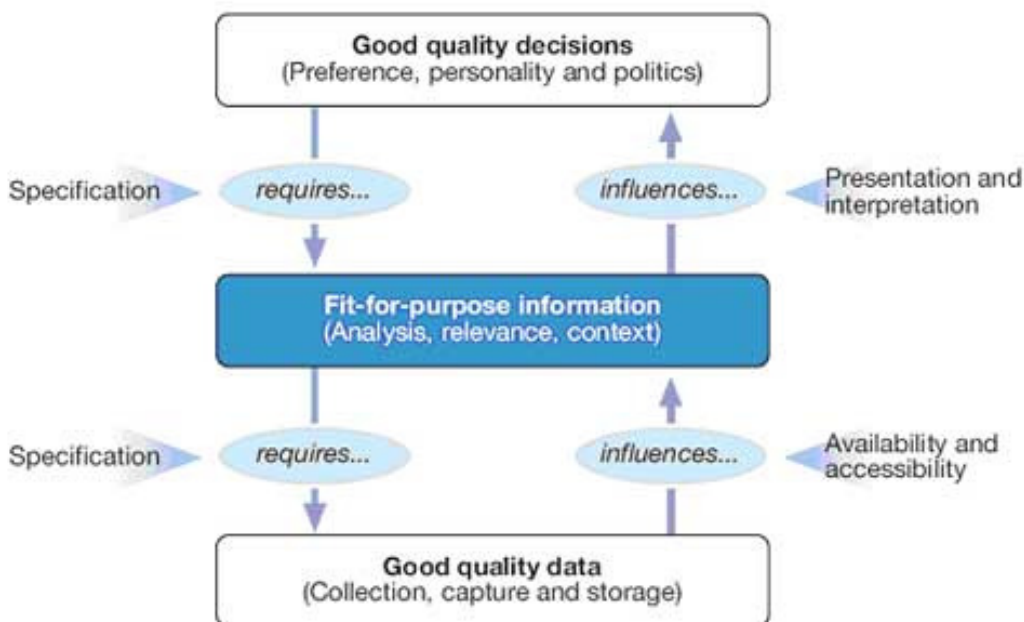


A Guide to Data Quality

Why is Data Quality Important?

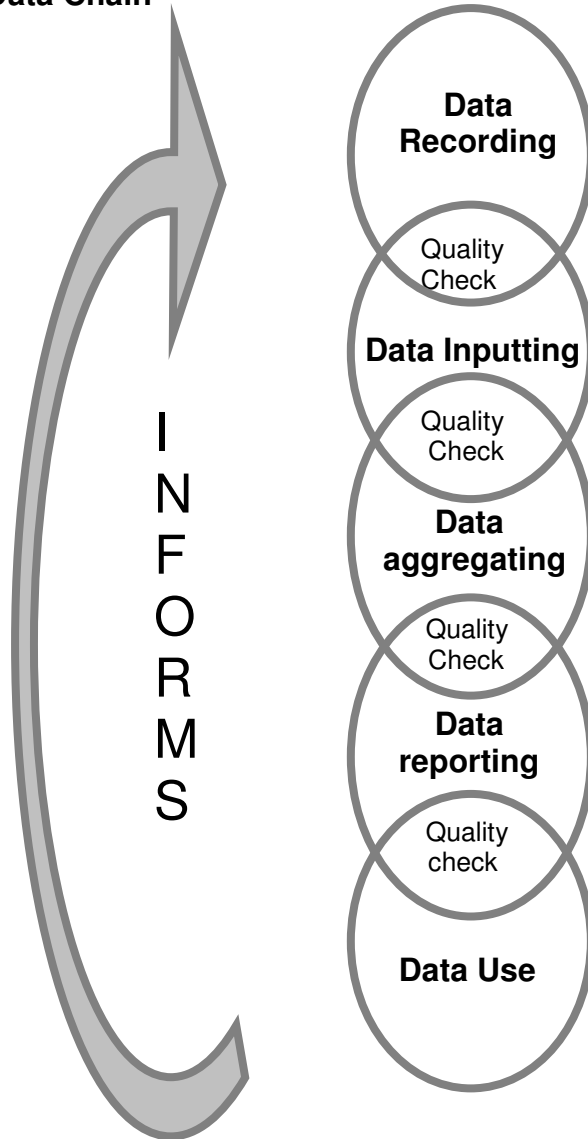
Public service organisations need information suitable for managing services, to evaluate their own performance and provide intelligence that will aid strategic decision-making. The information may be used by the organisation’s service managers, government departments, service users and other members of the public and strategic partnerships. All require accessible information on which to make informed decisions.

Clearly the data on which the information is based needs to be of the highest possible quality. Failure to provide such information or understand its limitations creates risks and diminishes the adequacy of the decision (see model of the Data-Decision cycle below):



Public bodies are accountable for the public money they spend, and the Audit Commission requires them to have a documented, evidenced and embedded Data Quality Strategy in place within the organisation. Somerset County Council’s own [Data Quality Strategy](#) sets out the steps to ensure good quality data and the systems and structures put in place to follow them. This paper aims to complement the Strategy by explaining in more detail what we actually mean by ‘good quality’ data.

The Data Chain



Producing and using good quality data needs the right culture, people and processes to support it, and the quality should be checked at each stage of the process (see above).

The quality of data should be assured as close to the point of origin as possible. Any errors can be compounded as the data moves along the chain or between organisations and it can become more difficult to locate any error that becomes apparent, if indeed it is noticed at all before decisions are made.

The Six Principles of Data Quality

The Audit Commission offers six key characteristics of good quality data:-

- Accuracy
- Validity
- Reliability
- Timeliness
- Relevance
- Completeness

1. Accuracy

This could relate to the way the data is recorded as well as the way it is captured. Data should clearly represent what is being measured as close to the event as possible (e.g. data on hospital admissions, incidences of crime) while memories are fresh. Maintaining standard data inputting procedures used at the time (or soon afterwards) should improve the accuracy of data.

However, there may be compromises involving accuracy and timeliness. For instance, some reported data requires validation, which may lead to a re-categorisation and subsequent re-release of data. If you use the most recent data release, you may need to be aware that the data may be revised later on. In this instance, it may be more advisable to use slightly older but more reliable and accurate data. Wherever there needs to be some sort of compromise in accuracy, the limitations of the data must be made clear to the end user. This means that the risks relating to using the data need to be stated, mitigated and reported alongside the data itself.

There are different ways of obtaining data, from observation (such as traffic counts), registering incidents or events (such as hospital admissions or crime) to asking a representative group of individuals or organisations for their opinions or details about their lives by means of a survey which, unless it is a census, can never be 100% accurate.

In a survey, how accurately the data represents the true picture of the target population will depend on the quality of the questions asked, the methodology used and the size of the sample. For instance, questions need to be clearly worded and unambiguous, otherwise answers will also be vague and probably of little use. If the respondent understands and engages with the purpose of the questionnaire, he/she may be more likely to give a considered answer or even any answer at all, thus improving the accuracy of the overall survey data. An incentive such as a Prize Draw can boost response. However, the choice of prize should not influence how people respond and should be appropriate to the target population.

There are many different ways of doing a survey. The pros and cons of using different methodologies are given in the Appendix.

As with any survey or research project, it is recommended that a small-scale 'pilot' should be carried out in order to identify and correct as many potential issues and threats to data accuracy as possible.

By testing the survey on a small number of individuals – ideally those in the target group – it may be possible to find problems with things like unclear or ambiguous questions, the way respondents are 'routed' through the survey and also whether all the required topics are covered and respondent data (e.g. age and sex) are captured.

It is also important that the people who collect or record the data engage with the process and realise that it is not merely a 'box-ticking exercise' and that if errors are made it can have a knock-on effect on the effectiveness of service delivery.

Another factor to consider is how to treat missing values. There may be a valid reason for an apparent omission, so it should not automatically be assumed there is an error. An understanding of the data collection process is important in judging how to treat a missing data value. For instance, perhaps there has been an issue with entering data at one location, or perhaps a question was irrelevant to, or could not be answered by a respondent.

Often data need to be provided to decimal places rather than whole numbers. In such cases, where data need to be grouped together or aggregated in some way, it is best to maintain the same number of decimal places. This is to avoid compounding any potential error caused by rounding numbers to a whole number.

2. Validity

There may be rules or guidelines governing the recording and use of data. These normally exist to ensure that you are measuring the right thing and indeed measuring what you think you are measuring. Following the correct regulations and data definitions should ensure consistency between different time periods and/or data provided by other organisations. Data will be of higher quality if they allow valid comparisons to be made.

The Audit Commission's recommendations on data governance, policies, systems, skills and ultimate use complement further professional guidelines, in particular the Office of National Statistics [Code of Practice](#) the CIPFA Statistical Information Service [Code of Conduct](#) and, regarding electronic delivery systems, the Government Interoperability Framework ([e-GIF](#)).

There are also Research Governance Frameworks pertaining to particular public services. For instance, principles, requirements and standards for

anyone involved in Health and Social Care research and data collection (including those set out in Somerset County Council's Children and Young People directorate's [policy documents](#)) are governed by the Department of Health's [Research Governance Framework for Health and Social Care](#).

An example of research which requires a consistent approach is the Place Survey. This needs to be carried out at a local level, but must comply with national guidelines so that each local dataset can be benchmarked against each other and regional/national averages. Similarly, any piece of research should comply with the Market Research Society (MRS) [Code of Conduct](#). While this does not take precedence over national law, the Code of Conduct is designed to support all those engaged in marketing or social research in maintaining professional standards.

Where proxy data are used to compensate for an absence of actual data, organisations must consider how well these data are able to satisfy the intended purpose. For instance, if you want to examine differences in crime levels across the county, you might have data on only one or two measures (such as acquisitive crime and assaults). Thought must be given to whether these provide a sensible representation of total crime and, if not, other data sources may need to be explored.

National segmentations such as Mosaic and ACORN can be useful as a way of describing local areas in terms of what makes residents different from those in other areas in their demographics, attributes and characteristics. However, please take care not to assume that all individuals are exactly the same. It is also possible that for any measure representing a group, none of the individuals will match the average figure. For instance, in a group of people aged 20, 25, 75 and 80, the average age would be 50. However, none of those in the sample are aged 50 or even close to 50, so merely quoting the average in this case would be misleading.

It is also good practice to check data against a comparable alternative data source; for example, where an identical question has been asked on a different survey. It can be reassuring if the data is similar but if there is a wider-than-expected discrepancy, it is important to understand the possible reason, such as a different methodology, sample or timeframe.

3. Reliability

Where and how data are collected should be consistent over time. It is important to know that if you take the same measure in, say, a year's time how likely it is that you will get the same result. By reducing changes in the data collection process (methodology, sample, time of year, etc) it is more likely that any observed changes in results reflect real movement or trends.

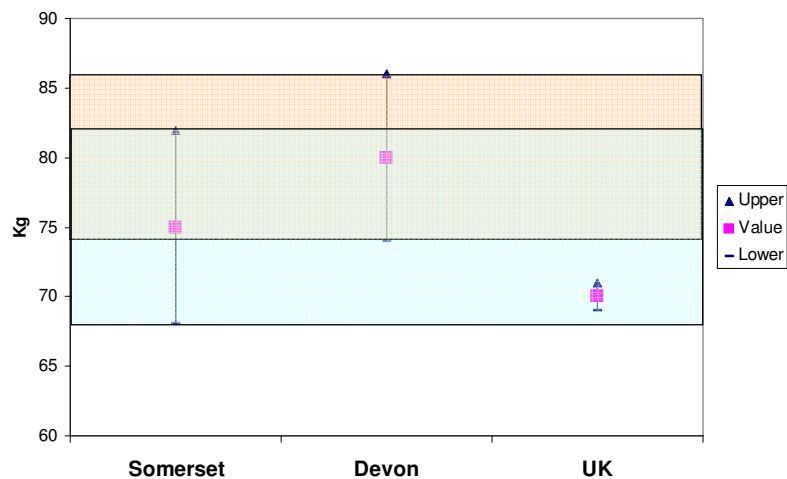
Where data has been collected from the whole population (as in the Census), you can be confident that the data reliably represent the population. However,

where data has been collected from a subset of the population (for instance, from a survey), it is extremely unlikely that you would obtain identical values from different subsets and in reality, the values would be different. The variation in values is defined as the sampling error. When interpreting high quality data, it is important to know the confidence interval for an estimated figure; in other words, the range of values within which the user can be confident that the actual figure lies.

The smaller the sample size, the wider the confidence interval and so the less reliable are the data as estimates of reality. This needs to be borne in mind when making comparisons between different data sets.

For instance, say the average weight of an adult in the UK is 70kg. In a sample of adults in Somerset, the average figure may be observed as 75kg. While this is larger than the UK average, we may only be confident that the true value for Somerset lies between, say, 68 and 82kg (the 'confidence interval'). In this case, the difference between Somerset and the UK is said to be 'not statistically significant' as the UK figure of 70kg falls within this confidence interval (see chart below).

If, however, the value observed for Devon was 80kg, with a confidence interval of 74-86kg, it is very unlikely that an average value of 70kg would have occurred by chance and thus the difference between Devon and the UK can be described as 'statistically significant'. Because the confidence intervals for Somerset and Devon overlap, we cannot say that there are statistically significant differences between adult weights in the two counties.



The same principles apply when comparing data over time to assess whether or not an increase is significant. This is why, when defining targets for local authority National Indicators based on a data source such as the Place Survey, the confidence interval is taken into account.

Most analysis software systems will calculate the confidence limits for estimated values.

4. Timeliness

The more up-to-date the data, the more useful they are. This could involve being captured as soon as possible after the event (e.g. a hospital admission or a service user's opinion of his/her interaction with the service provider) and be available for the intended use within a reasonable time period.

In the first example, maintaining up-to-date records enables strategic decision-makers to monitor performance trends in a more timely manner. In the arena of customer relations management, the availability of very recent data offers the potential to address any issues more quickly.

It may be faster to extract and make informed decisions based on data which is collected on a continuous basis than on data collected or released less frequently. For instance, some government statistics are released monthly while others only quarterly or annually. As with other factors of data quality, there is a trade-off between timeliness and accuracy where, for instance, complex data needs to be verified over a long period before departments can be confident it is of sufficiently high quality to be published.

Nevertheless, sometimes it is necessary to go back further in time for data. The latest year's figures may not have a sufficiently large sample size to be robust enough for reliable analysis, particularly at a sub-national level. For instance, county or regional data on teenage conceptions or serious road casualties are often published on rolling three-year periods.

Other data can take even longer because of the sheer volume of data that needs to be analysed. The national Census is a case in point. Because it aims to capture data on every person in the UK, the Census should be extremely accurate, valid and reliable, although by the time some of the figures are published, they are no longer especially timely.

Nevertheless, provided the reasons are understood and that accuracy is paramount in decision-making, such periodic large-scale datasets remain extremely important in understanding the populations we serve. For instance, the Office of National Statistics produces population estimates and projections based on Census figures and subsequent birth, death and migration data. These won't be precise but, supplemented by local knowledge, they give a good basis for strategic decision-making. On the other hand, Census data on ethnicity can become out-of-date quickly, and official figures on migrant workers also lose their value because of the often short-term nature of such migration. This in turn makes it harder for organisations to locate and count some of those with specific needs, and therefore to deliver services that meet those needs.

When reviewing performance it is imperative to capture timely data in order to properly assess whether objectives have been met within the specified timescale.

Time is also a factor when considering seasonal fluctuations. People's activities and attitudes may be different at different times of year. For instance, people may feel happier during summer months, poorer just after Christmas and those with children in education are likely to have different work and travel patterns during school holiday periods. Annual surveys should be conducted at the same time of year in order to reduce bias caused by seasonal factors often only indirectly related to the data being collected.

5. Relevance

Data captured should be relevant to the purposes for their intended use. Irrelevant data may cloud the issue and not only make it harder to analyse and interpret but also potentially break trust with the public who are being asked to provide such data.

However, it may be necessary to capture data which at that time may not be directly relevant. For instance, asking questions about disability may not be obviously relevant to a questionnaire on library book preferences other than to ensure a better representation of the local population or users. It may later prove appropriate to examining comparisons between those with physical impairment levels in terms of accessibility of resources. In any circumstances, when capturing personal information, we would need to justify the relevance of the task in hand under Data Protection legislation.

When designing a survey it is not appropriate to adopt an 'ask everything just in case' approach. The burden placed on the respondent should not be excessive and the respondent should be capable of answering all questions. When providing options to select, all possibilities should be included to enable the respondent to give his/her opinion and not feel constrained by the questioner. If it is difficult to list every possible option, then provide room to write additional comments, or include an 'Other' option, with a 'Please give details' section in which the respondent can do just that.

In any case, requirements should be reviewed periodically against potentially changing needs, and there needs to be opportunities for feedback to check that data is indeed of high quality and that this remains the most suitable source.

6. Completeness

Data requirements should be clearly specified based on the information needed and the means of collecting the data. Different organisations may attempt to collect data in different ways. Within individual organisations, the data may be of good quality but when combining the data from all organisations, the effect of having an inconsistent approach can make interpretation extremely difficult. This emphasises the importance of agreeing clear objectives for collecting the data.

When identifying the target population it is also vital to ensure all relevant groups are sampled to meet these objectives. For instance, when consulting on a new road scheme, it would be important to obtain feedback from a variety of potential users as well as local residents who would be affected by the scheme. By omitting key stakeholders, data relating to an important subset of the target group will have been missed, with the potential for misleading interpretation and policy recommendations.

It is vital to monitor missing, incomplete or invalid records, as these could highlight problems in the recording of data. For example, using an out-of-date postal address file can lead to a degradation of data because postcodes can become redundant and new ones created. If there is a new housing estate in an area of interest, it may not be represented in the address file. This could lead to a postal survey omitting an important section of the local community that aims to target, thus reducing the quality of the data.

Even the UK Census does not reach 100% of UK residents but it is important to understand what types of people do not respond. For example this could be done using a pilot survey or, in the case of the Census, a follow-up survey (details shown [here](#)).

Datasets often do not publish detailed data at very local levels. This is because by showing details of age, sex and/or postcode, it may be possible to identify the actual individuals to whom the data refer, which could contravene data protection regulations. Datasets may either show no figures at all or include 'proxy' data based on similar areas or aggregating data based on a larger area or longer timeframe. Knowing how the data is collected or displayed (from the 'metadata') is important in the interpretation of the information as well as assessing the data quality.

In conclusion

It is important that all those involved in the process of collecting, reporting and interpreting data, as well as setting, monitoring and reviewing standards of data quality, are aware of these principles.

APPENDIX: Pros and Cons of Using Different Types of Surveys

	PROS	CONS
Postal surveys	<ul style="list-style-type: none"> - Relatively cheap - Can reach target population relatively easily - Less intrusive; respondent can do it in their own time and at their own pace - Good for questions of personal nature 	<ul style="list-style-type: none"> - Response rates not high - Response bias towards older people with more time on their hands. Young people and ethnic minority groups less likely to answer - Need to be simple and easily understood, with clear instructions - Unsuitable for residents with low literacy levels or less familiarity with English; may need translations or a helpline to assist such people - Need up-to-date postal address files - Uses a lot of paper - Takes longer
Face-to-face interviews	<ul style="list-style-type: none"> - Allows more interaction with the respondent, including probing for more information. This can enable more honest and thoughtful answers - Selecting individuals at home reduces sample error and bias, allowing for call-backs - Survey can be more complex, including use of additional materials such as show cards, pictures, etc - Overcomes most literacy problems 	<ul style="list-style-type: none"> - Relatively expensive, especially if visiting people at home - Street-based interviews are cheaper but may be harder to reach representative sample of target population - Requires trained interviewers - Potential bias from the respondent or interviewer – may give a response they think you want to hear - Needs more planning time - More difficult in times of bad weather (e.g. winter)

	PROS	CONS
Telephone interviews	<ul style="list-style-type: none"> - Relatively quick and inexpensive to carry out, and also analyse results if using CATI (Computer-Assisted Telephone Interviewing) - Better response rates - More interaction than postal, with probing possible - Overcomes most literacy problems 	<ul style="list-style-type: none"> - Biased towards those with landline telephones - Increasing problem of people having TPS to screen phone calls or moving to mobiles - Requires trained interviewers - Can be intrusive, but overcome by arranging contact time to suit respondent - Not suitable for some subjects - Not suitable for interviews longer than 20 minutes
Web-based survey	<ul style="list-style-type: none"> - Relatively cheap and quick to carry out, and also to analyse, especially if using bespoke software - Good way of reaching the young, traditionally hard to reach by other methods - Little staff resource needed - Allows complex routing and potential to use additional materials such as images, audio or video clips, but overall design needs to be simple and engaging - Less intrusive; respondent can do it in their own time and at their own pace - No barrier to geography 	<ul style="list-style-type: none"> - Biased towards those with internet access, so harder to reach older and/or less wealthy people - Relatively low response rate, so answers less likely to represent target population - Unlikely to include those with low literacy levels - No guarantee about who has completed the survey - No easily available list of people's email addresses unless sample is, say, for college students or company staff. - Sample could include people from overseas which could distort results