OPENAIR

Operational Network of
Air quality Impact Resources

Best Practice Guide

BP402 | Manage and analyse data

Data labelling for smart air quality monitoring

Introduction

Data requires a wide range of supporting information to be useful, shareable, and reusable. Supporting information about data is often referred to as 'metadata' (i.e. data about data). Metadata can end up being more complex and detailed than the original data that it describes. Data labelling is the process of defining, collecting, and applying the metadata needed for your project. This OPENAIR Best Practice Guide chapter provides guidance on data labelling for smart air quality monitoring, assisting you to define and collate information about your data. Effective data labelling allows you to store, retrieve, interpret, and share data in a productive, safe, and secure manner.

Who is this resource for?

This Best Practice Guide chapter is intended for use by many of the parties involved in an air quality sensing project. It should be especially relevant to data producers, managers, and users, including:

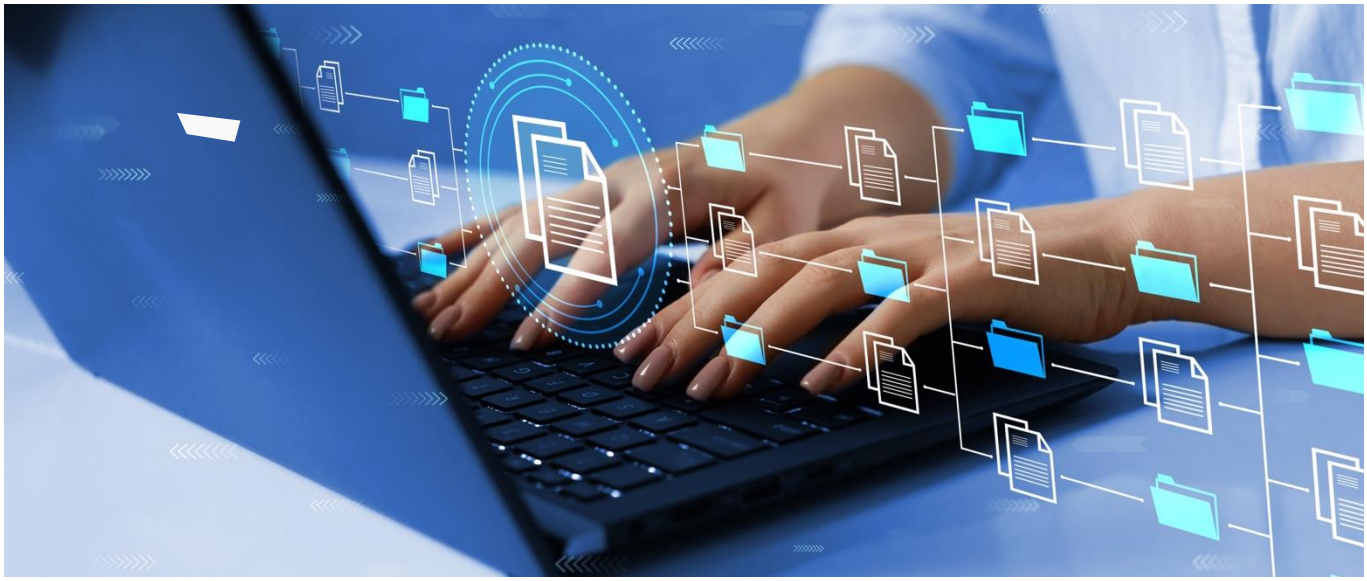
- people leading air quality monitoring projects
- smart city professionals
- information and communication technology professionals
- data analysts and custodians.

How to use this resource

This chapter is a high-level introduction to data schemas and data labelling. You should engage with it during the initial design stage of your project to develop an understanding of the importance of these processes, and to develop some general strategies to address them.

There are two main tasks introduced in this chapter:

1) **Creating a data schema:** a *data schema* clearly defines the telemetry and metadata fields that you will use for your project.
2) **Creating a master metadata record:** a *master metadata record* collates and tracks metadata field entries for all your devices and data sources. The creation and completion of a metadata record is a process referred to as *data labelling*.



Key terms

Metadata

Metadata is ‘data about data’, and is defined by ‘fields’, each of which describes a specific attribute of your data (or other aspects of your project). Each metadata field needs to serve a clear purpose that is tied to your data use case and the operation of your network. If you are delivering an air quality monitoring project, you will need to engage with metadata, regardless of the scale at which you are operating. Good metadata design will help to ensure that you are able to manage, retrieve, and utilise your data in the most effective way.

Telemetry

Telemetry refers to all dynamic information reported by a device, and includes sensor data as well as device functionality variables (such as battery voltage and communications signal strength). Telemetry can also include data from third-party sources, such as regulatory air quality monitoring stations, or weather reports. Telemetry values are dynamic and can change every time a device reports. They can be viewed as an archival data set, or as a near-real-time data stream.

Data schema

Metadata and telemetry fields are generally characterised in a *data schema* that defines their intended applications, validated field entries, and data formats. Each smart sensing project should develop its own data schema.

Data labelling

The practical application of a data schema within the context of a project is referred to as *data labelling.* This involves the creation of a Master Metadata Record, where entries are made against each field.

Creating a project data schema

A data schema is a document that provides detailed information about the data you will collect, how it will be collected, and what you can do with it in the context of a particular project. If you are setting up an air quality monitoring network, it is highly advisable that you develop your own specific data schema.

TIP: Develop your data schema collaboratively

Development of a data schema typically requires input from a range of project participants and partners, including those responsible for infrastructure, communications, asset management, and administrative and legal divisions within your organisation.

As you co-develop this data schema, you should keep in mind the following factors: when data is generated, standards and quality assurance measures, plans for sharing data, ethical and legal issues or restrictions on data sharing, copyright and intellectual property rights over data, data storage and backup measures, and data management roles and responsibilities.

The structure of a data schema document

A data schema lists and characterises all the telemetry and metadata that will be used in your project (see Figure 1). A data schema can be separated into a telemetry schema (which summarises data sources), and a metadata schema (which lists and characterises metadata fields).

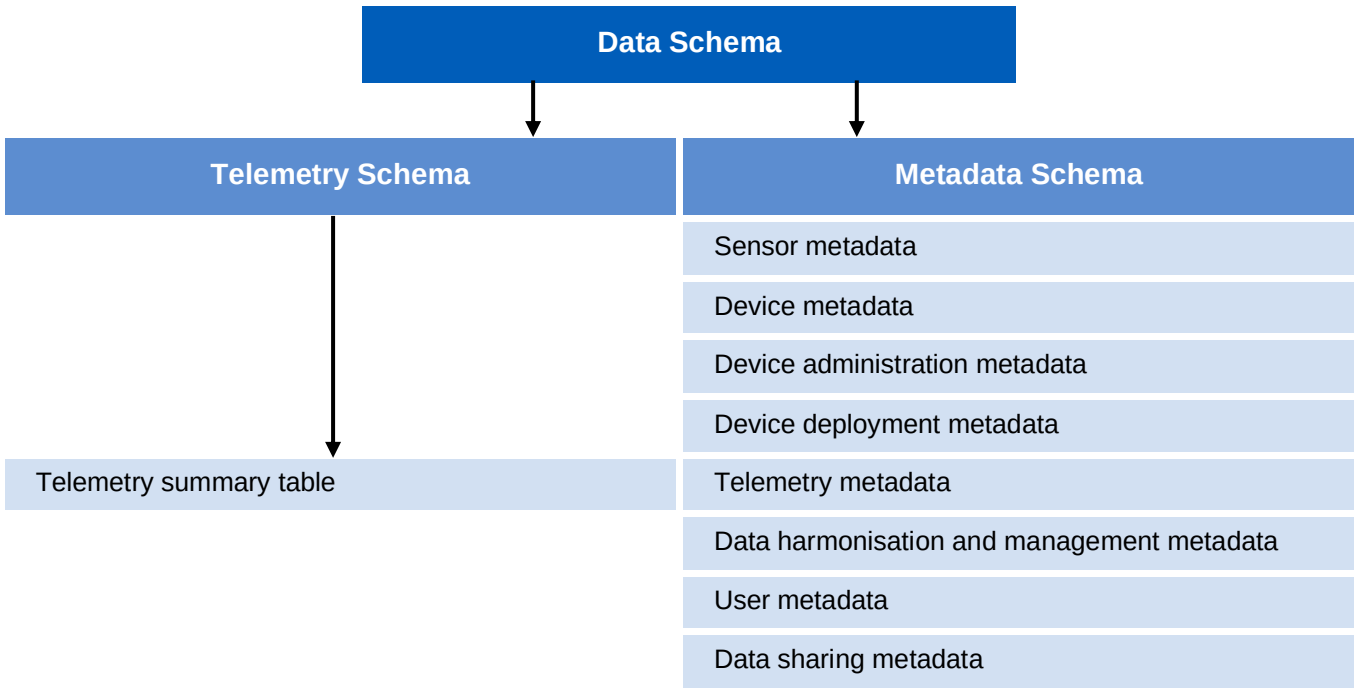


Figure 1. Data schema structure showing telemetry and metadata schemas and fields

The benefits of creating a project data schema

Table 1. The benefits of creating a project data schema

Benefit	Description
Plan metadata capture to support your data use case	A data schema helps you to consider all the metadata fields that are critical to your data use case. By doing this during the design phase of your project, you can ensure that you collect what you need in one go, supporting efficient and timely workflow.
Support device management and system operations	Device management and system operations require a range of contextual information (e.g. when and where a device was deployed, how it was installed, and by whom) that must be captured in metadata fields. Development of a data schema allows you to systematically consider which fields you will need to support these activities.
Support data interpretation	Well-planned data interpretation is supported by a data schema that characterises *all* your project's data. For example, your data use case might require interpretation of other data sources (e.g. regulatory air quality or weather reports) alongside your sensor data. A data schema also tracks contextual information (metadata) that helps you interpret your sensor data.
Support data storage and accessibility	A data schema defines the format for all data in your database, and will assist with establishing your database structure and management. It also informs the development of query, retrieval, and edit functions that make finding and working with data easier. For example, you might specify a parameter of interest, the location and name of a sensor, and a period of interest. By accessing this metadata, the user can retrieve exactly what they need from a larger data set.
Support visualisation and user interface design	A data schema provides a reference for displayable information that can be used by software engineers and user interface designers tasked with customising user interfaces and visualisations.
Support data sharing and platform integration	If you are using a stand-alone sensing solution, you may still want to integrate a live data feed with another platform (e.g. a data feed to an in-house, enterprise-scale smart city platform; or a data feed to an external data-sharing portal). In such cases, a data schema would be used by developers to support the design of the integration or interface between the two platforms.
Support technical staff to design, implement, and operate an integrated technology solution	A data schema provides a reference to any technical staff or contractors tasked with delivering and operating a complete integrated technology solution. The more sophisticated and bespoke your technology solution is, the more critical a data schema will be, because it establishes a common blueprint around which an integrated system is built.

I'M USING A STAND-ALONE SOLUTION – DO I STILL NEED A DATA SCHEMA?

Smart air quality monitoring networks can take many forms. Some are large, customised, functionally sophisticated, and comprise multiple integrated technologies, services, and platforms. In these cases, a detailed data schema is an essential piece of core documentation. But what if you are doing something much simpler?

An IoT (Internet of Things) Platform hosts and manages all of your devices. There are many stand-alone solutions for air quality monitoring that bundle devices together with an IoT platform and data storage, easily allowing you to access, visualise, and download your data via a single website. In this instance, you may wonder if you still need to create a data schema for your project.

Our advice is that you do. Here's why:

Stand-alone IoT platforms will log a small amount of 'fixed-in-platform' metadata, and may also support the integration of certain third-party data sources (e.g. weather data). However, you are unlikely to have control over what these fields and data sources are, and they are unlikely to meet all your project needs. A data schema can exist entirely separate to your sensor platform, and may be vital for a variety of reasons (see Table 1).

Practical advice for creating a data schema

The following critical project planning and design activities provide a foundation for the development of a data schema:

1. Ensure that you have developed a business case and a data use action statement (see the OPENAIR *Identify template*). This ensures that you have a clear understanding of your primary data use case. Most of your telemetry and metadata must serve this use case.
2. Ensure that you have engaged with your data users (both within your organisation as well as external stakeholders) to help you define a data schema that targets their needs.
3. Confirm (if possible) the complete technology solution for your project (such as sensing devices, communications, and platforms). This will heavily define a large amount of your metadata.
4. Develop an operational plan for your monitoring network, including an approach to device management, maintenance, service contracts, communications, and platform services. Operations can be thought of as 'secondary data use cases' that require a range of supporting telemetry and metadata. Make sure to include operational staff in your list of data users.
5. Ensure that you have developed a complete sensing device deployment plan (see OPENAIR Best Practice Guide chapters *Sensing device deployment planning: high-level design* and *Sensing device deployment planning: detailed design*).
6. Develop a data policy and/or data management plan that serves the needs of your smart sensing project and informs your approach to data management, access, and sharing.

Telemetry in your data schema

Telemetry refers to all dynamic information reported by a device, and includes sensor data as well as device functionality variables (such as battery voltage and communications signal strength). Telemetry values can change every time a device reports and can be viewed as a static/archival set, or as a near-real-time data stream.

Creating a telemetry summary table

A telemetry schema (which is part of a data schema) should feature a summary table of all telemetry sources that will be used in your project.

Table 2 provides an example of such a table. This example is a 'hybrid network' that features two different models of low-cost sensing device. Data from nearby ambient regulatory air quality monitoring stations is also being acquired or 'ingested' by the network. Note that the purpose of the table is to characterise incoming telemetry streams in the formats they are received (i.e. it shows the averaging period for a telemetry stream as it is received from the data source). For instance, ingested $PM_{2.5}$ data from the ambient regulatory station will have a 60-minute averaging period[1].

Table 2. An example of a telemetry summary table for all data sources used in a project

	Data sources		
Telemetry	**Temperature /humidity device X**	**Particulate device X**	**Ambient regulatory station**
Temperature - ambient (°C)	10 min. averaging		60 min. averaging
Relative humidity - ambient (%)	10 min. averaging		60 min. averaging
Temperature - internal (°C)		15 min. averaging	
Relative humidity - internal (%)		15 min. averaging	
PM_1 [2]		15 min. averaging	60 min. averaging
$PM_{2.5}$		15 min. averaging	60 min. averaging
PM_{10}		15 min. averaging	60 min. averaging
Voltage - battery (V)	10 min. averaging	15 min. averaging	
Voltage - solar (V)		15 min. averaging	
Timestamp	Yes	Yes	Yes

[1] For an explanation of averaging periods, please refer to the OPENAIR Best Practice Guide chapter *Data interpretation: correction and harmonisation*.
[2] PM (particulate matter) refers to airborne solids or liquids. Its size is measured in micrometres and is indicated by the subscript. E.g. PM_{10} has a diameter of 10 micrometres or less. (NSW Health, 2020)

INTERNAL VS AMBIENT TEMPERATURE AND HUMIDITY

Internal and ambient readings are not the same. It is important that you recognise this, and correctly label your telemetry for these readings in your data schema.

A temperature and humidity monitoring device featuring a Stevenson shield. Image source: UTS

Internal temperature and humidity readings are taken by sensors inside the housing of a device. These readings are vital for correctly interpreting air quality readings, but they do not accurately represent *ambient* conditions. During warmer conditions, the temperature inside a device housing can be elevated by several degrees above ambient levels (due to poor air flow, or the device housing heating up in direct sunlight). Under cooler conditions, residual heat generated by device electronics can also elevate internal temperatures by a degree or more.

Ambient temperature and humidity readings are taken by sensors that minimise localised effects created by a device. This generally involves positioning them within something called a Stevenson screen, which maintains ambient temperature and humidity while protecting equipment from the weather. Ambient readings provide an accurate representation of ambient environmental conditions, and are critical if you are studying urban heat. Some smaller sensing devices are entirely housed within a Stevenson screen, which means that internal readings are also ambient readings. Larger devices tend to mount a Stevenson screen on the outside of the main housing.

When a device delivers temperature or humidity telemetry, you must ascertain whether this is internal or ambient. Many low-cost devices will report internal readings only, and you should be careful to label these accordingly. More sophisticated devices that feature Stevenson screens may report both internal and ambient readings.

TIME AND DATE FORMATS

Any smart sensing device will report data with an associated timestamp (containing time and date). A timestamp is a type of telemetry that requires special consideration. There are several different ways that sensing devices can report time. There are also several ways that you might choose to record time in your database. Your telemetry schema should characterise the time format of each incoming data source, as well as stipulate a single common time format for data storage.

The three most common time and date formats are:

1. Coordinated Universal Time (UTC) is the primary time standard by which the world regulates clocks and time, and is the basis of time zones.
2. Local time (e.g. Australian Eastern Standard Time) is expressed relative to UTC as +/- hours.
3. Unix time (sometimes called 'Epoch time') is a standard computing timestamp that measures time by the number of seconds elapsed since 1 January, 1970. A variation on Unix counts in milliseconds.

Devices may report time in different ways, including UTC, local time, or Unix time. The standard approach for the Internet of Things (IoT) is to report and store in Unix seconds, and convert to local time for dashboard visualisation. Shared data should always include Unix time, even if local time or UTC are also included, and all time formats should be clearly labelled.

Defining telemetry metadata

Your data schema should characterise each type of telemetry using a telemetry metadata table.

Each type of telemetry requires a range of supporting information (metadata) that defines what it is, and how it is reported and interpreted. A data schema should include a telemetry metadata table (see Table 3) for each type of telemetry that you are capturing.

Table 3. An example of a telemetry metadata table

	Description of metadata field	Example
Telemetry type	Name used to refer to telemetry when displaying or referring to data in human-readable contexts.	Particulate Matter (2.5μm^3)
Description of telemetry	An open text descriptive field that explains telemetry.	The mass of suspended particles (measuring ~2.5 microns) within 1m^3 of air.
Telemetry abbreviation	The abbreviation used to refer to this type of telemetry. Often used for data visualisations.	PM2.5
Name in database	The name used to refer to this type of telemetry in your database.	pm_2.5
Unit of measurement	The units used to express telemetry data	Micrograms per cubic metre (μ/m^3)
Decimal places	The number of decimal places used to express telemetry	2
Reporting interval	The period of time between data reports made by a device	15 minutes
Outer threshold, upper	Thresholds to define an upper and lower limit for what is physically possible for this type of telemetry. Values outside of these thresholds must be erroneous.	1000 μ/m^3
Outer threshold, lower		0 μ/m^3
Inner threshold, upper	Thresholds to define an upper and lower limit for what is *expected* for this type of telemetry in your deployment context. Values above this threshold are possible but should be reviewed.	500 μ/m^3
Inner threshold, lower		0 μ/m^3
Humidity correction	A factor used to correct raw particulate matter data relative to ambient relative humidity	3%

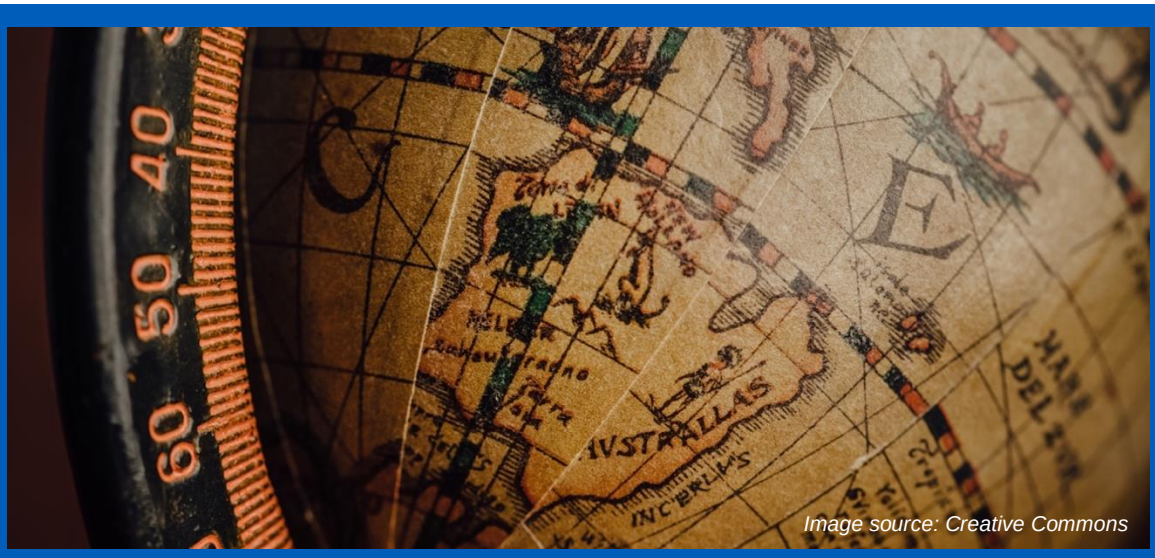


Image source: Creative Commons

LATITUDE AND LONGITUDE FORMATS

Latitude and longitude are coordinates that define a location on a two-dimensional map. They are a fundamental aspect of any geospatial data set, allowing you to associate a telemetry value with that location. There are three ways to record latitude and longitude:

DDD° MM' SS.S"	**Degrees, minutes, and seconds**
DDD° MM.MMM'	**Degrees and decimal minutes**
DDD.DDDDD°	**Decimal degrees**

The preferred format for entering latitude and longitude is decimal degrees (e.g. the decimal degrees location of the University of Technology Sydney is -33.88204, 151.20040).

LATITUDE AND LONGITUDE AS TELEMETRY…OR METADATA?

Are spatial coordinates telemetry or metadata? The answer is that they can be either.

If a sensing device features on-board GPS (as many do), it will report latitude and longitude as *dynamic* telemetry. A device can also have latitude and longitude recorded as *fixed* metadata fields. Any sensing device that is registered in a database and IoT platform must have a primary reference for latitude and longitude. This means that – in a situation where you have GPS telemetry *and* coordinates as metadata – you must choose one of these as your primary geospatial reference.

A GOOD RULE OF THUMB:

- If you are sensing in a fixed location, your primary geospatial reference should be manually entered metadata fields (even if your device reports GPS coordinates).
- If you are doing mobile sensing (e.g. devices mounted on vehicles), your primary geospatial reference should be GPS telemetry.

Coordinates reported by low-cost GPS are relatively inaccurate (particularly in Australia), and may indicate a point several metres from a true position. For fixed sensing, it is best to manually record latitude and longitude as metadata. Mobile sensing cannot use fixed metadata for coordinates, meaning that it must rely on less accurate GPS.

GPS coordinates for fixed sensors are still useful, and should be stored as part of your telemetry record. Their main use is for verifying the current location of a device as part of network operations. For large monitoring networks with many devices, it is possible for devices to be accidentally deployed in the wrong place, or recalled for maintenance. GPS data allows you to check where a device was when it last reported.

Creating a metadata schema

A metadata schema forms the majority of a data schema. It characterises all the telemetry that will be used in your project, as well as any other aspect of your project (e.g. hardware, data users, or services). Typical metadata fields relate to the following types of information:

- where, when, and how data was generated (e.g. details relating to a sensing device and its deployment, timestamp, and spatial coordinates)
- data interpretation applied (e.g. correction factors, averaging periods, and quality control)
- data management and storage
- data ownership, copyright, and intellectual property
- data access roles and responsibilities
- data sharing preferences.

Metadata is entered against fixed metadata fields. A field is a clearly defined information category that can have multiple entries made against it. Field entries may be updated manually (e.g. via a user interface or Application Programming Interface) or automatically in certain cases involving advanced data management (e.g. Artificial Intelligence applications).

Components of a metadata schema

A metadata schema should include a complete list of all metadata fields that will be captured and used for your project. Each field requires supporting information to characterise it (see Table 4). In effect, a metadata schema is a way to record metadata *about* the metadata.

A metadata schema typically involves a spreadsheet with metadata fields arranged in rows and columns to capture information about each field. Typical information includes:

- metadata fields (grouped into categories)
- a short description of each field
- the intended application or use case for each field
- data type (e.g. int, float, char, string, enum, date, bool, etc.)[3]
- formats for each field (e.g. units of measurement; number of decimal places)
- validated field entries (if required).

[3] Data types are a standardised set that constrain the format and possible values of a data field and relate to how that data will be used by a computer program. Click here for further guidance on data types.

Table 4. An example of supporting information for a single metadata field

	Description of metadata field	Example
Field name	The name that is used to refer to the field when displaying or referring to it in human-readable contexts.	Height above ground (metres)
Field name in database	The name used to refer to this field in your database.	height_above_ground_m
Field description	An open text descriptive field that explains what the field is.	The height that a sensing device is deployed above the ground immediately beneath it, measured in metres
Intended application	An open text description that explains how a field will be used.	Used to provide spatial context for data interpretation
Data type	A standardised type of data that constrains the format and possible values of the field, and relates to how that data will be used by a computer program.	Float
Unit of measurement	The unit of measurement used for entries in this field.	Metres (m)
Decimal places	The number of decimal places used for entries in this field.	2.0
Validations	A list of validated field entries used for a field.	*NA*

Categories of metadata

A smart air quality sensing project requires various categories of metadata, which should be captured in your metadata schema. Table 5 provides an overview of these, with examples of fields for each.

Table 5. A guide to all categories of metadata, with examples

Metadata categories	Description	Examples
Sensor metadata	Relates to the design, performance, function, and outputs of an individual sensor (a component of a device).	• Sensor performance metrics • Model details (make/model/version) • Sensor design/specific technology • Sensor telemetry outputs
Device metadata	Relates to the design, performance, function, and data outputs of a device.	• Device names • Make and model details • Device communications • Device telemetry outputs • Device calibration methodology • Device configuration
Device administration metadata	Relates to administrative details connected to devices, for asset management purposes.	• Owner (e.g. organisation, department) • Procurement details (e.g. date of purchase, cost, estimated lifetime)
Device deployment metadata	Relates to the physical deployment of devices.	• Power supply and mounting solution • Spatial context (e.g. height off ground) • Address (e.g. lon:lat, street, suburb, etc.) • Deployment status
Telemetry metadata	Relates to the formatting and interpretation of a specific telemetry stream.	See Table 3 for details
Data harmonisation and management metadata	Relates to the harmonisation and management of all data entering a database/platform. A single definition and approach to telemetry and metadata, decoupled from data sources.	• Data harmonisation and ontology alignments (e.g. units of measurement)

Metadata categories	Description	Examples
User metadata	Relates to users of an IoT solution or platform. May vary for different platforms/interfaces /databases.	• User identification (name, organisation, etc.) • Data management roles, access privileges, and responsibilities
Data sharing metadata	Relates to the decisions made about how data can be shared.	• Shareability status of the data • Data access settings • Ethical/legal restrictions • Intellectual property rights

Aligning with standards and best practice

Many aspects of data labelling are defined by standards or best practice documentation, which contain recognised terminologies and conventions that should inform the design of your own project data schema. By aligning with standards and best practice, you can ensure that your project data is recognisable and useable by others, which is critical for data sharing and interoperability.

A quick guide to relevant standards

As you develop your own project data schema, you might consider referring to the following types of standards.

Standards relating to devices

There are several standards or recognised best practice documents relating to the design of sensing devices that may influence the design of your data schema. These include:

- general hardware standards, such as IP rating (water and dust ingress protection)
- specific/recommended calibration methodologies (e.g. from a peer-reviewed paper, or government guide)
- standards that define sensor performance metrics
- communication technology standards (that vary by technology used, and relate to the type, quantity, format, security, and reliability of shared data).

Standards relating to device deployment

Standards exist for the deployment of regulatory ambient air quality monitoring devices for the purpose of capturing representative readings that can be applied to a wide surrounding area. These are the standards that are used for regulatory grade monitoring sites in your region, state, or country. If your smart sensing project aims to capture representative air quality for an area, then you may be referring to these types of standards, and you should align your metadata schema with them.

However, most use cases for smart low-cost air quality monitoring devices involve purposefully deploying them in locations that do not comply with such standards. In complex urban environments,

highly localised micro-climate effects are created, where pollution hotspots can appear for short periods, air movement is chaotic, and sensors are exposed to unmixed pollution sources (e.g. on busy roadsides). It is perfectly valid to study these spaces, but there are no standards relating to these types of deployment. For further advice on best practice deployment of smart low-cost sensors, please refer to the OPENAIR Best Practice Guide chapter *Sensing device deployment planning: high-level design*.

Standards relating to data sharing

One of the best existing references for data usability is the *Linked Data Rating* (LDR), which has been adopted by the Australian Government for assessing the quality of open data sets. If data sharing is a priority concern for your project, you should ensure that you align your data schema with the LDR. The rating assigns up to five stars to a given data release. Four- and five-star linked data ratings are awarded in cases where data sets are released with accompanying metadata that fully contextualises the primary data. LDR provides a benchmark for best practice data release by local governments.

The *FAIR Guiding Principles for scientific data management and stewardship* (2016) provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of data. The FAIR principles have seen widespread uptake across government, industry, and academia, and are particularly concerned with improving machine-actionability (i.e. the capacity of computational systems to find, access, interoperate, and reuse data with no – or minimal – human intervention). This is important in a smart city context, where we are dealing with large data sets, complex integrated systems, and increasingly automated data use. By aligning your project data schema with FAIR data principles, you can ensure that it is future-proofed for an expanding smart city sector.

A DATA ONTOLOGY FOR SMART AIR QUALITY MONITORING

A data ontology is a framework that defines all aspects of data labelling and management for a given topic. It is rather like a generic data schema that applies to a whole sector or type of activity, providing a common reference for all practitioners. A data ontology may be quite broad and non-specific (e.g. all local government smart city projects), or it can be quite domain-specific and detailed (e.g. smart air quality monitoring).

Creating a master metadata record

The creation of a data schema is a major but critical undertaking. The previous sections of this chapter should have convinced you of this and assisted you in the creation of a data schema for your project. The next important step is to create and complete a *master metadata record*.

A master metadata record is a direct reflection of your data schema that captures all the metadata field entries for your project. It should be a shared spreadsheet that logs metadata entries against all your devices, deployment locations, users, telemetry streams, and other entities – anything with metadata associated with it that appears in your data schema.

The initial role of a master metadata record is during the design and deployment of a new sensor network. During these early stages, there may be limited capacity to capture metadata in platforms and databases. There may be limitations imposed by a simple, stand-alone sensing platform that require you to manually collect much of your metadata. A more sophisticated platform solution may capture the majority of your metadata, but is likely to be under construction during early project phases, making manual metadata capture and management necessary. However, even sophisticated solutions that manage a large number of custom metadata fields rarely include *all* fields in your schema. Some information is only relevant to certain people, and does not need to be incorporated into a central platform (e.g. the contact details of the device installation contractor). A master metadata record captures *all* your metadata.

Designing a master metadata record

The design of a master metadata record is up to you. It is recommended that you use spreadsheet software, and that you structure your metadata record using various tabs. One approach is to divide metadata using tabs that correspond to your main metadata categories. Another approach is to divide metadata into operational views, such as the example shown in Table 6.

Table 6. An example of how to structure a master metadata record in a spreadsheet

Tab label	Description
Devices tab	Each row relates to a specific device and all metadata associated with that device across its lifetime. This view would be of most use to asset managers.
Deployments tab	Each row relates to a specific device deployment. If a device is ever recalled and redeployed, a new row would be created. This view can be useful for supporting more complex data management functions, and for long-term projects.
Locations tab	Each row relates to a device deployment location. This decouples locations from devices, meaning that a location can be associated with multiple devices at once, or over time. It may be useful for more complex or long-term projects.
Users tab	Each row relates to an individual user of the platform/database. Note: there may be ethical and privacy considerations relating to the capture of user data. It may be inappropriate to include all user data in a shared metadata record.
Validations tab	The place where all validated lists for metadata fields are stored.

A range of other tabs/views could be helpful, depending on your project. Consider who your data users are (including those tasked with operation and management of the monitoring network), and what information they are likely to require. Try to develop tabs/views on factors that relate to their specific needs, and remove information that they do not need.

AVOID DUPLICATION OF RECORDS: Use formulas to connect multiple spreadsheet cells to a single point of truth. There must only ever be one primary editable location for information within a master metadata record. You may wish to duplicate a field entry across multiple tabs, but these duplicates should automatically populate from a single primary instance. As soon as you have multiple editable instances of a single piece of information, you will create discontinuities.

Using a master metadata record

Keep it dynamic

Project metadata is dynamic, meaning that any static record of it will be a snapshot of a specific time. You can opt to make periodic saves of a master metadata record (adding a date to the file name), providing a comprehensive paper trail that can be very useful if you need to look back on past conditions (something that happens often). This is the simplest option, and it can be easily managed with respect to roles, access, and editing rights.

Another option is to use an online cloud-based file-sharing service that allows real-time, collaborative updates on a file. This turns your master metadata record into a dynamic document that should, in theory, be current and accurate. A challenge to note here relates to access and editing rights. You may wish to make the file 'read only' by certain users, and editable by a smaller subset of users. Make sure that your chosen file-sharing service can support this.

Instantiation of your metadata record and management of the 'source of truth'

Instantiation in computer programming means to create a concrete, working instance of an abstract record. Instantiation of your metadata involves creating a 'home' for it within a data platform, and shifting the primary 'source of truth'[4] for field entries to that platform.

All projects should create a master metadata record during their design and set-up phase. Many projects may go on to use that document as the source of truth throughout the operational lifetime of the monitoring network. However, certain more sophisticated systems may instantiate a majority of the record within their database, data model, user interfaces, and Application Programming Interfaces (APIs). This can, in theory, render a manual metadata record obsolete.

However, in reality, only a subset of your master metadata record would ever be instantiated in a data platform. This means that some form of ongoing manual record will always be required. Once you have developed a data schema, you should speak with your IT department or the software engineers working on your project to explore which elements of your project metadata should be instantiated, and where.

Any instantiated metadata should have a plan in place for managing field entry updates. You may opt to make updates in the manual record, and use it to update records in the platform – essentially maintaining the manual record as your point of truth. This is the simplest option, but it can get messy if platform users are able to edit entries via a user interface. You need to be clear about *where* the primary record is for all metadata fields, *how* it will be updated, and *by whom*. One option is to update your manual metadata record to exclude fields that are instantiated into your platform. This means that you track and manage non-instantiated fields in a manual record, and manage instantiated fields through your platform.

[4] A source of truth is a single source of information that is maintained as current and correct, superseding all others. Your source of truth may start out as a spreadsheet document, but later shift to being an instantiated record within a data platform or database (rendering the original spreadsheet obsolete).

References

NSW Health. (2020). *Particulate matter (PM10 and PM2.5)*. NSW Government. https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx

Additional resources

Microsoft* | *Common Data Model

Microsoft provides useful resources relating to a shared data language in the Common Data Model.

NSW Government* | *SEED website

The Central Resource for Sharing and Enabling Environmental Data (SEED) in NSW provides an excellent overview of metadata, and some essential metadata elements. Please see the SEED for more information. For details about the metadata standards used in SEED, please refer to *AS/NZS ISO 19115.1:2015 Metadata*, available at this **link**.

Associated OPENAIR resources

Best Practice Guide chapters

Data interpretation: correction and harmonisation

This chapter provides guidance for correction and harmonisation of data produced by smart low-cost air quality sensors. It introduces several types of correction factors that may need to be applied to raw sensor data, and explores how data formatting and labelling should be harmonised with a project data schema to support effective data management and sharing.

Sensing device deployment planning: high-level design

This Best Practice Guide chapter explores the high-level design of a smart air quality monitoring network. It provides general guidance for selecting where to deploy devices, what to mount them on, how to mount them, and how to support their operation.

Sensing device deployment planning: detailed design

This Best Practice Guide chapter explores the detailed design of a smart air quality monitoring network. It builds upon high-level design activities, and provides guidance for planning and documenting the details of specific device deployments.

Supplementary resources

Identify template

This template supports creation of a business plan and ‘data use action statement’ as strategic foundations for a smart low-cost sensing project.

Further information

For more information about this project, please contact:

Peter Runcie

Project Lead, NSW Smart Sensing Network (NSSN)
Email: peter@natirar.com.au

This Best Practice Guide section is part of a suite of resources designed to support local government action on air quality through the use of smart low-cost sensing technologies. It is the first Australian project of its kind. Visit www.openair.org.au for more information.

OPENAIR is made possible by the NSW Government's Smart Places Acceleration Program.

Document No: 20231026 Data labelling for smart air quality monitoring Version 1 Final