Guide to reporting on administrative data quality

New Zealand Government

Citation
Statistics New Zealand (2016). *Guide to reporting on administrative data quality.* Retrieved from www.stats.govt.nz

ISBN 978-0-908350-29-2 (online)

Published in February 2016 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600
www.stats.govt.nz

Contents

Access the next two appendixes from 'Available files' on this document's webpage.

List of tables and figures

List of tables

List of figures

1 Introduction to assessing administrative data quality

Purpose

This guide presents a framework for understanding how well different datasets meet their intended purpose, including their strengths and limitations. It also explains how to determine what effects these strengths and limitations may have on the quality of a statistical output that uses administrative data, survey data, or a combination of the two.

Quality assessments carried out using this error framework should help answer the questions arising from Statistics New Zealand's push to be an 'administrative data first' organisation: how do we decide which administrative data should be used for which purposes, and how can we be sure that direct surveying is not necessary?

Measuring data quality

No statistical dataset perfectly measures exactly what we want it to. At present we cannot provide a single generic measure to summarise data quality, but this guide's error framework can produce a comprehensive list of the strong and weak points of datasets and outputs. Instead of judging a dataset as 'good' or 'bad', the framework identifies the strengths and weaknesses of a dataset in an objective way, with reference to its original purpose. Such analysis can guide design decisions and ensure we collect the right amount of data to produce fit-for-purpose outputs.

The framework facilitates reusing both existing data and previous quality assessments.

Structure of this guide

The first part of the error framework focuses on how well a dataset meets its original, intended purpose – useful information when wanting to investigate whether the data can meet other needs. We hope the framework provides a common language for talking about data quality issues, and is a valuable decision-making resource for the organisation.

The second part addresses problems that can arise when combining datasets from different sources (eg transforming raw variables to match statistical needs and identifying and creating statistical units from integrated datasets). The outcome of such an assessment is useful to test different design options, or to identify quality risks that need to be mitigated or checked over time to ensure the consistency of the resulting statistics.

This guide also supports measures and indicators to quantify key aspects and concerns of data quality in a detailed way. While these measures do not cover all situations, they give ideas for more detailed or technically complex measures that could be developed for a specific output.

See Quality indicators for phase 1 [and 2] errors in the ‘Available files’ on this webpage.

The framework in this document cannot solve quality problems on its own, but it will highlight aspects of the datasets most in need of further work – so investigations can focus on the most crucial quality issues.

Contents of this guide

The main parts of this guide are:

- An explanation of the error framework. This details the framework and describes how to apply it to different datasets and outputs.
- A practical example (see section 5) to help explain how to use the framework. The tables in this example provide useful templates for other assessments.
- Our current plans for implementing the error framework, and future work.
- The metadata information template. This Excel spreadsheet (see ‘Available files’) to use to capture the key information about datasets being assessed.
- Detailed lists of quality measures. The two quality indicators files (see ‘Available files’) list indicators and measures categorised by error type. Select and use the ones most relevant or useful to assess a specific dataset.

A glossary at the end of this guide defines the terms we use in this guide. Use this alongside the other documents, and also as a guide for which terms to use when writing up a quality assessment or report on work done using the error framework.

2 Overview of making a quality assessment

This section summarises how to use the error framework and where to begin a quality assessment.

You can apply the error framework to almost any dataset and/or statistical output. Some of the thinking involved can be complex and difficult, so it's important to consider your quality assessment's scope – start with these questions:

- What are your aims? Some possibilities include:
  - evaluating the quality of a new administrative dataset as it arrives at Statistics NZ
  - developing better measures of quality for an existing output
  - understanding the impact of design choices on quality when using administrative data
  - getting a better understanding of the trade-offs between more administrative data use and final output quality.
- What are the relevant data sources for your output (whether planned or existing), including both survey and administrative data? Which are the most important for your purposes?
- Has the dataset been used before within Statistics NZ, and is there any earlier work that could save you time? Look at what is available on Colectica and check relevant internal documents and databases to see if meta-information templates or other studies have already been completed for the dataset.
- What are the variables in the different data sources? Which are the most important for your intended data use?
- What population do the relevant datasets cover? Is the basic unit people, businesses, or something else?
- How are the dataset's raw variables combined or transformed to produce your final data?
- How are the dataset's basic units converted into the statistical units in the final data?
- What are the main quality problems you know of or guess might be relevant to your purpose, based on your understanding of the original data?

Stages of quality assessment

This section lays out the main steps to carry out using the error framework so your time spent on quality assessment is as effective as possible.

It starts with the most important and generally useful aspects of the framework and works down to the details. Your aim should be to produce a quality assessment that gives enough information to make design and other decisions confidently.

Metadata information template

The metadata information template encourages thinking about the key aspects of quality in an organised way. It is also a convenient way to record a standard set of information – to compare different datasets. See 'Available files' for this template.

The first step of a quality assessment is to briefly answer the main questions in the template. The most important are:

- **General information**: Items 1.1–1.6 including source agency, purpose of collection, summary of variables, and time span of the data.
- **Population**: The target population, admin population, and reporting units. The items relating to coverage might not be possible to answer with a quick assessment but note anything you do know.
- **Variables**: A short description of key variables. As work progresses, record the target concepts for the variables under investigation as they become known.
- **Collection**: The timing/delay information and method of collection are important and should be easy to find out and record.

Note: Colectica may have much of this information for datasets already used at Statistics NZ. All items are crucial to a sound understanding of a dataset's quality and the issues that might arise from using it for a different purpose. For example, understanding the original purpose of the data collection can guide you to which variables might be of higher quality than others, and to the likely coverage of the data.

Record any useful information for other questions but ignore any non-relevant boxes in the template. If you uncover relevant information later in the assessment, then add it – ideally the meta-information template for a given dataset should be improved and expanded as different people in Statistics NZ find out more about it.

Phase 1 of the error framework summarised

The error framework, explained in detail in the next section, has two phases. Phase 1 deals with datasets in their raw state – as they look when originally produced. The key questions in this stage of the assessment are "what information did the creators of this dataset want to capture?" and "how well does the final dataset capture this ideal information?"

The framework is split into two sides:

- 'measurement', which deals with the variables in the data
- 'representation', which deals with the respondents or other reporting units, generically labelled 'objects'.

The phase 1 assessment should give you a detailed understanding of the issues that arise during the original data creation processes, and how they affect data quality for the original purpose.

You need to define or describe each boxed term in the figure 1 phase 1 error framework diagram (eg target concept, harmonized measure) for the datasets being assessed. Use the general information in the meta-information template to do this. Think about each step from the point of view of the original data producers and what their goals were when they created the dataset.

Once you've defined the terms, categorise known data quality issues or strengths according to the error source (the ovals on figure 1). This shows exactly where any quality issues arise.

Phase 2 of the error framework summarised

Phase 2 of the error framework aims to determine how well a given combination of datasets meets a statistical need.

Firstly, describe or define the boxed terms in figure 2 (phase 2 error framework). This requires knowledge of the output design and the processes that transform the source data into the final statistical output. Apply the phase 2 framework to your proposed designs to help decide on the design, or to existing designs – to understand the current strengths and weaknesses of an output and where improvements might be possible.

Throughout your phase 2 assessment, the target concept and population (see figure 2) are the ideal statistical information you would like to have. You must identify a clear statistical need to carry out an effective assessment.

Using the results

Record your phase 1 and phase 2 descriptions in a simple table.

See Case study: the Quarterly Building Activity Survey (tables 1–6) for a guide to the appropriate level of detail needed.

Once you've completed these first steps, conduct a more detailed investigation of specific error sources for any causing a problem in the final statistical output. However, a comprehensive evaluation of every source of error in a complex output could be time-consuming. Time spent on these tasks should reflect what is needed to meet each project or assessment's goals – focus on areas where you can make useful mitigations or improvements.

The lists of quality measures and indicators (see 'Available files' on this webpage) for phase 1 and phase 2 may be useful at this stage, to help understand some ways to measure various aspects of data quality. These lists are not intended to be universally applied, but are meant to give some ideas and potentially prompt thinking about more specialised measures that might be useful for a particular output. Use these measures to help form an objective picture of the quality of a dataset at a particular point in time. They can also be used as ongoing monitoring checks for the output, to ensure that its quality is consistent.

Record your completed quality assessment or template centrally, for others to reuse the results and analysis.

See Central repository of quality assessments section for more details on doing this.

3 Explaining the framework in detail

Purpose of the error framework

Any statistical output contains imperfections or uncertainties. These can arise from choices in methodology, limitations of input data sources, processing problems, or many other sources. The effect on customers will depend on how they use the output – a data issue may be irrelevant to one customer but make the output useless for another.

To fully understand how good a final output is for a given need we need a comprehensive list of its limitations. The error framework gives us a way to categorise and understand the sources of these limitations and how they affect the final output.

How the framework operates

Li-Chun Zhang (2012) developed the error framework. It breaks down the steps between the ideal concepts and population we would like to capture in our dataset and the final unit-record data that we obtain in practice.

Zhang’s framework builds on the Total Survey Error framework developed by Groves et al (2004, figure 2.5). This model examines all possible sources of error in survey data, from design right through to the data’s use in producing statistical outputs.

The framework has two phases – each has separate flows for 'measurement' (relating to target concepts and values obtained from population units) and 'representation' (relating to target sets of units and the objects measurements are obtained from). These are explained in more detail below.

Note: steps in the error framework are not arranged in order of production processing steps or data flows from data receipt to statistical output, as in the Generic Statistical Business Process Model (UNECE, 2013). The framework is trying to capture compromises needed to produce the output; for example, in translating an ideal concept into a question or variable we can measure in a well-defined way. Identifying these compromises and limitations helps to understand the differences between the final data and the perfect data we would wish for.

By using Li-Chun Zhang's framework we can compile a comprehensive list of error sources for a given dataset. Use the quality measures (see ‘Available files’) to try to quantify or monitor each error source. The framework and quality measures assist your decision-making about cost/quality trade-offs when designing new outputs and improving old ones.

Elements of the framework

The error framework separates the 'life cycle' of statistical data into two phases. This division makes it easy to categorise sources of error and understand their causes. The idea is to first evaluate datasets against their original purposes, and then consider how well the combination of datasets making up the final dataset fits the target concept and population of the intended statistical output. This is very important when combining several administrative or survey datasets to produce an output, but it is also useful for single-dataset outputs – it allows us to separate source data issues from the problems caused by trying to reuse the data for a purpose it wasn't designed for.

The framework is also split into two sides, 'measurement (variables)' and 'representation (objects or units)', which are explained below.

Phase 1

Phase 1 allows us to evaluate a single data source against the purpose for which the data was collected. For a survey dataset, this purpose is defined for a statistical target concept and target population. For an administrative dataset, the entries or 'objects' in the dataset might be people or businesses, but they could also be transaction records, or other events of relevance to the collecting agency. At this stage, evaluation is entirely with reference to the dataset itself, and does not depend on what we intend to do with the data.

Phase 2

Phase 2 categorises the difficulties arising from taking variables and objects from source datasets and using them to measure the statistical target concept and population we are interested in. In this phase, we consider what we want to do with the data, and determine how well the source datasets match what we would ideally be measuring.

Dividing assessment into two phases has benefits. Firstly, it separates out the information about the source dataset, which means we can reuse the phase 1 assessments for other possible outputs without repeating a lot of work. This also lets us explain why an administrative dataset can be fit for purpose for one output, but inadequate for another.

Secondly, it makes it easier to identify the real cause of a quality issue and to come up with a solution or mitigation strategy that addresses the error at its source. For example, undercoverage in our final output could have many causes, such as poor quality processing at the source agency, mismatches between how matching variables are defined on different datasets, or overly strict edits in our system. Being able to determine which of these is the true cause is far more valuable than simply knowing there is undercoverage.

Measurement

The measurement side of figures 1 and 2 sets out steps that connect the target concept (ideal information we want about each object) with the final edited values in the dataset. Sources of error on the measurement side include the degree to which the operational measure used captures the target concept, and how many and what kind of errors are introduced by respondent misunderstanding or mistakes.

Example of measurement evaluation: look at taxable income recorded in the Employer Monthly Schedule administrative dataset as a measure of personal income. In phase 1 we see how well the figures in the administrative data meet their administrative purpose, whereas in phase 2 we evaluate the issues the administrative variable has for our ideal statistical variable or concept.

Representation

The representation side looks at the objects or units in the dataset and how well they match the desired target set (note: we use 'set' instead of 'population' because some administrative datasets are based on capturing events or transactions rather than a well-defined population of people or businesses). Ideally every object in the target set has a corresponding object recorded in the data. In phase 1, the focus is on objects, which could be events, transactions, or other entries in an administrative dataset, whereas phase 2 is concerned with units (the final statistical units in the dataset), which may be created artificially – based on a combination of objects from several linked datasets.

The representation side of figures 1 and 2 could be used to evaluate errors arising from combining administrative datasets to create a household register. Coverage problems, timing issues, data matching uncertainties, and problems in actually generating a list of household units are all included in the framework.

Steps for using the framework

The steps we recommend to assess the quality of an output or dataset using this framework are:

1. Determine which datasets are relevant and collect basic information about each, such as the original purpose of the data collection, the set of objects or units in the target population, and definitions of the variables and how they are collected.
2. Use phase 1 of the framework to collate detailed information about each dataset that relates to: processing the variables, rules used during collection, any specific restrictions on the units that make it into the final data, and any other known issues with the dataset. The aim is to define and explain each box in figure 1 (eg accessible set) in detail for each dataset. This includes categorising the known issues into the correct error types.
3. Use the information gathered in step 2 to create a list of known or potential error sources, categorised according the framework.
4. Use the list of measures and indicators (see ‘Available files’) to find ways to quantify or control each important source of error. Also consider the effect of each type of error on the final output for the most important error sources.

Once you’ve completed the phase 1 assessment for each source dataset, complete phase 2 using a similar process. Defining the statistical target population, concepts, and variables very clearly is important – so you can accurately compare the individual datasets assessed with the phase 1 framework to the statistical use for the data.

4 Sources of error in each phase

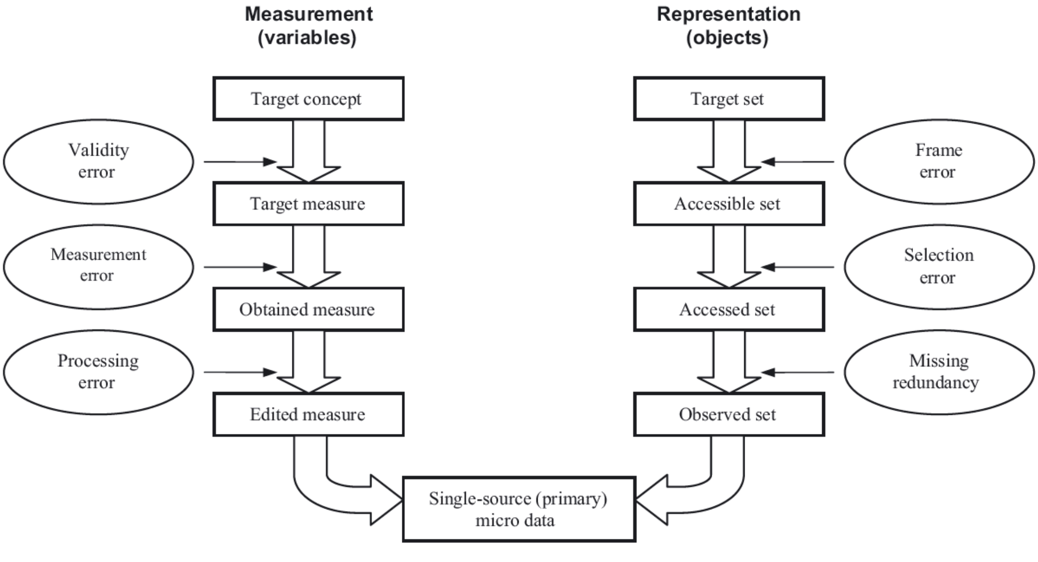
This section illustrates the possible sources of error the framework can identify.

Li-Chun Zhang's 2012 framework built on earlier work by Groves and Bakker (2004), which was more focused on survey-only collections. Figures 1 and 2 are based on the steps that connect the abstract, ideal measurements (or objects) to the final data actually obtained. To apply the error framework, we need to clearly define the boxed terms. Once they're defined, list and categorise issues that arise in each step into the error types (ovals pointing to each transition).

Phase 1 errors

Phase 1 applies to a single dataset in isolation. For a complex statistical output from many different datasets, carry out the phase 1 evaluation separately for each source dataset. The framework can be used for both administrative and survey data. Once we have a comprehensive list of phase 1 errors, we determine the effect of these errors on the final output we want to produce using the phase 2 framework.

Figure 1
Phase 1 error framework showing the different types of error that can arise



In phase 1, define the target concept and target set by the dataset's original purpose, whether it is a stand-alone sample survey or a transactions database held by a retail store. For a traditional sample survey designed to produce a statistical output, phase 1 errors mean the final outputs are not perfect estimates of the true population values – there is always some uncertainty due to sampling error, imputation, non-response, and other issues. For an administrative source, we focus on evaluating how well the source meets the purpose intended by the collecting business or agency.

Below we explain the terms used in figure 1.

Measurement (variables) terms

The measurement side describes the path from an abstract target concept to a final edited value for a concretely defined variable.

Validity error

Measurement begins with the **target concept**, or ‘the ideal information that is sought about an object’. To obtain this information, we must define a variable or measure that can be observed in practice. Validity error indicates misalignment between the ideal target information and the operational **target measure** used to collect it. Typically, administrative variables are collected for a definite purpose and defined in a very concrete way. The error arising in this step refers to the translation from an abstract target concept to a concrete target measure; it doesn’t include errors such as misunderstanding terms on a form.

Measurement error

Once the target measure is defined, we collect actual data values. The values for specific units are the **obtained measures**. When the data is obtained from people responding to a survey or filling out a form for a government agency many errors can occur. People may misremember details or interpret questions differently from what was intended.

Note: for some administrative sources the objects that data is being collected about may not be ‘respondents’ in a traditional sense.

An example: a retail chain might record the values and times of all transactions made in their stores. In this case the object is a transaction whose value is recorded automatically, but measurement error could still occur. For example, a fault in the reporting system that delayed processing of a week’s transactions and ended up recording them as occurring on the day the system was fixed would be a measurement error.

Processing error

The **edited measure** is the final value recorded in the administrative or survey dataset, after any processing, validation, or other checks. These checks might correct errors in the values originally obtained, but can introduce additional errors. For example, in a survey, dividing a response by 10 because it appears a magnitude error was made by a respondent, when in fact the original response was correct.

Representation (objects)

The representation side of the flowchart deals with defining and creating ‘objects’ – the basic elements of the population being measured.

Frame error

The **target set** is similar to the target concept – it is the set of all objects the data producer would ideally have data on. An important distinction between the usual statistical concept of ‘units’ and ‘objects’ in this context is that in some administrative datasets the base units could be records of individual events (eg transactions with customers). Statistically, we may want to create a list by customer that links many transaction events into one statistical unit, but the administrative agency may only care about the events themselves. To avoid confusion, we say that the final dataset after all phase one transformations is organised into ‘objects’ rather than ‘units’.

Frame error refers to the difference between the ideal target set of objects and the **accessible set**, the set from which we can take measurements in theory. These concepts are clarified under ‘Selection error’ below.

Selection error

Many collections have objects in the accessible set that don’t end up in the data. For instance, our accessible set could be all people eligible to vote, but the **accessed set**, the set we actually obtain information about, includes only people who actually registered on the electoral roll. The missing, unregistered people are a source of selection error.

The distinction between frame error and selection error can be confusing, especially when the collection is designed with restrictions already in mind.

An example: crime statistics. If the target set is all crimes committed, and the accessed set is all crimes reported in the Police database, then it is probably best to treat unreported crimes as selection errors, and ‘unreportable’ crimes (crimes that could never be reported even in theory – if there are any) as the frame error.

Another example: a retail chain wants to produce statistics on the transactions across all its stores, but their system can only record purchases using electronic cards. Cash transactions could be said to be ‘inaccessible’ since they will never be in the database – they cause a frame error. However, if a store manager forgets to run the reporting tool for a week, the transactions missing from the dataset due to that mistake will be selection errors: they were accessible, but were not accessed and do not appear in the dataset.

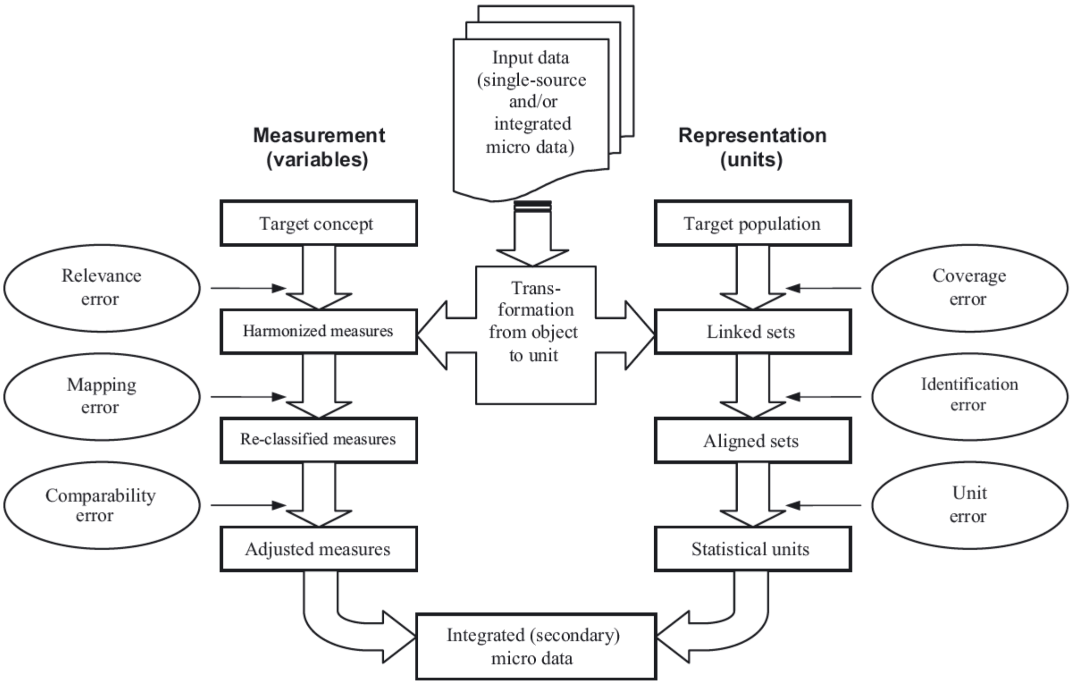
Missing/redundancy error

The observed set comprises objects in the final, verified dataset. Most checks an agency does are likely to remove objects that shouldn’t have been in the selected set to begin with (eg someone trying to enrol to vote who is under 18); these types of errors are selection errors. The incidence of errors where the agency mistakenly rejects or duplicates objects due to their own processing is fairly rare, but this category of error exists so we keep such errors distinct from reporting-type errors.

Phase 2 errors

Phase two of the error framework covers errors arising when existing data is used to produce an output that meets a certain statistical purpose. Often this involves combining different datasets for different parts of the population, or integrating several datasets together. However, phase two can also be valuable when a single administrative dataset is used to produce an output on its own – the process allows us to distinguish between quality problems in the original data and errors resulting from trying to make the data measure something it wasn't intended to. In phase two, the reference points are the statistical population we would ideally access, and the statistical concepts we want to measure for the units in the population.

Figure 2
Phase 2 error framework showing the different types of error that can arise



Measurement in phase 2 is concerned with how to reconcile variables from each source dataset, which may differ from the target concept or from each other. Representation is about creating a set of statistical units from the objects in the original datasets.

Figure 2 indicates possible sources of error in phase 2. Note that errors arising in phase 1 can also propagate through to the final data, and that movement is not necessarily directly related to specific or sequential steps in a statistical process. We need to carefully consider the effect phase 1 errors have on the final data, which depends on the intended statistical purpose.

Below we explain the terms used in figure 2.

Measurement (variables)

Relevance error

The **target concept** in phase 2 is similar to that in phase 1 (the ideal information sought about the statistical units). The **harmonised measures** are the practical measures decided on in designing the statistical output, such as a survey question aligned with a standard classification. In some cases they could be the same measures as in one of the datasets to be combined, but the harmonised measures may also be a standardised statistical measure that does not align perfectly with variables in the original datasets.

As in phase 1, relevance errors are entirely conceptual, and don't arise from actual data or values. Harmonisation can be thought of as "consist[ing] for the greater part of the formulation of decision rules, in which the measurement of a concept is determined as precisely as possible, given the existing information in the data sources" (Bakker, 2010).

Mapping error

We transform measures in the source datasets into harmonised variable values. The values we assign in this process are called **re-classified measures**. Practical difficulties encountered in this stage lead to mapping errors.

An example: our building consents data, where the 'job description' field of the consent must be assigned to a specific code in the building type classification. The job description the builder enters is free text and may be ambiguous or unclear – the resulting reclassified measure may not be the correct one.

Following from Bakker's description of harmonisation above, mapping errors may result from the decision rules chosen, which won't work perfectly in every case. Mapping error also includes 'modelling errors'.

See Understanding errors arising from modelling for more on this important source of error.

Comparability error

Regardless of how the reclassified measures are derived, we may need editing and imputation to obtain consistent outputs. The final values after these processes are our **adjusted measures**. In addition to the usual imputation, for units with missing variables in the source datasets, we may need extra checks to reconcile values that are correct for each individual dataset but disagree with each other for the output measure.

An example: someone loses their job and applies for a benefit just before their employer refiles their employee tax returns. If we link the benefit data with the tax data, the person could be recorded in both, since they were paid taxable income but also registered as unemployed in the reference period. Both datasets are individually correct, but we would need to resolve the inconsistency for our final data (eg by looking at application and filing dates).

Representation

Representation in phase 2 deals with creating a list of statistical units to include in the output data, based on the source data's objects. Here is where the object/unit distinction is most important – the individual datasets may be based on transactions or events we need to connect then place into newly created statistical units that relate to customers, stores, or other entities of interest in the statistical target population.

An example: someone whose hiring was recorded on a register of jobs, but whose dismissal from the job was not recorded. If the jobs register is based on events, failing to record the dismissal is a selection error in the jobs register. In phase 2, if we define a harmonised employment measure that classifies people as employed if they were hired and not dismissed, then we misclassify this person's employment status – this is a mapping error. The distinction between phase 1 and phase 2 allows us to understand complex situations such as this.

Coverage error

The **target population** is fairly familiar from survey statistics – it is the 'set of statistical units that the statistics should cover'. The **linked sets** are the units that are connected across the relevant datasets. Note that these units will not necessarily be the final statistical units of the output. For instance, the target population might be households, but the linked sets could be individuals we link using an address variable across different administrative datasets.

Coverage errors are the differences between the units actually linked in practice and the full set of units we include in the (ideal) target population. These errors arise in several ways. For instance, the datasets themselves may not cover the whole target population, or linking errors may mean we don't identify some members of the linked sets. Measurement errors in the source data can also cause coverage errors.

An example: if the date-of-birth variable on an administrative dataset is not of good quality and we filter on age to select our population, we could end up with undercoverage even though the units aren't missing from the source data.

Working through an example: imagine we want to build a dataset that includes qualifications and income for every person living in New Zealand – to study how these variables are related. Our source datasets are individual Inland Revenue tax records and university enrolment data. The target population would be all people in New Zealand. If we try to link people based on name, date of birth, and address across these two datasets coverage errors could occur, including:

- out-of-date addresses, spelling mistakes in names, or other errors, so we can't link a person's Inland Revenue record with any enrolment data (missed links)
- people who studied overseas, so have a qualification but don't appear in the New Zealand enrolment data (undercoverage)
- people who are linked in the two sets but have moved overseas, so are not actually part of the target population (overcoverage).

Identification error

Depending on the type of units in the linked sets, we may want to create ‘composite units’, which are made up of one or more ‘base units’.

An example: the Quarterly Building Activity Survey, where our target units are construction jobs, but we receive data on individual consent approvals. Usually one consent corresponds to one building job, but some complex jobs file separate consents for different stages of the job. Conversely, some consents can be for two buildings of different types, which we would like to have separate statistical units for. We can consider the **aligned sets** as a table that records the consents relating to each construction job. Failure to recognise a consent as the next stage of a job already in progress, and not recording it as related to the previous consent in the job, would be an identification error.

In more complex cases, different datasets may conflict and we must decide how to resolve this.

An example: we have person-level data linked by a common identifier across several datasets, and want to form groups of people living at the same address. If the different datasets contain different addresses for the same person we may make identification errors when we are forced to decide on a single address for each person.

Unit error

The final statistical units in the output dataset could be created from scratch, without a direct correspondence to any of the units in the source datasets. In the example above about addresses, we may create a dwelling unit that consists of all the people living at each unique address. The conceptual difference between linking errors and unit errors is that we are not just connecting people to a known list of addresses – we are simultaneously determining which addresses should actually be given a dwelling unit and which people should be connected to each dwelling unit.

Understanding errors arising from modelling

The variables in administrative datasets typically differ from the ideal data we would like to use to measure our statistical target concepts. In Li-Chun Zhang’s error framework, he gives examples that involve reclassifying the raw values in an administrative variable, such as a free text ‘job title’ field into an official statistics occupation classification. Any errors arising from this process are mapping errors.

A conceptually similar, but often more complex situation arises when we want to estimate a numerical target variable from one or more administrative variables that don’t precisely capture the information we really want.

An example: using GST returns from businesses to estimate the sales and purchase variables as defined on our subannual business surveys. One way we do this is to calculate the ratio of survey sales to GST sales for the larger units we survey, and use this ratio to estimate sales for small, non-surveyed units (for which we only have GST data).

Many sources of error arise from this kind of modelling. Because they generally occur in the step from harmonised measures to reclassified measures in phase 2 of the framework, they come under mapping errors.

Measuring and minimising these errors is crucial to deciding how to make more use of administrative data, and how much survey data we might still require in an ‘administrative data first’ design.

An example: we need to answer questions such as “for which units does the model perform poorly?”, “how stable are the model parameters and how can we monitor them over time?”, and “how large is the uncertainty in our modelled estimates?”

To help understand modelling errors, we consider two types of error that can arise when we use a statistical model to estimate a target variable:

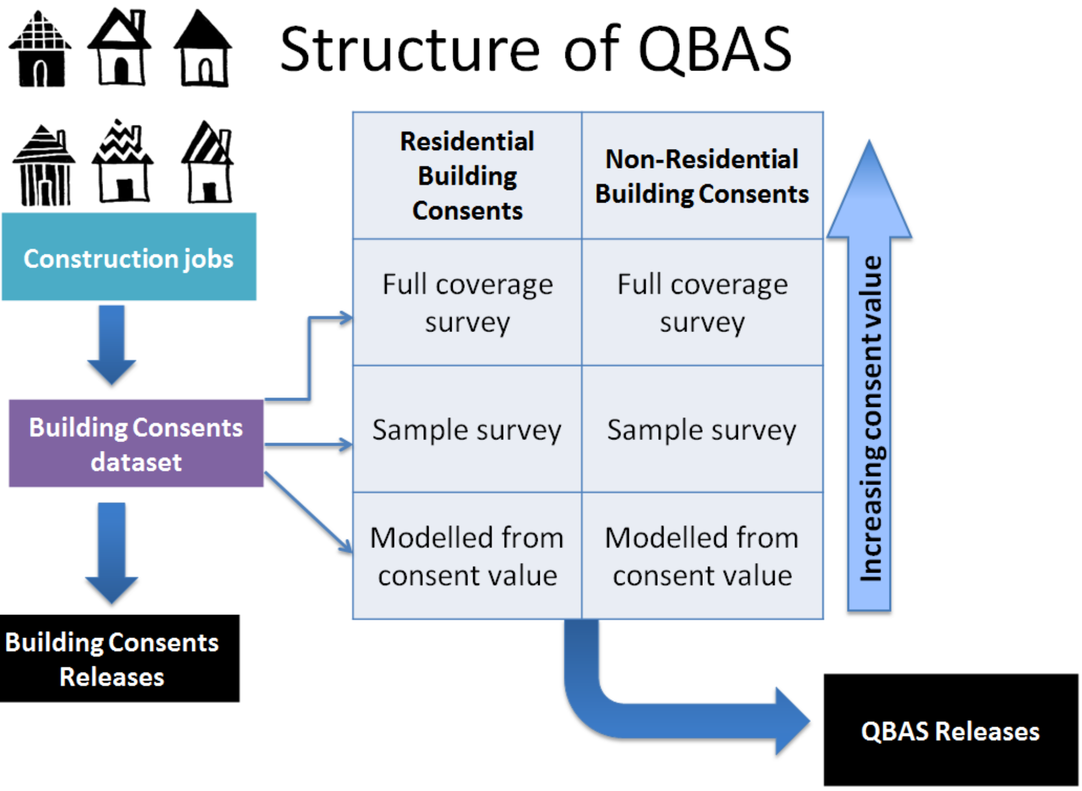
- Model structure error – the model specification chosen may not capture the real relationship between the variables. For example, we might use a simple linear model to predict one variable, using another, but in reality the relationship between these variables is non-linear. Common techniques for assessing this type of error include goodness-of-fit tests and residual plots.
- Parameter uncertainty – when we estimate the values of the parameters in a model, there is always some uncertainty. We need to measure parameter uncertainty and propagate it through to the final results that rely on the model. Techniques such as bootstrapping or Bayesian estimation are often used.

We also consider whether an overall model uncertainty can be determined. If we have more than one possible model, we might combine the results of the different models to provide an overall measure of uncertainty. Bayesian model averaging is one way of doing this.

5 Case study: the Quarterly Building Activity Survey

To explain how to apply the error framework, we use the Quarterly Building Activity Survey (QBAS) as an example. The current QBAS design is a sample survey that uses an administrative dataset (Building Consents) as the sampling frame, the source of some variables, and to aid editing and imputation. Figure 3 shows the QBAS structure.

Figure 3



The building consents output measures the number and value of all consents each month, with breakdowns by area and building type. It serves primarily as an economic indicator of likely activity in the construction sector and the wider economy. It's a full-coverage dataset, and we spend considerable effort on processing and coding the data supplied by all territorial authorities.

Building consents data includes information on the location, consented value, building type, floor area, and some other variables, for each construction job above $5000. This information is published monthly.

QBAS aims to measure the actual value of construction work done in New Zealand each quarter, and is an important component of national accounts series. Based on the monthly building consent datasets, we select a postal sample for the quarterly QBAS. This sample is stratified by residential/non-residential consents and value. We estimate the low-value (under $45,000 for residential and $80,000 for non-residential) jobs with a simple model that assumes the job starts and finishes in the quarter it is consented. We measure the highest-value jobs with a full-coverage survey, and estimate the middle jobs from a sample survey. The survey asks for a single variable – the value of work put in place on the job up to the reference quarter.

QBAS has now moved to this a new design, which makes more use of modelling based on building consents data and significantly reduces the number of construction jobs surveyed. The main change is to model the former sample survey strata from consents data and historic survey data.

See Methodology and classification changes to Value of Building Work Put in Place statistics for details of the changes and what they mean for the published output.

Datasets for phase 1 of the error framework

We combine two unit record datasets to produce the final QBAS output: building consents and the survey data. Each needs a separate assessment. We focus on two variables: work put in place to date (WPIP) from QBAS, and building consent value from the building consents dataset.

First we look at the measurement (variables) side of the phase 1 framework. In tables 1–4 we define the target concept, and the measures for each dataset, and briefly note the most important sources of each type of error. Each table covers one dataset.

Table 1
Measurement side of phase 1 framework for building consents

Measurement (variables)	Building consent value variable	Error type in measurement side	Potential errors arising in building consent value variable
Target concept	Value recorded on each building consent approved by the territorial authority (TA).	Not applicable	Not applicable
Target measure	Measure takes the value from the consent form, as recorded by TA.	Validity error	Alignment between target concept and target measure is very good – the output simply reports building consents data.
Obtained measure	Values that end up in the datasets supplied by each TA.	Measurement error	Values could be wrongly entered on forms (eg $20,000 not $200,000). Rounding could affect responses (eg consent for $285,000 is entered as $300,000). For the building type variable, an example is consenting a building that could be either retail or office space. Consent may say 'retail/office', but the building's true use can't be determined at any finer level. Missing values for consent value (item non-response) are also measurement errors.
Edited measure	We check consent values supplied by TAs. Suspicious or missing values are followed up with TA. Edited measure is the final value after checking and confirmation.	Processing error	Where more than one building type is in a consent, we assign the value of the consent to each building type using a predefined formula – some errors for the target measure will arise, because the exact value of each construction job by type can't be determined.

Table 2
Measurement side of phase 1 framework for Quarterly Building Activity Survey respondent data

Measurement (variables)	QBAS work put in place to date (WPIP) variable	Error types in measurement	Potential errors arising in QBAS WPIP variable
Target concept	Work put in place: the actual dollar value of work done on a construction job in the reference quarter.	Not applicable	Not applicable
Target measure	Measure used is the survey question, "What is the total cost of work put in place on this job, from the start of the job until now" (it also explains what costs to include and exclude).	Validity error	Alignment between question and target is good. WPIP is a well-defined dollar value, so it is relatively easy to create a practical question to measure it. If we know WPIP for each quarter we can easily subtract previous work and obtain the quarterly work put in place. No additional transformations should be necessary.
Obtained measure	Actual responses we receive on questionnaire forms.	Measurement error	Respondents can make mistakes in their estimates, or round off values. They may also not understand the instructions and include costs (eg legal fees) that shouldn't be counted towards WPIP. Category includes item non-response, but QBAS only really asks one question so not much difference between item and unit non-response. Non-response is around 10–15% by value for each category/building type and overall.
Edited measure	From WPIP and previous responses, we derive the work put in place for the reference quarter. QBAS responses have edits (eg checking for magnitude errors), and	Processing error	Errors in imputed values mostly result from the regression imputation assumption that the relationship between WPIP and consent value is the same for all jobs in the imputation cell. Errors in

Measurement (variables)	QBAS work put in place to date (WPIP) variable	Error types in measurement	Potential errors arising in QBAS WPIP variable
	we impute missing values. The edited measure is the final value after this processing is done.		imputation flow to the next quarter, since WPIP depends on the previously reported/imputed values. If we impute too low a value for WPIP one quarter and get a true response next quarter, WPIP based on subtracting the imputed value from the response will be too high.

Note: although the WPIP and consent value variables differ considerably, from the phase 1 perspective they are both valid measures of the intended concept of each dataset.

We also need to compare the representation (objects) side of the phase 1 framework.

Table 3
Representation side of phase 1 framework for building consents

Represent-ation (objects)	Building consents units	Error types in representation	Potential errors arising in building consents
Target set	All building consents issued in NZ with a value greater than $4,999 in specified month.	Not applicable	Not applicable
Accessible set	All building consents actually recorded by territorial authorities (TAs) and sent through to us.	Frame error	We assume we get information on all building consents each TA processes. Any consents that can't get into the system would cause frame error (eg manual errors in date of consent so it is not in the monthly data we receive, or missing records due to TA data supply problems). Could include the $5000+ restriction if we consider the target set to be all building consents.
Accessed set	Building consents is a census of the consents that arrive, so the accessed set is the same as the accessible set: all consents that end up in the TA data sent to us.	Selection error	There should be no selection errors (eg sampling errors). Depending on exactly how we define the target population and where very small consents are removed (eg by TA or by Statistics NZ when loading the data), some errors mentioned under frame errors could be selection errors.
Observed set	Includes all units we have data for, so is the same as the accessible set.	Missing/ redundancy error	Once we 'select' a consent (we select 100% of consents) we always get a response – the consent both forms the target population and contains the responses we want.

Table 4
Representation side of phase 1 framework for Quarterly Building Activity Survey respondent data

Representation (objects)	QBAS units	Error types in representation	Potential errors arising in the QBAS survey
Target set	Active construction projects in NZ during ref qtr	Not applicable	Not applicable
Accessible set	For a survey this is the sampling frame. Construction jobs with building consents approved during the months of the reference quarter.	Frame error	Construction work is likely to happen outside the consent frame (eg people do small home renovations without getting a consent or realising they need one). Errors in the consents systems could mean a job doesn't appear in our consents for the relevant months (eg a manual error puts it in the wrong month, or another mistake when TAs' prepare data for us). The frame may also be in error for staged or split jobs. Sometimes stages are missed, so the corresponding job is not in the correct stratum. For split jobs, the building types could be difficult to determine or the value apportioned may be uncertain – jobs might not have a corresponding consent for the correct value/type.
Accessed set	For a survey this is the sample. Includes units selected into sample from the building consents each month, including the modelled, sample, and full-coverage strata.	Selection error	For full-coverage strata, selection errors should be minimal (but see staged or split jobs mentioned above). Same applies for the lowest strata, which can be treated as full-coverage – WPIP data for them comes from administrative data rather than survey. For sample strata there are sampling errors, which we calculate routinely for QBAS releases. Typical values are around 3% in the total WPIP across all buildings, and around 4% for residential/non-residential categories.

Representation (objects)	QBAS units	Error types in representation	Potential errors arising in the QBAS survey
Observed set	Final set of responding units in the dataset, which includes survey responses and modelled units. Non-response causes this set to be smaller than the accessed set.	Missing/redundancy error	QBAS treatment of unit non-response is very similar to that for item non-response – there is only really one target variable. According to the latest tech description, non-response is around 10%.

This example demonstrates how some types of error tend to affect survey data more than administrative data, and vice versa. For example, validity and measurement errors are often an issue in surveys, particularly in social surveys where concepts such as ethnicity or well-being are very difficult to define and measure, and respondents may not understand the questions in the way the designer intended. Administrative datasets, because they are created for an operational purpose, tend to aim to collect strictly defined information that matches their rules. They are often less affected by validity and measurement error.

Phase 1 assessments of administrative data sources are valuable because they are easy to pick up and reuse when we evaluate the dataset for a new statistical purpose. The building consent–QBAS example also demonstrates how to use the framework to assess a stand-alone survey that doesn't use any administrative data. This is useful if our aim is to compare an existing survey design with a new administrative data-based design, where we want to know all the quality issues and trade-offs involved.

Phase 2 of the error framework

Phase 2 of the error framework is where we evaluate the combination of separate data sources against a specific statistical purpose. The first step is to define the target population and target concepts clearly. We also need to understand the processes by which the source datasets are transformed into the final dataset.

Because QBAS uses the building consents dataset as a frame, reconciling the units in the two datasets and creating statistical units is fairly simple, and the most important sources of error are in the measurement side. Table 5 explains how the framework matches up with the QBAS design and the structure of each dataset.

Table 5
Measurement side of phase 2 framework for final QBAS output

Measure-ment (variables)	Output QBAS dataset combining building consents and survey responses	Phase 2 error types	Potential errors arising in final QBAS output
Target concept	Work put in place (WPIP) in each job during the reference quarter. Other important variables are secondary: building type, floor area, location/region, and institutional sector. These come from consents data (the frame).	Not applicable	Not applicable
Harmonised measures	The final measure on QBAS units is WPIP, as defined by QBAS questionnaire. For other variables, we have Statistics NZ’s building type, institutional sector, and location classifications.	Relevance errors	Refer to concepts, definitions, metadata, not actual data. WPIP variable: We assume conceptual alignment between QBAS WPIP question and target concept is excellent because it is a direct survey collection designed with the target concept in mind. For building consents and WPIP modelling we consider the conceptual alignment between consent value and WPIP. Major discrepancies at the conceptual level are: Consent value is the estimated total value of the project, while WPIP is the actual work done in a given period (quarter). Consents indicate confidence/ intentions at a point in time, while QBAS measures real economic activity over a certain period. Consent value includes GST; WPIP excludes GST. Other variables: QBAS questionnaire includes consent details for location, value, nature of job, and asks

Measure-ment (variables)	Output QBAS dataset combining building consents and survey responses	Phase 2 error types	Potential errors arising in final QBAS output
			respondents to correct details if wrong. This results in differences between QBAS and consents for these variables (due to timing, updates, and fixing mistakes rather than conceptual misalignment). The underlying concepts for all variables other than WPIP are the same for building consents and QBAS. These 'harmonised measures' may differ from the target concepts for these variables in similar ways to WPIP, because the consent is only a plan/estimate.
Reclassified measures	Not a lot of conversion is necessary for QBAS data, since we collect it with harmonised measures. The main conversion is for building type, where the job description on the consent has to be converted into our classification. In the new design, WPIP is modelled from the consent value and the age of the consent, based on historic survey responses. The reclassified measure includes details of the modelling methodology, including the rules that determine which jobs will be modelled and how.	Mapping error	For WPIP: for QBAS responses the WPIP we collect is already for the target harmonised measures (apart from adjusting from total WPIP to date to WPIP in the previous quarter using earlier responses, which could also be part of adjusting the measures below). When we model WPIP from building consent values, this raises further mapping error possibilities. Although modelling units from administrative data is similar to imputation, we distinguish the two because modelling is designed to deal with the conceptual mismatch between the administrative consent value variable and the target statistical variable. In contrast, imputation corrects for non-response in otherwise well-aligned variables. Building type is the most likely other variable to have mapping errors (eg free text

Measure-ment (variables)	Output QBAS dataset combining building consents and survey responses	Phase 2 error types	Potential errors arising in final QBAS output
			consent descriptions in administrative data are not clear or are ambiguous – the analyst has to judge the best fit). Ideally we'd like to measure WPIP by building type; this may be classified when the building consent comes in but left for the rest of the project even if the project changes slightly.
Adjusted measures	For QBAS data, our main adjustment is calculating quarterly WPIP by subtracting the previous WPIP from the latest survey response. We also edit and impute at this stage. Imputation uses a combination of auxiliary consents data, previous responses from the imputed unit, and responses from similar units. Editing, imputation, and adjustment for other variables at this stage is relatively minor.	Comparability error	WPIP imputation will never be perfectly accurate, which contributes to errors. The imputation method we use for QBAS respondents assumes all units in the imputation cell have a certain relationship between WPIP to date and the consent value. This means jobs that run slower or faster than average, or have complications during construction that increase costs, will be in error to some degree.

Table 6
Representation side of phase 2 framework for final QBAS output

Represent-ation (units)	**Output QBAS dataset combining building consents and survey responses**	**Phase two error types**	**Potential errors arising in final QBAS output**
Target population	Building projects in NZ that did work during the reference quarter.	Not applicable	Not applicable
Linked sets	The unit record dataset that matches up data from consents with (if the consent was in the survey) QBAS responses.	Coverage error	Coverage of the two data sources for the target population. We expect the building consents frame to cover nearly all significant building projects, except any lost due to clerical or other errors (as mentioned in phase 1). We don't expect coverage errors for QBAS to be an important contributor to overall errors in the output. Linkage of QBAS responses to their consents is a fairly trivial process. Note: the design doesn't cover consents < $5000.
Aligned sets	Alignment sorts the relationships between different sets of units in different datasets. QBAS has little distinction between linked and aligned sets, because the survey frame and statistical units all come directly from building consents. In most cases alignment is already achieved by the link between the QBAS form and consent number. Split and staged consents are the main problem, where we want to find all the consents, updates applying to a single job.	Identification error	These could result from staged or split consents that aren't identified. All consents relating to a single project should be linked to produce the aligned set of statistical units. Similarly, ideally we want to treat a split consent as two separate projects and count work done on the two different building types separately.
Statistical units	Creating a statistical unit corresponding to a building job/consent is	Unit error	Because the fundamental statistical units are based on building consents, unit errors

Represent-ation (units)	Output QBAS dataset combining building consents and survey responses	Phase two error types	Potential errors arising in final QBAS output
	simple for QBAS, because of the relationship between consents, sampling frame, and target population.		are minimal. Extra statistical units might be created if we miss staged or updated consents and treat them as new consents. However, they are more accurately identification errors, where we haven't correctly connected later consents to the original consent/statistical unit.

.

6 Implementation and future work

This section answers the following questions:

- What is this error framework missing?
- How should this framework be implemented and how will a central repository work?
- How do we see it being used, and who will be responsible for it?

Out-of-scope work

The error framework presented in this guide provides a general way to understand the quality of datasets and outputs. However, it is high-level and we haven't found solutions to many of the difficult technical problems that are the natural next steps to enhance this framework.

At the start of this project we identified key methodological gaps that were out of this guide's scope.

As part of developing this framework, we listed the methodological areas excluded from our scope. The main areas we identified were:

- time-series quality measurement
- the effect of confidentiality rules on quality
- measuring conceptual and validity errors
- quality measures for linked data
- combining different measures into a single overall measure
- weighting in the presence of combined survey and administrative data
- quality measures for statistical outputs that use administrative data for benchmarking or calibration
- quality measures for apportionment (eg of GST and rolling mean employee counts on the business register)
- quality measures for editing and imputation
- combined assessment of costs, quality, and respondent burden.

Statistics NZ, along with other statistical agencies and statistics researchers, has investigated many of these areas. A full summary of these complex areas is out of the scope of this framework.

Central repository of quality assessments

To maximise the benefits from the error framework in this guide, we must make the outcomes of quality assessments available in a central repository. Statistics NZ uses Colectica as a standard corporate metadata tool, and it can naturally be extended for use as a quality assessment repository.

- It already contains a lot of basic metadata, links to existing studies, and other information, so we avoid duplication.
- Detailed documents, such as feasibility studies and in-depth quality assessments can be linked into Colectica (many already are).
- It is possible to add 'quality statement' templates to Colectica that can capture more detailed information in an organised way.

- The standard metadata items can be modified over time so we capture and organise key information. Our template contains good questions that can be easily integrated into Colectica.
- We can create phase 1 and 2 assessment templates in a format to be directly uploaded to Colectica.

Centralising quality assessments has the following benefits.

- Eliminates duplication of work and lets new studies build on the old (especially for phase 1 assessments).
- Ensures all relevant work and knowledge about a given dataset is easily found in one place for reuse.
- Encourages certain basic information to be understood and recorded for all our collections in the same way.
- Provides an easy way to find quality information for releases and to answer queries.
- Assists analysts new to undertaking quality assessments (with examples).
- Provides guidance for performing quality assessments of sample and census survey data.

Implementing the error framework

Using administrative data requires our statistical analysts to change the way we organise our work. Since the providers produce administrative data primarily for their own use, the data have to be assessed before we can use them in statistical outputs. The assessment replaces the controls we generally rely on during the initial phases of a survey. The error framework guides the assessment process and leads the user of the administrative data to decide on its potential uses.

We envision that more and more datasets and outputs will be assessed using this error framework. Although the framework and measures we've presented here are very detailed, the time we've available to undertake quality assessment for an administrative data source may be limited. Initially, data quality assessments may only focus on the most important and useful aspects of the framework.

Tracking the changes

New users may need to update an existing quality assessment when considering their data use or when the administrative data changes. We recommend that analysts include an audit trail when updating the meta-information template or when computing additional phase 1 quality indicators. The audit trail should indicate who did the update and its date. The current meta-information template includes an audit trail section, but now we have loaded the template into our Colectica tool we can more easily keep track of changes and share them. We're continuing to develop our systems for recording and updating quality information about datasets and outputs.

Future work

This guide is not intended to be the final word on administrative data quality, but it should provide a consistent language and structure for assessments. As we gain more experience applying it in different contexts, we will probably discover gaps or ambiguities in the types of error and the aspects of datasets we need to consider.

A major area for future research that we've found through recent Census Transformation work is 'Phase 3' for the framework and assessment process. The idea is to build knowledge of the sources of errors in the output microdata into a model that would attempt to correct for, and quantify, the uncertainty these errors introduce into our statistical estimates. Bryant and Graham (2015) describe this sort of model for use in population estimation from administrative datasets.

We hope that a ‘virtuous cycle’ can be created where we use the errors identified to help correct and improve the model, and use the model to measure and test the effect of the errors on the output. This would take us closer to the goal of using the framework to compute a general ’total survey error’ quantity for the statistical outputs we produce from administrative data.

References

Bakker, B (2010). *Micro-integration: State of the art*. Paper presented at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Bryant, J, & Graham, P (2015). A Bayesian approach to population estimation with administrative data. *Journal of Official Statistics* 31: 475–487. DOI: 10.1515/JOS-2015-0028.

Groves, RM, Fowler, FJ Jr, Couper, MP, Lepkowski, JM, Singer, E, & R. Tourangeau. (2004). *Survey Methodology*. New York: Wiley.

UNECE (2013). The Generic Statistical Business Process Model. Retrieved from www1.unece.org.

Zhang, L-C (2012). "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica* 66: 41–63, DOI: 10.1111/j.1467-9574.2011.00508.x.

Appendix 1: Glossary

Here are clear definitions and explanations for the technical terms used in this guide.

Term	Explanation
Accessed set	The set of objects for which measurements are obtained in practice. For example the electoral roll doesn't include people who fail to enrol despite being legally entitled, or whose forms get lost in the mail.
Accessibility	Statistics are presented in a clear and understandable manner and are widely disseminated. One of the six quality dimensions.
Accessible set	The set of objects from which measurements can be taken in theory.
Accuracy	Source data and statistical techniques are sound and statistical outputs sufficiently portray the reality they are designed to represent. One of the six quality dimensions.
Admin(istrative) data	Admin data is all data collected by government agencies or private organisations in conducting their business or services Such data is not collected primarily for statistical purposes. Rather, it is collected or captured for operations such as delivering a service, registering members, events, or activities, or as legally required records.
Aligned sets	The groups of base units that are determined (after linking and other processing) to belong to each composite unit in a final output dataset. For instance, we might create household units based on dwelling units and person units – the aligned sets could be represented by a table containing all these relationships (eg household 1 consists of dwelling A and persons X, Y, Z; household 2 consists of dwelling B and person W)
Base dataset	Where data integration is carried out by linking one or more datasets to a single large dataset we call this central dataset the base dataset.
Base units	The lowest-level units created after linking within and across datasets. These units often represent individual people, businesses, or dwellings.
Comparability error	An error arising from editing and other treatment methods applied to values obtained from reclassified measures – to correct for missing values, inconsistencies, or invalid values.
Composite unit	A unit made up of one or more base units. These are not necessarily the final statistical units in the output: intermediate composite units may be created and further combined or arranged into the final statistical units.
Consistency	Statistics are consistent and coherent within the dataset, over time, and with other major datasets. One of the six quality dimensions.
Coverage error	The differences between the units actually included in the linked datasets in practice (linked set) and the full set of units included in the (ideal) target population. Coverage errors can arise in several ways. For example, the datasets themselves may not cover the whole target population, or linking errors may mean some members of the linked sets are not identified. This error type may also be caused by measurement errors. For example, if the date of birth variable on an admin dataset is not of good quality and we filter on age to select our population, we could end up with undercoverage even though the units aren't missing from the source data.
Dimensions of quality	A guide to help a national statistical office manage quality in their operations, to ensure customers can have confidence in the statistics published. These dimensions are: accuracy, relevance, timeliness, accessibility, consistency, and interpretability.

Term	Explanation
Edited measure	The final values recorded in an admin or survey dataset, after any processing, validation, and other checks. This term is only relevant in phase 1 of the error framework.
Final dataset	An output micro-dataset after all processing and checks are completed.
Frame error	The difference between the ideal target set of objects and the accessible set. These errors refer to objects that are inaccessible, even in principle. In a survey context the accessible set is the sampling frame. For an admin source objects may be inaccessible for many reasons.
Harmonised measure	The operational measures decided on in designing the statistical output to capture the target concepts. They include elements such as questions, classifications, and variable definitions. For example, a survey question aligned with a standard classification.
Identification error	Misalignment between the linked set and the aligned set. This type of error also includes situations where the target statistical units cannot be adequately represented using combinations of base units. For example, to measure the economic activity of all manufacturing businesses by industry, we would ideally have separate statistical units to capture different types of manufacturing done by a single company. However, in practice we might have to define statistical units via legal entities. Changes in company or legal structures might result in statistical units being absorbed into others, despite no real-world change in economic activity occurring.
Indicator	A numerical or descriptive value that can be used to measure or report on an aspect of quality. An indicator can be either a quantitative or qualitative measure.
Interpretability	Processes and methods used to produce official statistics, including measures of quality such as estimated measurement errors, are fully documented and available so customers can understand the data and determine whether it meets their needs. One of the six quality dimensions.
Input dataset	Any datasets assessed in phase 1 that are combined and processed to produce the final statistical dataset.
Input quality	Aspects of the quality of an original data source at the point where it is finalised by the admin agency. The quality of a dataset is assessed to determine any treatment required for it to be used in the statistical production of outputs. Input quality is best assessed for the data's original purpose. This allows a dataset to have a single assessment that can be used by anyone trying to use the dataset for different purposes – quality issues have different effects depending on what is done with the original data.
Li-Chun Zhang's framework	The framework developed by Li-Chun Zhang provides a well-defined list of errors that can occur when producing statistics, using a given dataset or combinations of various datasets. Phase 1 of Li-Chun Zhang's model allows a single data source to be evaluated for the purpose for which data was collected. This evaluation is entirely for the input dataset itself, and does not depend on what we intend to do with the data. In phase 1 the focus is on 'objects', which could be events, transactions, or other entries in an admin dataset. Phase 2 of the error framework covers errors that arise when existing data is used to produce an output that meets a certain statistical purpose. In phase 2 the reference point is the statistical population we would ideally have access to, and the statistical concepts we want to measure about the units in the population.

Term	Explanation
Linked set	Includes all the basic objects from across all source datasets that are matched together to make base units. These units will not necessarily be the output's final statistical units.
Mapping error	These arise from transforming variables on the input datasets into defined output variables (the harmonised measures). Such transformations include: • Reclassifying from a non-standard classification, or coding a free text field. • Deriving a numerical variable from a source dataset, such as removing GST from a transaction value. • Modelling a target variable using a combination of several variables on a source dataset, and some model parameters. In each instance the value of the output variable may differ from the true value – these differences are mapping errors.
Measure	When used in the context of the error framework (eg target or adjusted measure) 'measure' refers to the practical definition and method for capturing a variable value. For example, a question on a survey or admin form, including the instructions and definitions given to respondents. A quality measure is a value derived by analysing a dataset or metadata that captures information about an aspect of the dataset's quality.
Measurement (variable) in phase 1	Refers to Li-Chun Zhang's framework. The measurement side of the diagram is the path from the (possibly abstract) target concept the data is intended to capture, to a final processed value for a concretely defined variable. Sources of error on the measurement side include the degree to which the operational measure used captures the target concept, and how many and what kind of errors are introduced by respondent misunderstanding or processing difficulties.
Measurement (variables) in phase 2	Refers to Li-Chun Zhang's framework. The measurement side in phase 2 is concerned with how well the final values of the output statistical variables capture information about the target statistical concept. Measurement errors in phase 2 are mostly result from a mismatch (eg in concept, definitions, classifications) between the variables on the original source datasets and the target concept the final output aims for. Ideally, variables are collected using classifications and questions that match what we would use to collect the same information in a specialised survey.
Measurement error	This occurs when the obtained measure (value actually recorded in the dataset) differs from the measurement intended. Errors could include people misremembering details or interpreting questions differently from their design. In more automated admin systems, such as electronic transaction records, measurement errors could include computer system problems that corrupt some values or introduce ambiguity.
Missing/redundancy error	Misalignment between the accessed set and the observed set. For example, errors where an agency mistakenly rejects or duplicates objects, due to their own processing, could mean objects are missing from the dataset even though correct data was received about them. This category of error exists so such errors are kept distinct from reporting-type errors. (Compare with selection error.)
Object	This could be events, transactions, or other entries in an admin dataset. The final dataset at the end of all the phase 1 transformations is organised in terms of 'objects' rather than 'units' – to avoid confusion.

Term	Explanation
Obtained measure	The values initially received for specific variables against objects in the dataset.
Observed set	The set of objects that end up in the final, verified dataset after all processing by the source agency.
Output quality	The quality of the final statistical product for the intended statistical purpose. Reporting on output quality could involve many quantitative measures and explanatory notes about possible limitations.
Processing error	These arise from editing and other processing done by the source agency to correct or change the initial values received (the obtained measures). This kind of processing is done to improve the quality of the data for the target concept, but it is important to understand how much improvement it makes, as well as any limitations introduced by the processing.
Relevance error	These are the phase 2 errors analogous to validity errors. They are errors at a conceptual level that arise from the fact that the concrete harmonised measure usually fails to precisely capture the abstract statistical target concept. For example, we want to find out about personal income but we only measure taxable income – this creates a relevance error, since non-taxable income is part of our target concept but not our harmonised measure.
Representation (objects) in phase 1	Representation concerns creating the final list of objects in an individual output dataset. This part of the framework deals with how well the objects in the dataset match the objects in the target set (or target population). Ideally every object in the target set has a corresponding object recorded in the data.
Representation (objects) in phase 2	The representation side of phase 2 deals with creating a list of statistical units to be included in the output data, based on the source data's objects. Sometimes this list is created directly from the list of objects in a source dataset, but in complex cases different types of linked units created from several datasets might be combined into new statistical composite units.
Selection error	These errors arise when objects in the accessible set do not appear in the accessed set. For example, if a store manager forgets to run the reporting tool for a week, the transactions missing from the dataset due to that mistake are selection errors: they were accessible, but were not accessed.
Source agency	The business, organisation, or group originally responsible for the design and creation of an individual dataset.
Target concept	This is 'the ideal information sought about an object' for phase 1, and 'the ideal information sought about the statistical units' for phase 2. The target concept is usually connected to the underlying purpose of the collection and may be quite abstract. Examples are: household income, political views, advertising effectiveness, or population counts.
Target measure	The operational measurement used in practice by a source agency to capture information. A target measure includes elements such as variable definitions, classifications, a questionnaire, or rules and instructions for people filing out forms.
Target set	The set of all objects the data producer would ideally have data on. For example, people, businesses, events, and transactions.
Target population	The ideal set of statistical units that a final dataset should cover.
Timeliness	Data is released within a time period that permits the information to be of value to customers. One of the six quality dimensions.

Term	Explanation
Unit error	Creating the final statistical units for the output dataset can introduce unit errors. For instance, to create household units from aligned sets of dwellings and people, we must simultaneously decide which dwellings should have a household created, and which people should go into which household unit. Because the statistical units may not correspond to any of the units in the source data, a variety of errors can arise at this stage.
Validity error	This error refers to misalignment between the ideal target information and the operational 'target measure' used to collect it. The error arises from translating from an abstract target concept (the ideal information sought from the admin dataset about an object) to a concrete target measure, which can actually be observed in practice, and does not include issues such as misunderstanding a term used on a form.

See 'Available files' for appendixes 2 and 3 (the quality indicators for phases 1 and 2).