directly. This means that quality dimension named accessibility is crucial in this step. Next is to prepare algorithms to create key-value pairs. This is often based on regular expressions. When preparing such algorithms a major issue is to ensure that the quality dimension timeliness will be in effect.

One of the most important things when working with unstructured data sources such as social media is to ensure that during mapping key-value pairs any ambiguous observations are identified. Other issues concern: mappings failure, records misplaced, de-duplication (to eliminate duplicates), data integration.

In that phase any correlations in the data must also be checked and reported.

During data and metadata processing the results must be oriented to output which means that the quality dimension relevance is the key issue of the Big Data processing. It is also important to ensure that data and metadata may be reused in the future. To accomplish that task, the results of Big Data processing must be stored in traditional data set.

Future Directions in throughput quality

The approach to throughput quality in statistical production described here is a general, broad one.

A more detailed treatment of quality issues in this part of the business process needs to take into account the wide and expanding range of data sources and the uses to which they can be put. This involves an expansion of statistical quality control to a wider range of data sources and data types.

However it also will involve a treatment of different kinds of statistical products and outputs. A lot of current data collection activity has the end goal of providing insights into population characteristics that are inferred from the data but not directly measured, such as demand, growth, sentiment, consumer confidence, and so on. Given an increasingly diverse range of data sources to choose from, the future of statistics may lie in obtaining correlates of these parameters of interest without sample survey techniques. Instead, parameters of interest may be modelled using data sources without the use of intermediate population estimates.

In such a statistical landscape, it is not sufficient to simply expand our understanding of data quality to a wider range of data formats and sources. Rather, more general conceptions of data quality must be developed that encapsulate new techniques as well as old, and that are flexible enough to be applicable to the full range of outputs and products that are possible from Big Data.

8. Output Quality

An output quality framework should be applicable to reporting, dissemination and transparency. It is information about the quality of the statistical product that a consumer of that product would ideally have. In terms of the General Statistical Business Process Model, 'output' is equivalent to the *disseminate* and *evaluate* stages of the GSBPM.

The following table provides a summary overview. New dimensions are bolded.

Table 2. Dimensional Structure of the Output Phase of the Big Data Quality Framework

Hyperdimension	*Quality dimensions*	*Factors to consider*
Source	Institutional/business environment	Type of data source Arrangements and quality assurance Type of use of the BD source
	Privacy and security	Legislation Actual limitations in the use of data Actions undertaken
Metadata	**Complexity**	Data treatment; output limitations
	Accessibility and Clarity	Data and metadata accessibility Clear definitions, explanations Conformity to standards
	Relevance	Extent to which the data measures the concepts meant to be measured for its intended uses
Data	Accuracy and **Selectivity**	Traditional measures of accuracy Selectivity
	Validity	Correlation with similar metrics Utility Conceptual soundness
	Coherence - linkability	
	Coherence - Consistency	
	Time-related factors	Timeliness Periodicity

Output quality dimensions tend to be more holistic than the dimensions of input or throughput quality. As a result, specific indicators for Big Data output quality are not always relevant or useful. It should also be noted that the factors and indicators described here are intended to have a Big Data focus. Quality indicators developed for statistical outputs can be applied as well to Big Data and have not been reported in this framework for the sake of simplicity.

Institutional/Business Environment

Factors to consider

1. The nature of the input Big Data source (e.g., social media, satellite data; see below for further elaboration of this);
2. The arrangements under which the data was transferred to the NSO
3. What quality assurance processes were applied, if any, to the incoming data
4. The role that the data played in the final output product (e.g., whether it was used for benchmarking, imputation, etc.)

Possible Indicators

1. What institutions contributed to the data, and under what arrangements?

Privacy and Security

Factors to consider

1. Legislation related to the production of the data, its maintenance and access.
2. Restrictions (privacy, security, confidentiality) limiting the use of the data, if any
3. Actions taken to mitigate potential negative perceptions on the use of data from stakeholders

Possible indicators

1. Does the NSO have clear legal authority to obtain the data?
2. Was the data collected in accordance with relevant privacy laws?

Complexity

Factors to consider

1. Data treatment: how complexity of the input data has been dealt with during the input and throughput stages with regard to data structure, format, data and hierarchies
2. Actual limitations to the use of statistical outputs caused by complexity of Big Data used as input, if any

Possible indicators

1. Uniform and consistent metadata standards and classifications across the dataset
2. Presence or absence of nested hierarchies

Accessibility and Clarity

Factors to consider

1. Data and metadata accessibility
2. Extent to which data are accompanied by clear, unambiguous definitions, explanations and quality indicators
3. Conformity to metadata standards

Possible Indicators

1. Cost of access
2. Presence of supporting documentation

Relevance

Factors to consider

1. Whether the data measures the concepts meant to be measured for its intended uses. In assessing data relevance, key aspects include scope and coverage; reference period; geographic detail; use of standard classifications; types of estimates available and any other relevant issue or limitation in the use of the data.

Accuracy and Selectivity

Factors to consider

1. Traditional measures of statistical accuracy such as standard error, bias, etc.
2. Selectivity issues

Possible indicators

1. Measures of distance between Big Data population and the target population (e.g. Kolmogorov-Smirnov Index, Index of dissimilarity)
2. Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by Big Data source
3. Assessment of spatial distribution of measurement instrument and of periodicity of observations

Validity

Factors to consider

1. Convergent validity: how well the metric aligns with other, similar metrics
2. Conceptual utility: the extent to which the metric is able to provide insight into real-world phenomena

3. Methodological validity: the extent to which the methods underlying the metric are transparent and theoretically sound

Possible indicators

1. Correlations between big-data-derived metric and population parameters
2. Robust and transparent methodology underlying derivations
3. "Predictive power:" the ability to predict movements or trends in variables of interest

Time factors

Factors to consider

1. Timeliness
2. Periodicity

Possible Indicators

1. Time between receipt of data and when the data was collected; a longer time lag is considered to be an indicator of lower quality.
2. Time between collection of data and reference period to which the data refers. If at the time of data capture, the data was referring to past events (for example, upload or manual entry into a database of historical transactions), then there is a reduction in the quality of the data, for two reasons: first, additional delays make the data further out of date; and second, delays in capture introduce additional possibilities for error.
3. Whether data is collected and available periodically. Recurring data provides the opportunity for time series.

Appendix I: Secondary Sources

Given that big data products involve the synthesis and transformation of data from a wide variety of data sources, the nature of these sources becomes a more prominent consideration than for statistical products that are generated entirely in-house.

Where the NSO obtains data from an external organisation, transparency is needed about the nature of the acquisition. A complete evaluation of the institutional credibility and trustworthiness of third parties is not a feasible requirement in reporting quality; but an output should provide sufficient information for the user of the data to make an informed decision about the impact of the third party on the quality of the data.

Factors to consider in output quality reporting:

- The nature of the original source data (e.g., social media, satellite data; see below for further elaboration of this);
- The arrangements under which the data was transferred to the NSO
- What quality assurance processes were applied, if any, to the incoming data
- The role that the data played in the final output product (e.g., whether it was used for benchmarking, imputation, etc.)

Types of sources

The following is a list of possible sources for third party data acquisitions that an NSO might make. This is not intended to be an exhaustive list.

1.Sensors/meters and activity records from electronic devices

This kind of information is produced in real-time, the number and periodicity of observations of the observations will be variable, sometimes it will depend of time intervals, whereas on other occasions it will be a record of the occurrence of some event (e.g., a car passing a roadside camera) and on still other occasions it will depend of manual manipulation such as meter checking. Sensor accuracy is a critical component of the quality of this type of data.

2. Social interactions

Electronically captured social interactions, such as those from online social networks, provide a rich source of quantitative and qualitative data. The quantitative aspects are easier to capture, using techniques such as counting the number of observations grouped by geographical or temporal characteristics. Analysis of qualitative data such as unstructured language include techniques such as sentiment analysis and trend topics analysis. These rely on algorithms which should be subject to scrutiny and analysis from a data quality point of view.

3. Business transactions

Data produced as a result of business activities can be recorded in structured or unstructured databases. A common problem with the analysis of structured databases in this context is the large

volume of information. For example, in recording sales data, a large retail chain can produce up to thousands of records a second.

An additional complication is that this kind of data is not always produced in formats that can be directly stored in relational databases. Electronic invoices are an example of this problem. The invoice has a structure, but to store it in a relational database, a large amount of processing needs to be applied. If the data is not in plain text (for example, if it is in formats such as picture, PDF, Excel, etc.), additional processing is required. It is conceivable that the rate of production of the data outpaces the rate of processing. Strategies to overcome this include not storing the data on a relational database, discarding some observations, and using parallel processing. The quality of information produced from business transactions is highly dependent on the ability to get representative observations and to process them.

4. Unstructured documents

Unstructured documents may be statically or dynamically produced, and include electronic files such as Internet pages, video and audio files, and pdf files. While their content may be informative, that content can be difficult to extract. Techniques in this environment include text mining and pattern recognition. Data quality will be dependent on the capacity to extract and correctly interpret all the representative information from the documents.

5. Broadcasting

Broadcasting data refers to video and audio produced in real time. Obtaining statistical data from this kind of electronic data is not currently feasible, as it is highly complex and requires very large levels of computational power. However breakthroughs in processing techniques along with increases in computational power may make this possible in the future.