Geospatial data quality training for the South African Spatial Data Infrastructure – Lessons learnt from training geospatial data custodians

Antony K Cooper [1,2], Nicolene Fourie [3], Serena Coetzee [2,4], Marinet Blom [2], Maroale Chauke [5] & Vutomi Ndlovu [5]
[1]Smart Places, CSIR, Pretoria, South Africa: acooper@csir.co.za
[2]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa: marinetblom@gmail.com
[3]Next Gen, CSIR, Pretoria, South Africa: nfourie@csir.co.za
[4]United Nations University Institute for Integrated Management of Material Fluxes and of Resources (UNU-FLORES), Dresden, Germany: serenacoetzee@unu.edu
[5]NSIF, Department of Agriculture, Land Reform and Rural Development (DALRRD), Pretoria, South Africa: maroale.chauke@dalrrd.gov.za ,vutomi.ndlovu@dalrrd.gov.za

DOI: https://dx.doi.org/10.4314/sajg.v14i1.8

Abstract

Standards play an important role in achieving the objectives of a spatial data infrastructure. However, standards can be difficult to understand and implement for those with limited exposure to them. The South African Spatial Data Infrastructure (SASDI) aims to facilitate the capture, management, maintenance, integration, distribution and use of spatial information. To decide whether a SASDI data set is fit for a specific purpose, users need information about its quality. SANS 19157:2014, Geographic information – Data quality, specifies how the quality of geospatial data can be described and assessed. The Committee for Spatial Information (CSI), responsible for implementing SASDI, identified the need to train geospatial data custodians in implementing SANS 19157. While custodians were eager to learn, several barriers prohibited presentation of training in a 'traditional' classroom setting. These barriers included the costs and time to travel from remote areas of the country to a training venue and challenges with scheduling the training at a time suitable to all participants. Online training was therefore delivered − however, structured in a way to overcome general 'online fatigue' after the pandemic. In this paper we present our experiences in presenting training on SANS 19157 to professionals responsible for geospatial data sets. We also share the lessons learnt from the novel structure for online training.

Keywords: *standards, data quality, geospatial data, spatial data infrastructure, training, South Africa*

1. Introduction

South Africa's Spatial Data Infrastructure Act established the Committee for Spatial Information (CSI) to oversee the implementation of the South African Spatial Data Infrastructure (SASDI), "*the national technical, institutional and policy framework to facilitate the capture, management, maintenance, integration, distribution and use of spatial information*" [South Africa 2003]. The Directorate: National Spatial Information Framework (NSIF) within the Department of Agriculture,

Land Reform and Rural Development (DALRRD) is the Secretariat for the CSI tasked with supporting the CSI, facilitating policy and legislation development and implementing the technical platforms for SASDI. One objective of the SDI Act is to facilitate access to, and the sharing of, geospatial data sourced from public bodies [Fourie, 2023].

Standards play an important role in achieving the objectives of an SDI. However, standards can be difficult to understand and implement for those with limited exposure to them. To decide whether a geospatial data set (such as in SASDI) is fit for a specific purpose, users need information about its quality. SANS 19157:2014, *Geographic information – Data quality*, is a South African National Standard (SANS) and is identical to ISO 19157:2013, *Geographic information – Data quality*. These standards establish the principles for describing the quality of geospatial data, including types of quality, evaluation procedures and quality measures. There appears to be a lack of accessible training materials on ISO/SANS 19157 and on understanding the sharing and curating of data, so the European Open Science Cloud (EOSC) Association has called for "*training material to create a common ground understanding of what quality is*" [Lacagnina *et al.,* 2022].

A decade ago, for the Mapping Africa for Africa (MafA) initiative, in collaboration with the United Nations Economic Commission for Africa (UNECA) and the International Cartographic Association (ICA), the Chief Directorate: National Geo-spatial Information (NGI) in DALRRD issued a tender for a guideline *on standards for best practice for the acquisition, storage, maintenance and dissemination of fundamental geospatial data sets*. The guideline was published [Coetzee *et al.,* 2014] and converted into a wiki hosted by the ICA [ICA & ISO/TC 211, 2024]. While the guide and wiki provide a good overview of ISO 19157, it was not developed with the intention of providing training material.

During 2023, DALRRD awarded a tender for the *Digital Training Manual and Virtual Data Custodian Training for SANS 1957* to the University of Pretoria and the CSIR. DALRRD issued the tender to support the CSI. The training was provided online and was aimed at the geospatial data custodians within SASDI but was also made available to others within the public sector. As the training was online, it enabled participation from across South Africa, including by local municipalities that are largely rural.

This paper shares our experiences and findings from conducting the online training sessions. The remainder of the paper is structured as follows: the next section provides background and context on SASDI, data quality and standards for geospatial data quality. Section 3 presents the training design, while section 4 discusses the training implementation, together with lessons learnt and recommendations for future trainings, before the paper is concluded.

2. Background and context

2.1. The South African Spatial Data Infrastructure (SASDI)

Generally, the CSI has met quarterly, though much of its work has been done through subcommittees. Currently, the CSI has three subcommittees, each of which has three working groups (WGs): Subcommittee on Governance (Governance and Institutions; Policy and Legal; and Financial WGs), Subcommittee of Technology (Data; Innovation; and Standards WGs) and Subcommittee on People (Partnerships; Capacity and Education; and Communication WGs).

The Regulations for the SDI Act require that the CSI identify and publish a list of relevant national and international standards for geospatial information [DRDLR, 2017]. These lists have been published: DALRRD [2023] and DALRRD [2024].

In practice, the data custodians are those responsible for the base (or fundamental or core) geospatial data sets (base data sets) for SASDI [Chauke *et al.,* 2021]. The CSI has identified ten themes for these, which include 40 different geospatial data sets [CSI, 2020]. Currently, all the custodians are national departments and state-owned entities. Local governments were identified as potential base data set custodians for high-resolution imagery, cadastral and transport-related data. However, these custodianship responsibilities are still to be formalized by the CSI, the responsible base data set coordinators and the relevant authorities. According to the SDI Act,

data custodian means-

(a) an organ of state: or

(b) an independent contractor or person engaged in the exercise of a public power

which captures, maintains, manages, integrates, distributes or uses spatial information [South Africa, 2003].

Section 6.1.7 of the CSI's *Base Data Set Custodianship Policy* specifies what the custodians are meant to do to ensure that their base data sets are of a suitable quality [CSI 2015]. The implications of these requirements are examined in Table 1. Based on this, we recommend that this policy be updated to align with SANS 19157 and other relevant standards.

Table 1. Implications of the Base Data Set Custodianship Policy [CSI 2015]

Section 6.1.7. Quality	Implications
(a) The base data set custodian will ensure that the base data set is accurate and current concerning the determined user needs for the purpose for which it was captured. Where probable errors exist, the degree of probability of its correctness must be made available.	While this refers explicitly to only accuracy and currency (part of temporal quality), presumably, it is meant to encompass all the elements of quality in SANS 19157. This also requires the custodian to produce a quantitative report on quality that is hopefully statistically valid.
(b) The base data set custodian will ensure that the base data set or spatial information is free from ambiguities.	This requires a sensible specification for the data set and adherence to it, such as according to SANS 19131:2012, *Geographic information – Data product specifications*.

(c) Base data set custodians will ensure that the quality and resolution of their base data sets and other spatial information meet the needs of their intended users.	This matches the usability element of SANS 19157.
(d) Base data set custodians of specific base data sets will ensure that base data set updates are sent to base data set custodians of derived data sets.	This could be done automatically by SASDI. A key issue is the incremental updating and versioning of the data sets [Cooper & Peled 2001], whereby changes to this base data set affect others.
(e) Base data set custodians of derived data sets should ensure that their data is derived from the latest base data sets [CSI 16 Feb 2015].	The problem of incremental updating and versioning is more acute, as external changes can affect the integrity, quality and spatial referencing of the data set [Cooper & Peled, 2001].

One complication is that a base geospatial data set can consist of features and attributes that have separate custodians. For example, for the cadastre, the Chief Surveyor General is responsible for the land parcels (the geometry) and the Chief Registrar of Deeds is responsible for the ownership records (attributes) [CSI, 2020].

2.2. Data quality

Beauty is in the eye of the beholder [Hungerford 1878]

Quality is in the eye of the user.

The quality of any resource (such as data) depends on who will use it and for what purpose. For example, the details of a tree (species, age, canopy size, etc.) are more important for a forester or ecologist than for a civil engineer routing a road; or the positional accuracy for a stream needs to be better for a local planning application than for a regional development analysis. This aligns with Sections 6.1.7(a), (b) and (c) in Table 1. Hence, only the user can really understand the context for using the resource, and thus the desired and actual quality of the resource, to balance aspects such as being comprehensive *vs* being timely *vs* being accurate *vs* being free [Ashley, 2013].

As with metadata, the quality of a resource can be explicit or inferred. Further, the documented or reported quality is only part of the inherent quality of a resource, be it data, product, service, process, transaction, operation, etc. "*The inherent quality can be obvious or subtle; easy or difficult to comprehend or describe; crucial or practically irrelevant; and qualitative or quantitative, or both*" [Cooper, 2016]. Hence, only a subset of the inherent quality gets to be assessed, measured or estimated and then reported, because of limitations such as cost and complexity. As a data set approximates the universe of discourse (the real or hypothetical world that is of interest), so the reported quality approximates the inherent quality [Cooper, 2016].

The producer (or custodian) of the resource needs to document it sufficiently for the user to be able to assess the resource. Such documentation covers the metadata of the resource and the assessment of its quality against documented criteria. So, quality has two key aspects:

- *Truth in labelling*: the producer must be truthful in identifying the quality of the data.
- *Fitness for use*: the prospective user must evaluate the quality against their needs [Moellering, 1985].

Unfortunately, users can be unaware of inaccuracies inherent in the data [Zargar & Devillers, 2009], such as the limitations of GNSS (global navigation satellite systems) receivers and other tools; or being unable to distinguish between an error and a distortion. One common error made in South Africa is transposing coordinates, because the coordinate values for latitude and longitude are similar for a large part of the country [Cooper, 2016]. For example, the intersection of Thabo Mbeki Drive and Buchanan Street in Lichtenburg is at 26.152°S and 26.152°E.

The user also needs to consider the quality of the abstract model used for creating a data set as that determines what should be included in, and excluded from, the data set. Aesthetics can also be part of quality, and quality an aspect of aesthetics [Morita, 2015]. Most metrics of data quality are surrogates for the quality measures that really matter, such as truth [Ashley, 2013].

2.3. Standards for geospatial data quality

The International Organization for Standardization Technical Committee, ISO/TC 211, *Geographic information/Geomatics*, developed ISO 19157:2013, *Geographic information – Data quality*. SANS 19157:2014, *Geographic information – Data quality*, is identical to ISO 19157 and was published by the South African Bureau of Standards (SABS). SANS 19157 specifies seven elements (or components or dimensions) of data quality: logical consistency, completeness, positional accuracy, thematic accuracy, temporal quality, usability and metaquality.

ISO 19157:2013/Amd 1:2018, *Geographic information – Data quality – Amendment 1: Describing data quality using coverages*, has not yet been adopted by the SABS. However, this amendment will only become important for the base geospatial data sets when their custodians have established their data quality assessment processes and then wish to report on the results by a coverage (polygons or areas of interest). An example of this would be to assess the quality of a data set in only one province and not for the whole of South Africa.

Because ISO 19157-1 was still under development when the training was designed, the training could not be based on it. However, ISO 19157-1 does not introduce significant changes to the dimensions and components of data quality in SANS 19157. The changes were detailed in the manual to ensure that the training manual will remain current when ISO 19157-1 is adopted for South Africa.

There are several challenges that can affect quality that might not be surfaced when applying a standard such as SANS 19157. These include dependence on the purpose and context (which can go beyond the specification), bias by the producer (that can be masked by detailed but false metadata) and qualitative aspects (such as the weather affecting what gets logged in the field). Further, those who are not involved in developing a standard might not accept or understand it correctly [Cooper *et al.,* 2011].

Ariza López *et al.* [2020] suggest that some elements of quality are ambiguous or not well defined, such as usability being described as "*an aggregation of producer's measurements and conformance to requirements*". They suggest that usability could be determined by feedback from users. Further

problems with implementing ISO 19157 are the availability of metadata, ignorance of ISO 19157 and access to the standard as it is not available for free [Ariza López *et al.,* 2020].

ISO/TC 211 has published over 90 model-driven, integrated standards [Parslow & Jamieson, 2024], including others relevant to geospatial data quality. SANS 1878-1:2011, *South African spatial metadata standard Part 1: Core metadata profile*, is the South African implementation of ISO 19115:2003, *Geographic information – Metadata*, which specifies Lineage for SANS 19157 and for the training. Subsequently, ISO 19115-1:2014, *Geographic information – Metadata – Part 1: Fundamentals*, has become the metadata standard and has been adopted as SANS 19115-1:2016 but has not yet been adapted for South Africa. A revision of SANS 1878-1 is required for this. ISO 19131:2022, *Geographic information – Data product specifications*, uses the concepts and elements of ISO 19115-1 to describe a user's requirements for a data set, making it easy to compare the metadata of a data set with the user's requirements.

3. Training design

3.1. Learning outcomes

The training was designed to achieve three learning outcomes:

1) The participant appreciates the importance of assessing and documenting the quality of their geospatial data set.
2) The participant can specify the relevant quality elements of their geospatial data set and how the elements should be assessed.
3) The participant can prepare a data quality report based on simple measurements of selected relevant elements of the data quality of their geospatial data set.

The first learning outcome was mainly addressed on the first day and included the importance of knowing what the base geospatial data set is; the standards related to SANS 19157; an overview of evaluating data quality in accordance with SANS 19157; the CSI, SASDI and standards generating organisations; and an overview of a standard and how to read it. Participants were instructed to study the manual before their first training day, and they had to complete a multiple-choice assessment at the end of the first day. The second learning outcome was addressed on the second day, while the third learning outcome was achieved through an assignment, with a due date after the training.

3.2. Workflow for evaluating data quality

Before a custodian can assess the quality of their geospatial data set, they need to know what the data set is – or at least, what it is meant to be:

- What is the context for creating and using the data set? Is it a legislated mandate or something else? This will determine what expertise, funding and other resources are

available to create, maintain, update and disseminate the data set, and hence what can realistically be expected of the data set.

- What is the data set? What is the specification for the data set (hopefully there is one!) and how is the specification documented (as a text file, or using a formal specification, such as SANS 19131)? What are the feature and attribute types in the data set? How were the data collected (in the field, from existing data sets, from modelling, etc.)?
- Has the data set been comprehensively described through metadata, whether as text files or using a formal specification, such as SANS 1878-1 or SANS 19115-1?
- Who are the main users of the data set? What do they use the data set for? Does the custodian engage with them regularly? What are their expectations of the data set? Does the custodian know who the other users are? Does the custodian know of any unexpected uses of their data set?

To ensure the above are considered when assessing quality, we added two steps to the SANS 19157 workflow for evaluating data quality (Figure 12 in SANS 19157): knowing what the data set is and then what can be invested in assessing its quality. See Figure 1 (our additions are highlighted in red).

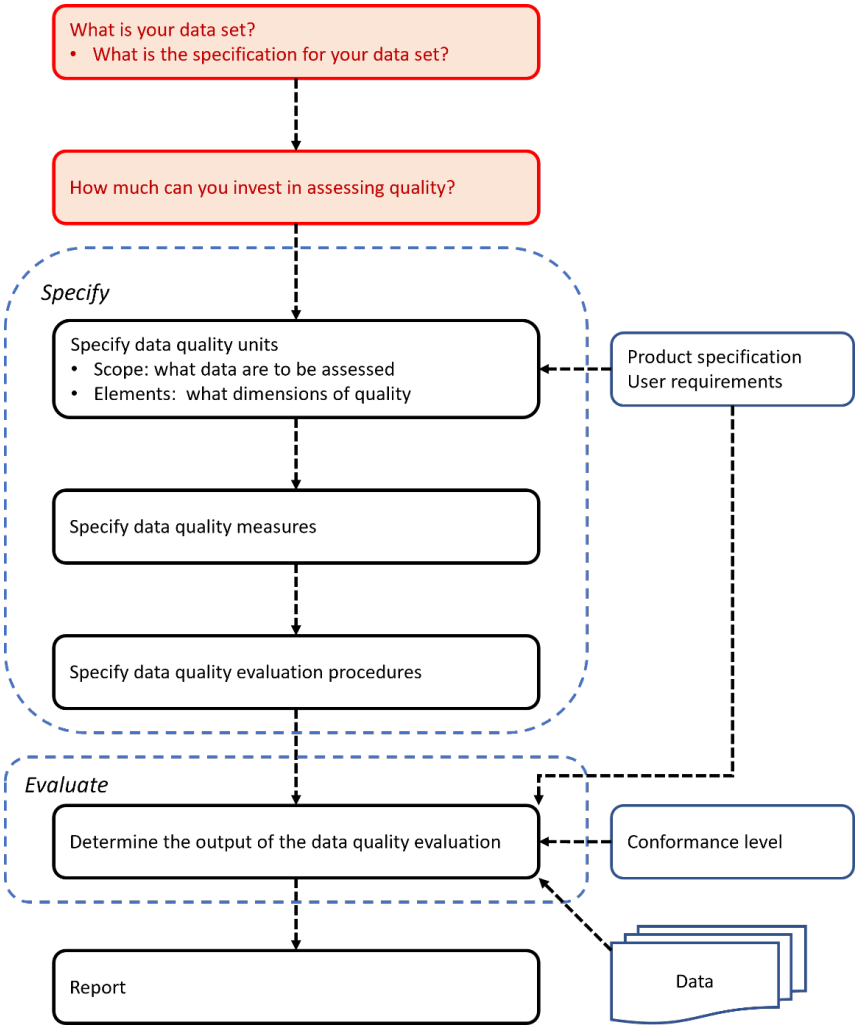

Figure 1, Evaluating data quality, adapted from SANS 19157: Figure 12

3.3. Training material

Based on the workflow above, a set of training materials aimed at achieving the learning outcomes was developed. An iterative approach was followed to ensure the manual had a broad reach and application scope. Training materials, based on feedback from the participants, were updated during the training. For example, one participant requested the documents be provided in a larger font to help those with limited vision. As some participants were uncertain about the requirements for the second assignment, we prepared and shared a rubric to show how the data quality report would be marked. There were also requests for a template and/or examples of quality reports, but we did not provide these to allow delegates to think creatively through the process of preparing a quality report.

Key elements of the training materials are provided below.

Training Manual, Part 1: SANS 19157:2014, Geographic information – Data quality covers the background, introduction to data quality and SANS 19157; implementing SANS 19157; and examples of assessing data quality, data quality reporting, and conformance. These examples were implemented in two geographical information systems (GISs), ArcGIS® Pro [ESRI, 2024] and QGIS

[OSGeo, 2024], selected because they are widely used. The examples provide detailed steps for assessing each of the seven elements of quality, plus lineage, which is specified in the metadata standard, ISO 19115 [2003]. Finally, a glossary of over 90 terms, covering SANS 19157, data quality in general, and the CSI and SASDI, assisted participants to understand the specific meanings a term may have in a standard that might not be the most common usage of the term.

Training Manual, Part 2: SANS 19157:2014, Geographic information – Data quality – Examples includes further examples for all the elements of quality, focusing on specific features and attributes in test data sets, with screen shots to illustrate the steps and results.

Slide Pack, Day 1 introduced SASDI, data quality and the test data sets. It provided an overview of SANS 19157 and the structure of a standard. Then, it went through all the elements (dimensions) of data quality in detail and introduced the data quality measures.

Slide Pack, Day 2 began with a review of the first day of training, followed by a detailed look at the data quality measures, with selected examples from the manual. Next were details of the data quality evaluation procedures, including how to extract random samples and ensure representativity. It concluded with an introduction to the assignment to produce a data quality report covering the assessment of some quality elements for a suitable data set.

Training data sets were prepared from diverse samples of vector and raster data with metadata from four data custodians: the Agricultural Production Health and Food Safety and Disaster Management Branch of DALRRD, the Western Cape Government, the City of Johannesburg and the South African National Biodiversity Institute (SANBI). These were generally subsets and intermediate versions of the base data sets and are now out of date. We also made a few changes to the data to highlight different aspects of the elements of quality and to ensure coverage of the different types of geospatial data, and incorporated errors into the test data sets to facilitate learning about data quality assessments. The training data sets do not, therefore, reflect the actual quality of the base data sets provided by these custodians and do not reflect the quality of their work.

For the data quality report assignment, participants were required to submit a standalone free text data quality report containing results of evaluating any three quality elements – it seems that few organisations are implementing all the elements [see, for example, Petrosyan *et al.* 2024 and Vukalić *et al.,* 2024]. We advised participants to use one of their own data sets because they would be familiar with it, have ready access to who specified and created the data set, and have access to the documentation on the data set. However, this was not feasible for some of the participants, so they were allowed to use one of the test data sets.

3.4. Learning management system

Training materials and data sets were uploaded to the learning management system (LMS), Moodle, to make them available to the participants. The multiple-choice assessment was managed through Moodle, which presented each participant with the questions and possible answers assembled

randomly. The participants could retake the assessment repeatedly, to reinforce their learning. The participants submitted the second assignment through Moodle, consisting of the data set and its data quality report so that they could be compared, when necessary.

4. Implementation, lessons learnt and recommendations for the future

4.1. Implementation

The training was advertised widely, including through the NSIF's mailing list; the mailing list of the statutory body, the South African Geomatics Council (SAGC) and by the SABS. Invitations were sent to at least 552 people. The training was much anticipated, with the result that the training was oversubscribed.

Ten online training sessions were delivered via Microsoft Teams to geospatial data custodians from 31 October 2023 to 7 December 2023. 140 people registered for the training via the registration link, and 20 registered via email after the sessions had been booked to capacity. Initially, the capacity was set at 10 participants per session. Due to the additional requests received, the attendance capacity was increased to 15, then finally to 20, to compensate for no-shows. Some participants had to reschedule their sessions due to changing work commitments, but the increased capacity meant they could be accommodated. While increasing the capacity did not disrupt the training sessions, it created additional administrative complexities.

Each online training session ran over two days. Each morning involved presentations by one of the training team and interactive discussions with the participants. The first afternoon was allocated to a multiple-choice assignment and the second afternoon to starting a practical exercise on implementing some quality elements from SANS 19157. On both afternoons, the team was available online to help participants with queries, including accessing the test data sets and training materials, and with setting up their GISs. The team could also be contacted afterwards to help with any problems.

Training support was provided on MS Teams and through email, in the hour before each morning session (to help participants get set up), during the session and through to 16:30. Thereafter, post-training support was available via email until the end of January 2024. Some participants also phoned for support. We used feedback from the participants to keep improving the training materials, to deal with aspects that were confusing and to make them more accessible.

It quickly became clear that the participants would need much more time than we had expected, for completing the practical exercises and assignments. The due date for all attendees for both assignments was extended until 14 January 2024. The attendees could redo the multiple-choice assignment as often as they liked, as the emphasis was on the learning; not on the grade achieved. They also had ample time to understand and complete the assignment to assess a data set and produce a data quality report, with the post-training support available as needed. Nevertheless, some

participants had problems with the volatility of their work programmes and end-of-year work pressures.

An online feedback session was held on 29 January 2024 and an online feedback survey using Google Forms was then circulated to all attendees. The results are discussed under lessons learnt.

Finally, a training video was prepared for the whole course, essentially a presentation of the slide packs for Day 1 and Day 2. The actual training sessions were not recorded, to preserve privacy.

Of the 153 people who attended part or all of the training, 50 completed the training, 52 attended (but did not complete the assignments) and the remaining 51 participants attended only parts of the training.

4.2. Risks

Initially, there was a concern that there might be limited support or even resistance to the training. Fortunately, this was not the case with much support for the training from across the community. Several potential risks were identified at the start, but most of them did not materialize. For example, rather than being unwilling to participate due to fear over the quality of the geospatial data being exposed publicly, some custodians welcomed the opportunity. The key problems are outlined in Table 2.

Table 2. Risks and how they were addressed

Lack of access by custodians to copies of ISO/SANS 19157	DALRRD had a licence with the SABS for several standards (including SANS 19157) for the base data set custodians. However, the licence had expired and could not be renewed by the time the training started. Thus, we ensured that the SANS 19157 content relevant for the training was included in the training manual. The assignments tested the content of the training manual (not SANS 19157). Note that many of the participants were from public organisations not covered by the licence, so DALRRD should consider extending its coverage or engage the South African Bureau of Standards (SABS) to enable temporary access to published standards for capacity-building purposes.
Stakeholders not attending after agreeing to participate in user community workshops	This was a common occurrence, unfortunately. To follow-up on no-shows, participants were contacted on their contact number and then via email. In some cases, the contact number and email address were incorrect. In other cases, participants were unable to join due to unexpected loadshedding. The risk was minimised by oversubscribing the number of participants in the later sessions. From the feedback questionnaire, it was also evident that some attendees were interrupted by their supervisors or line managers during the training. We were not aware of this. For the future, a signed off development plan and/or physical attendance at training workshops could prevent this.
Too many participants in a workshop	This risk did not materialize as we increased the number of participants that could register for each training session, and this did not impact on the experience of the participants.
Electricity (load shedding) and internet (connectivity) problems, including firewalls or policies preventing use of webinar software	Participants with problems due to firewalls or other policies were assisted by the team member to find a workaround and attend. Some delegates experienced connectivity and loadshedding problems during the sessions for which they were subscribed. These participants were given an opportunity to join again when they could on the training day or join another session to obtain the knowledge required and to fulfil the certification requirements. In addition, participants could contact the facilitators at any stage of the training period to ask questions and resolve any queries. Further, a participant who missed part of the training could obtain a copy of the recording and training materials from NSIF.

4.3. Lessons learnt

The positive response to the training was not only of interest to geospatial data custodians but also to other organizations, such as National Treasury and private sector companies. This was confirmed during the feedback session on 29 January 2024, in the responses to the feedback questionnaire and from comments at the National Control Survey Imagery and Mapping and Cadastral Information Consultative Forum on 6 February 2024. We also received requests for dedicated training sessions for individual organizations (e.g., National Department of Transport, or for a group of organizations, such as the Base Data Coordinator and Geospatial Data Custodians) responsible for a specific theme. This shows that more SANS 19157 training sessions are needed, and additional training should be more specific in the implementation of SANS 19157. Training is also needed on other standards.

The LMS, Moodle, worked well, except that some delegates wished to submit large data sets for the second assignment, which exceeded the space limits of the Moodle license. Instead, they were asked to email the files. In future, submitting just the data quality report should be sufficient, unless a Moodle license allowing more storage is bought. In future, one could consider using other LMS platform functionalities, such as announcements or more complex tests.

Delegates used the post-training support to clarify content and to solve technical problems. The feedback enabled us to refine and adjust subsequent training sessions. Some attendees preferred to send an email or make a phone call, instead of going back into the MS Teams call that stayed open until 16:30 in each session. This shows that delegates need different modes of communication to get support. The key issues dealt with during the support were:

- Most of the support was needed for logistics issues, especially access to Moodle, and access to MS Teams for a particular session, generally because of corporate firewalls.
- Access problems due to limited internet bandwidth, sometimes caused by loadshedding.
- Accessing specific functions in the GIS the participant was using, such as the plugins used for some of the examples.
- Understanding and interpreting the theory, training materials and examples, which was the main purpose for providing the training support.
- Interpreting and submitting the multiple-choice assignment.
- Interpreting the requirements for the data quality report (second assignment) and submitting it, especially when deciding on a suitable data set to use and submitting assignments that were too large to upload to Moodle.

We were pleased to find that participants adopted diverse formats and content styles for their reports, some of which used their organisation's corporate style. This provides the CSI with several sensible alternatives that they can use to develop guidelines or templates for the data quality reports

the custodians should submit to SASDI. The lack of a systematic approach to compiling the reports was observed in some submissions.

The submitted data quality reports did not always use SANS 19157 terminology and standardized names for quality measures. Access to a copy of SANS 19157 might have improved the situation, though many of the terms were defined in the glossary in the training manual. This does indeed emphasise the fact that standards are difficult to understand because of the need to be precise.

Despite the extension, some delegates ran out of time to complete Assignment 2. However, some submitted reports of very high quality, quite a few achieving a mark between 90% and 100%. Our impression after marking Assignment 2 is that everybody learnt something new, which was also evident from the feedback questionnaire. For some, it was their first opportunity to produce a data quality report according to SANS 19157. For others, they could improve on their reports. After completing the training, they had a better understanding of what quality reporting entails and the benefits of a quality report accompanying their data set and metadata. There was also improved appreciation for using quality reports internally, before a data set is shared or published. For the future, one could consider that the second assignment be completed in a group, particularly by the team responsible for a specific base data set.

The above shows that while the training was about the SANS 19157 standard, delegates acquired knowledge and skills about data quality generally. Such a training opportunity does not necessarily enforce conformance to standards in a top-down fashion but enables delegates to understand and appreciate the principles of data quality and reporting, based on the data quality principles embedded in SANS 19157. This confirms the value of the good or best practice documented and embedded in any standard.

A feedback questionnaire was circulated to participants, to which we received 23 complete responses. Key insights are:

- Half of the participants work for an appointed geospatial data custodian and a quarter expect such an appointment soon.
- Two thirds are registered with the South African Geomatics Council or another professional body.
- Over two thirds spend more than half of their working day on technical geospatial data and related tasks, with only an eighth being mostly involved in managerial and/or administrative work.
- For less than a third, this was their first exposure to standards for geospatial data.
- Respondents appreciated the training manual, especially the examples, and the discussions during the training.

- The assignments were useful for strengthening and reinforcing the theory presented on the first day, and for helping them to understand how a data quality report should be prepared.
- Most delegates strongly agreed they have a better appreciation of the importance of assessing and documenting the quality of a geospatial data set.
- While most strongly agreed the training prepared them adequately to apply the SANS 19157 standard in their professional contexts, fewer were convinced they had confidence in their ability to implement SANS 19157 in their organization. All delegates agreed that they would recommend the training to colleagues or peers.

Selected challenges identified by respondents:

- Loss of connection during loadshedding.
- Not having direct interaction with other participants.
- One was interrupted during the virtual training by their supervisor, and another found it impossible to complete the assignments on account of their heavy workload.
- Training should not be offered in the busy time of November to January.
- Selected suggestions from respondents to improve the training:
- Provide a one pager or cheat sheet that helps one to read and navigate through the standard.
- Provide good and bad examples of a data quality report, or just an example of a data quality report.
- Present the training as a hybrid session or as a physical contact session.
- Consider sharing videos of the training on YouTube so that they can be shared with colleagues.
- Similar training for other standards is needed.
- A WhatsApp channel or group where people can support each other is needed.
- Consider having a team that can provide support or even audit data sets on request, so that the participants can learn how to improve.
- The training team was commended for making a "*boring (boring does not mean unimportant) topic easy to understand and follow*".

Coetzee *et al* [2018] reported on the results of a survey that included investigating motivators and barriers for implementing quality standards. Some of these aspects correlate with those highlighted by the participants, such as the complexity of assessing quality, standards being difficult to read and understand, the costs of standards, and lack of expertise and funding. The motivators included wanting better quality for their data sets. Other aspects from the survey that could be considered by the CSI and the custodians are that implementing SANS 19157 can help meet compliance issues with

international agreements, facilitate the integration of geospatial base data sets together, reduce liability costs and attract more users. Standards also serve to implement good practices and incorporate the wisdom of all those who contributed to developing the standards [Coetzee *et al.,* 2018].

4.4. Recommendations

Based on the above, we would recommend the following to the CSI:

- The license with SABS should be renewed and extended so that geospatial data custodians have access to South African national standards for geospatial information. This would also give them access to the numerous examples of data quality measures in Annex D and the example of a full data quality report in Annex E of SANS 19157.
- The CSI should consider developing guidelines for data quality reports by custodians of data sets, particularly those to be included in SASDI.
- Additional SANS 19157 training sessions should be provided to selected custodians or on request. For example, some private companies are interested in the training.
- Advanced training sessions are needed to cover SANS 19157 implementation in more depth: more time should be provided to explore more detailed examples assessing all the elements of data quality. Longer courses (perhaps over five days) or as a two-day introductory course, followed later by a three-day advanced course, could then be presented.
- Training sessions could also be dedicated to a specific theme, such as hydrology or transport, or to a specific base geospatial data set. For example, Vukalić *et al.* [2024] found that quality can vary between different environments (e.g., urban, rural or mountainous regions).
- It might be useful to present the training in hybrid mode (online and in-person) to accommodate those with financial constraints to travel, while allowing the benefits of physical contact and networking.
- Training sessions on other standards prescribed by the CSI should be provided, as there is an interconnectedness between these standards.
- The CSI should develop a strategy for supporting custodians, not only through training but also by establishing interactive communication channels, such as via a WhatsApp group or support desk.
- Section 6.1.7, Quality, of the CSI's Base Data Set Custodianship Policy [CSI 2015], should be updated to correlate with SANS 19157, specifically to include all the quality elements.

The South African Statistical Quality Assessment Framework (SASQAF) is to improve the quality of all official statistics [StatsSA 2010]. Many of these statistics depend on geospatial data. Thus, the

CSI should ensure that SANS 19157 and SATS 19158:2022, *Geographic information – Quality assurance of data supply*, are embedded in the SASQAF processes. Equally, given the significant similarities and the need to integrate geospatial data and statistics, the integration should be reciprocal. SASQAF has prerequisites for quality, with eight indicators for the minimum requirements in the institutional and organisational context for producing statistics of a good quality − as in the legal mandate and appropriate policies − and in maintaining privacy. SASQAF then has eight dimensions of quality: relevance, accuracy, timeliness, accessibility, interpretability, comparability and coherence, methodological soundness, and integrity [StatsSA 2010].

Finally, several South African standards need to be updated or revised. Specifically:

- SANS 19157 needs to conform to ISO 19157-1.
- SANS 1878-1 needs to conform to ISO 19115-1 and its amendments. There are also some improvements to be made.
- SANS 19131 needs to conform to ISO 19131 and ISO 19115-1.

5. Conclusions

In this paper, we report on the findings made from presenting an online training course on SANS 19157 :2014, *Geographic information – Data quality*, to custodians of geospatial data sets. This paper covers the background to the Committee for Spatial Information and the South African Spatial Data Infrastructure, data quality and standards. It then describes the learning outcomes, workflow for evaluating data quality, the training materials developed and the implementation of the training. This paper concludes with a review of the risks, the lessons learnt and recommendations for the future.

The response to the training was overwhelmingly positive, evident from the large number of participants who registered for the training, from the responses during the feedback session, the feedback questionnaire and comments made at other meetings. The training increased awareness of geospatial data quality (and to some extent, metadata), helping key staff members among the custodians to understand the details of SANS 19157 and how to implement them. Key problems were loadshedding, firewalls (impeding access to the training materials) and supervisors who took participants off the training for other tasks.

The main contributions of this paper are:

- Aligning the CSI's Base Data Set Custodianship Policy with SANS 19157.
- Enhancing the workflow in SANS 19157 to include the context for the creation and assessment of a data set.
- Encouraging SABS TC211 to adopt *ISO 19157-3, Geographic information – Data quality – Part 3: Data quality measures register*, as soon as it has been finalised.

- Acknowledging the lessons learnt from the training such as the need for more (and more advanced) training on SANS 19157 and the need for training on other standards needed for SASDI.
- Making detailed recommendations to the CSI to develop guidelines and templates for quality reports; to providing ongoing support; and to renew and extend the licences for accessing standards.

Finally, several South African standards need to be revised.

6. Acknowledgements

We are grateful to organizations which provided test data sets as part of the training material: the Department of Agriculture, Land Reform and Rural Development, the Western Cape Government, the City of Johannesburg and the South African National Biodiversity Institute (SANBI). We appreciate the guidance of the Project Steering Committee (PSC) and their constructive comments on the training material and pilot training sessions. We would like to thank DALRRD for providing us with this opportunity to conduct training on SANS 19157. Finally, to all the delegates who made time in their busy schedules to attend the training and to submit the assignments – without you this training would not have taken place. A presentation of the project and the results was made to the Gauteng branch of GISSA on 14 March 2024.

7. References

Ariza López FJ, González PB, Pau JM, *et al* (2020) Geospatial data quality (ISO 19157-1): evolve or perish. *Revista Cartográfica* 129–154. https://doi.org/10.35424/rcarto.i100.692

Ashley K (2013) Data quality and curation. *Data Science Journal* 12:GRDI65–GRDI68. https://doi.org/10.2481/dsj.GRDI-011

Chauke M, Fourie N, Ndlovu V & Moema Y (Dec 2021) Implementing Base Data Set Custodianship – South Africa. *30th International Cartographic Conference (ICC 2021)*, Florence, Italy. https://doi.org/10.5194/ica-abs-3-51-2021.

Coetzee S, Behr F-J & Cooper AK (6 Feb 2018) Implementing geospatial data quality standards – motivators and barriers. Second *International Workshop on Spatial Data Quality*, Valletta, Malta. https://eurogeographics.org/app/uploads/2018/06/4-SDQ2018_Coetzee_V1e.pdf

Coetzee S, Cooper AK & Rautenbach V (13 Jun 2014) Part C: Standards for fundamental geo-spatial data sets. In: Clarke, DG (ed) (2014) *Guidelines of Best Practice for the Acquisition, Storage, Maintenance and Dissemination of Fundamental Geo-Spatial Data sets*, Mapping Africa for Africa (MAfA), 124p. United Nations Economic Commission for Africa (UN ECA). Also: ISO/TC 211 document N 3805. http://hdl.handle.net/10204/11702

Cooper AK (2016) *An exposition of the nature of volunteered geographical information and its suitability for integration into spatial data infrastructures*. PhD thesis, University of Pretoria, South Africa. http://hdl.handle.net/2263/57515

Cooper AK, Coetzee S, Kaczmarek I, Kourie DG, Iwaniak A & Kubik T (May 2011) Challenges for quality in volunteered geographical information. *AfricaGEO 2011*, Cape Town, South Africa, p. 13. https://researchspace.csir.co.za/dspace/handle/10204/5057

Cooper AK, Coetzee S & Kourie DG (Oct 2012) Assessing the quality of repositories of volunteered geographical information. *GISSA Ukubuzana 2012 Conference*, 2-4 October 2012, Kempton Park, South Africa. https://researchspace.csir.co.za/dspace/handle/10204/6377

Cooper AK & Peled A (Aug 2001) Incremental updating and versioning. *20th International Cartographic Conference (ICC 2001)*, Beijing, China. https://icaci.org/files/documents/ICC_proceedings/ICC2001/icc2001/file/f19007.pdf

CSI (16 Feb 2015) *Base Data set Custodianship Policy and Policy on Pricing of Spatial Information Products and Services, made in Terms of the Spatial Data Infrastructure Act, 2003 (Act 54 of 2003)*. Committee for Spatial Information (CSI). Government Gazette 38474. https://www.gov.za/documents/notices/spatial-data-infrastructure-act-base-data-set-custodianship-policy-and-policy

CSI (30 Nov 2020) *A list of base data sets and corresponding Base Data set Coordinators and Custodians*. Committee for Spatial Information (CSI). http://www.sasdi.gov.za/sites/SASDI/Acts%20Policies%20and%20Procedures/List%20of%20appointed%20base%20data%20set%20custodians%20-%2030%20November%202020.pdf

DALRRD (10 March 2023) *Spatial Data Infrastructure: Spatial information standards*. Department of Agriculture, Land Reform and Rural Development (DALRRD). Government Gazette 48187.

DALRRD (14 Jun 2024) *Spatial Data Infrastructure Act: Spatial information standards*. Department of Agriculture, Land Reform and Rural Development. Government Gazette No. 50825.

DRDLR (27 Oct 2017) *Spatial Data Infrastructure Act: Regulations*. Department of Rural Development and Land Reform (DRDLR), Pretoria, South Africa. Government Gazette 41203.

ESRI (2024) *ArcGIS® Pro, The world's leading desktop GIS software*. ESRI, Redlands, CA, USA. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Fourie, N. (2023). Public bodies compliance to PAI and SDI Act: An enabler for geospatial information freedom. *Abstracts of the ICA*, 6, 65.

Hungerford, MW (1878) *Molly Bawn*. Smith, Elder & Co, London.

ICA & ISO/TC 211 (2024) Standards, *ICA Wiki. International Cartographic Association (ICA) and ISO/TC 211, Geographic information/Geomatics*. Available at: https://wiki.icaci.org/index.php?title=Standards

ISO 19115:2003, *Geographic information – Metadata*. International Organization for Standardization, Geneva, Switzerland.

ISO 19115-1:2014, *Geographic information – Metadata – Part 1: Fundamentals*. International Organization for Standardization, Geneva, Switzerland.

ISO 19131:2022, *Geographic information – Data product specifications*. International Organization for Standardization, Geneva, Switzerland.

ISO 19157:2013/Amd 1:2018, *Geographic information – Data quality – Amendment 1: Describing data quality using coverages*. International Organization for Standardization, Geneva, Switzerland.

ISO 19157-1:2023, *Geographic information – Data quality – Part 1: General requirements*. International Organization for Standardization, Geneva, Switzerland.

Lacagnina C, David R, Nikiforova A, Kuusniemi M-E, Cappiello C, Biehlmaier O, Wright L, Schubert C, Bertino A, Thiemann H, Dennis R (2022) *Towards a data quality framework for EOSC Authorship Community*. EOSC Association. https://hal.science/hal-04017152

Moellering H (Jan 1985) *Digital cartographic data standards: an interim proposed standard. Report 6*, National Committee for Digital Cartographic Data Standards. 164p.

Moodle (2024) *Moodle LMS / Small (Version 4.2.2 LMS)*. Moodle Pty Ltd, Western Australia, Australia. https://www.moodle.com/

Morita T (2015) Evolution of Concepts in Ubiquitous Mapping. *27th International Cartographic Conference (ICC 2015)*, Rio de Janeiro, Brazil. https://icaci.org/files/documents/ICC_proceedings/ICC2015/papers/7/979.html

OSGeo (2024) *QGIS, a Free and Open Source Geographic Information System*. The Open Source Geospatial Foundation (OSGeo), Beaverton, OR, USA. https://www.qgis.org/

Parslow P, Jamieson A (2024) 30 years of geospatial standards: ISO/TC 211 celebrates three decades of standardizing geographic information. *GIM International* 28–30. https://www.gim-international.com/content/article/30-years-of-geospatial-standards?sid=40485

Petrosyan M, Piloyan A, Efendyan P (2024) Enhancing Quality Control Standards for Armenia's National Spatial Data Infrastructure: A Python-based Approach with Emphasis on Road Spatial Data Layers. *Abstracts of the ICA* 7:1–2. https://doi.org/10.5194/ica-abs-7-125-2024

SANS 1878-1:2011, *South African spatial metadata standard Part 1: Core metadata profile*. South African Bureau of Standards, Pretoria.

SANS 19115-1:2016, *Geographic information – Metadata – Part 1: Fundamentals*. South African Bureau of Standards, Pretoria.

SANS 19131:2012, *Geographic information – Data product specifications*. South African Bureau of Standards, Pretoria.

SANS 19157:2014, *Geographic information – Data quality*. South African Bureau of Standards, Pretoria.

SATS 19158:2022, *Geographic information – Quality assurance of data supply*. South African Bureau of Standards, Pretoria.

South Africa (2003) *Spatial Data Infrastructure Act* (Act No 54 of 2003).

South Africa (21 May 2010) *Appointment of members of the Committee for Spatial Information, by the Minister of Rural Development and Land Reform*. https://www.gov.za/documents/notices/spatial-data-infrastructure-act-committee-spatial-information-appointments-21-may

South Africa (2013) *Spatial Planning and Land Use Management Act* (Act No 16 of 2013).

StatsSA (2010) *South African Statistical Quality Assessment Framework (SASQAF). Second edition*. Statistics South Africa, Pretoria, South Africa, p. 92. https://www.statssa.gov.za/standardisation/SASQAF_Edition_2.pdf

Vukalić A, Triglav Čekada M, Petrovič D (2024) OpenStreetMap data quality assessment according to ISO 19157-1:2023 for Slovenia and Bosnia and Herzegovina. *Abstracts of the ICA* 7:1–2. https://doi.org/10.5194/ica-abs-7-182-2024

Zargar A, Devillers R (2009) An operation-based communication of spatial data quality. In: *2009 International Conference on Advanced Geographic Information Systems & Web Services*. IEEE, Cancun, Mexico, pp 140–145.