Creating a Data Quality Model

EuroGeographics Quality Knowledge Exchange Network

Author: Jonathan Holmes QKEN Chair

1. Purpose of this document

This document was created by the EuroGeographics Quality Knowledge Exchange Network (QKEN) with the purpose of explaining what a Data Quality Model (DQM) is; why you would need one; how you would implement it; and what the risks to a successful implementation are. It is intended as a high-level overview, with references out to more detailed information.

2. What is a Data Quality Model?

Before delving too deeply into the topic of Data Quality Models, it is useful to understand what a DQM is.

The QKEN have defined a Quality Model as:

"A framework for defining, evaluating, documenting, and presenting the quality of spatial data sets and geoservices according to their specifications."

It is important to realise that the creation of such a model is just the beginning. The people, processes and documentation will all need continuous update and management as requirements change, errors are identified, and new people work on a dataset, process, product or service.

3. Why would you need a Data Quality Model?

A good DQM will ensure that the Users' needs are met in a timely and efficient manner. User needs are always changing, and a good model will be adaptable and be able to quickly respond to changes.

Implemented properly a DQM will deliver:

- A common understanding of the data across the organization
- Improved efficiency
- Lower costs
- Insight and confidence in the data
- The ability to manage and monitor data quality

4. What is in a Data Quality Model?

A DQM consists of a number of distinct parts. The depth to which each of the parts should be implemented depends on the user needs, the dataset and how important the product or service is to an organisation.

Use Cases

The most critical part of any DQM are the Use Cases. Without these, a producer is only guessing at what the end user is trying to do with the product or service. This in turn means that everything that follows could be wrong.

Data Requirements / Specifications

Data Requirements turn the use cases into something that can be developed and measured against. They should inform the developers of what should be included in the product or service and also identify what quality is required of the data. The quality is usually expressed in terms of acceptable quality levels (AQLs). Both the data requirements and AQLs may be documented within the specifications.

Evaluation Methods

Evaluation methods identify how the quality of the data or service will be measured. There should be a direct correlation with what is documented in the requirements / specification which, in turn, should be aligned with the Use Cases.

Reporting

With any dataset it is important to identify when, where and how often the results are published. This is an important step as these reports highlight to management where issues may exist and when action needs to be taken.

Maintenance

A fundamental part of any DQM is the definition of how quality will be maintained during any maintenance phase. In a maintained dataset, the greatest risk to data quality is the process and people who maintain it. The maintenance section of a DQM defines where, when and how quality is controlled when the data is maintained.

Metadata

The metadata provides valuable information to the user on the quality of the data, the data model and the production process. It helps the user to confirm that the data will meet their use case.

5. How to implement a Data Quality Model

People

The most critical part of a DQM are the people involved in it. They will ultimately determine whether a DQM is successful or not.

The first step in this is to ensure support from the management team. They will need to understand the benefits and risks in order to support the development. Without their buy-in, the money and resources will not be made available to implement the DQM.

Quality Concept

Once support from the management team is secured the next step is to ensure that there is a common understanding of what is trying to be achieved amongst the team that are implementing the DQM.

In a quality concept the basic outline of quality management can be summarized. It defines its main principles and points out measures for quality assurance and quality control. The quality concept describes the process of implementation and will have great influence over the way that it is developed and implemented.

Once the use cases have been developed, it is useful to share them with the implementation team as they will be able to point out any issues.

Identifying Use Cases

It cannot be stressed how important this phase is. Identifying what the users are trying to do with the data is critical. By understanding this, any producer will then understand what the end outcome is expected to be and will then be able to identify any instances which will affect the outcome.

There are various methods (which we won't touch on here) available for documenting use cases but even if they are written down in a word document then that is better than nothing. A web search will quickly identify a number of different methods and technologies.

Developing Data Requirements

Once the use cases have been developed, these should provide the basis for the data requirements. This phase will include the identification of what data is required in the dataset. This is of interest to the DQM but is not as important as the required quality levels for that data. In order to meet a use case there will be a point where the data is of such a poor quality that it no longer produces an acceptable result. Agreed quality levels need to be documented, below which the data will not be acceptable for release. These will form the basis for future evaluation and are important to identify early. Acceptable quality levels around completeness of the dataset, positional accuracy, logical consistency, thematic accuracy and possibly temporal accuracy could be created at this stage.

Again, there are methods and systems available for documenting data requirements, but anywhere will do as long as they are written down.

Define Evaluation Methods

Once the acceptable quality levels are known, it is necessary to identify and document the procedures by which the quality will be measured. These will vary depending on the type of evaluation being made. For examples, a completeness check may require ground truthing whereas a check on the values in a field when compared with the permitted values, can be done directly on the data. EuroGeographics QKEN has produced a document that covers the definition of evaluation methods. See section 2.7.4 of ISO 19100 Guidelines NB. You will need to be logged in to access this.

The methods and software for implementing the evaluation methods are many and varied and will often depend on what type of database is used in the implementation. Sometimes it may be possible to query the database directly and other times it will be necessary to view the data and compare it against another dataset. There is no single method that will cover all bases. For this reason, evaluation tools and methods are often focused on just one product.

Decide on Reporting Mechanisms

The required quality levels are known, the evaluation methods have been determined and assessment of the data made. You have results so how you are going to report them? There is no point in taking the measurements if they don't get reported to those people who are responsible for the data and can take action should the acceptable quality levels not be met.

There are a variety of methods available for this from simple documents; to more elaborate spreadsheets with graphs; to dashboards which are immediately presentable on every screen or even html pages online. The choice is enormous, but it doesn't matter how you get the message across as long as the message is received and understood by the people who need to see it.

Identify what metadata to report

Users of a dataset often need to know a little bit more about the data than what can be identified from the data itself. It's often useful to identify how the data was created, when it was created, when it was last updated and who created it. They may also need to know what quality levels have been achieved for the data. All of this can be reported in the metadata.

Further guidance on this can be found in the [ISO 19100 Guidelines] and in the ISO 19115 and 19157 standards.

Identify and implement a regime for maintaining quality

Identifying how to maintain the quality of a dataset once it goes under maintenance is perhaps the most difficult aspect of all. How will you ensure that the people and processes working on the data will not degrade it below the acceptable quality levels? This phase of the work should not be underestimated.

There are many ways of ensuring that quality is delivered as required. Experiences within the QKEN suggest that those errors that can be trapped by software are the ones that are easy to resolve and that those that rely on humans are the most difficult.

People

How will you ensure that your production team is capable of producing data of the right quality? What training / checking will you put in place at the start of production? How much of their work will you look at and at what point do you start looking at less? What is your criteria for just having an occasional look at their work? If their work is found to be below the standard required, what is your process for getting them back up to standard?

Once a process has been developed, how are new joiners tested? Do they go through the same process as those who started at the beginning of production? (They probably should.)

Process

Is the process that has been developed capable of meeting the user's needs? For example, there is no point in creating an update cycle of three years if the acceptable quality level for currency is two years. Does the update process maintain all the aspects of the data that it is required to?

What is your process for updating the documentation / specification? It is unlikely that you will have made a perfect release before production began and there will be clarifications and updates. How do you ensure that all producers get to see the updates?

Do you have processes for fixing errors? Evidence shows that people make fewer mistakes when they have to fix their own errors as it becomes a learning opportunity.

Is there a documented process for fixing improving or correcting data following feedback from customers?

Technology

This is the easiest area to ensure that you have the correct quality. There are a number of well-developed technologies and methods that can be deployed to ensure that the data is correct. These include:

- Coded domains
- Key constraints
- Automatic validation routines that prevent the data being passed to the next stage in the process. These could include schema, topology and other validation routines covering thing such as unlikely combinations, valid ranges
- Schema validation

6. Risks to a successful implementation

Failure to get management support

The biggest single risk to a successful implementation is the failure to get support from management. Getting the right quality requires an investment in both time and money and without support from the senior stakeholders, the implementation is likely to fail.

Failure to understand the use cases

Failing to understand the uses cases is likely to lead to incorrect assumptions being made. This in two will lead to incorrect acceptable quality levels being set which will ultimately mean that the dataset is not capable of meeting the users needs.

Losing sight of the user

Any production flowline is likely to involve a number of steps, often covering involving different production areas. Those closest to the user are likely to understand their requirements better than those at the end of the flowline yet decisions on the data will be made right through the flowline. It's important that all aspects of the flowline understand what the use cases are so that everyone involved in making decisions on the data, know what is important to the user.

Testing

Whilst testing is vital to ensure that the correct data is supplied, a balance has to be struck between the amount of testing and the time it takes. Some of the risks with testing include:-

- Checking takes too long. This delays the release of the data and, depending on the type of testing, may add cost
- Checking too many things. This can result in a product that exceeds the requirements of the user and is likely to lead to increased costs and time delays.
- Failure to check the right things. The risk here is that something that is important to the user is not checked to the level required leading to the use case not being met.
- Duplicate testing. This is an additional cost on the flowline that can also have the impact of causing time delays.

Unclear documentation, processes, training

Having clear documentation, processes and training is vital to ensure that the production team has the correct information prior to data capture. Errors as a result of poor documentation, processes or training could possibly be trapped at the testing stage but only after the cost has been incurred. Preventing these from happening in the first place is far better than having to fix them later.

7. Conclusion

This document has provided a high-level overview of the what a Data Quality model is and how to implement it. Following its advice will provide the user with a dataset that is maintained to the right quality, at reduced cost and is able to meet the users needs. Further guidance on these matters can be sought through the Quality Knowledge Exchange Network or in the references below.

8. References

ISO 19100 Guidelines Document produced by the EuroGeographics Quality Knowledge Exchange Network on how to interpret and implement the ISO 19100 series of standard relating to quality.

ISO 19115 International standard on Metadata for Geographic Information

ISO 19157 International standard on measuring the quality of Geographic Information