Best Practices for Interoperability for the Air Quality Community

Table of Contents

1. Introduction

1.1 Background

Effective air quality management requires the acquisition and analysis of a variety of types of data, such as atmospheric concentrations of various species from ground- and satellite-based instruments, information about emissions and emissions-generating activities, meteorological variables parameters, and other environmental characteristics. Moreover, because many important quantities cannot be measured at the needed times and places or with the needed quality, numerical models provide essential information. Various organizations, both government and non-governmental, collect and compile observational and modeled data at the local, national, and international levels. To better understand relationships among individual air pollutants or among problems that occur at different spatio-temporal scales, it is necessary to share, integrate, and analyze data from ambient monitors, satellites, and numerical models. However, finding, accessing, using, and sharing data from diverse sources is often difficult and resource-intensive. Interoperability practices shared across the community can improve access, use, and management of data from diverse sources to make air quality management easier, more accurate, and cost effective.

The air quality community is making progress in more synergistically sharing data and analytical capabilities, and thereby providing analysts, decision-makers, and researchers more insight and value from existing and future air quality data sets. These efforts are progressing toward an air quality “cyberinfrastructure” that aims to improve the use of measurement and modeled data in air quality science and management (Figure 1).

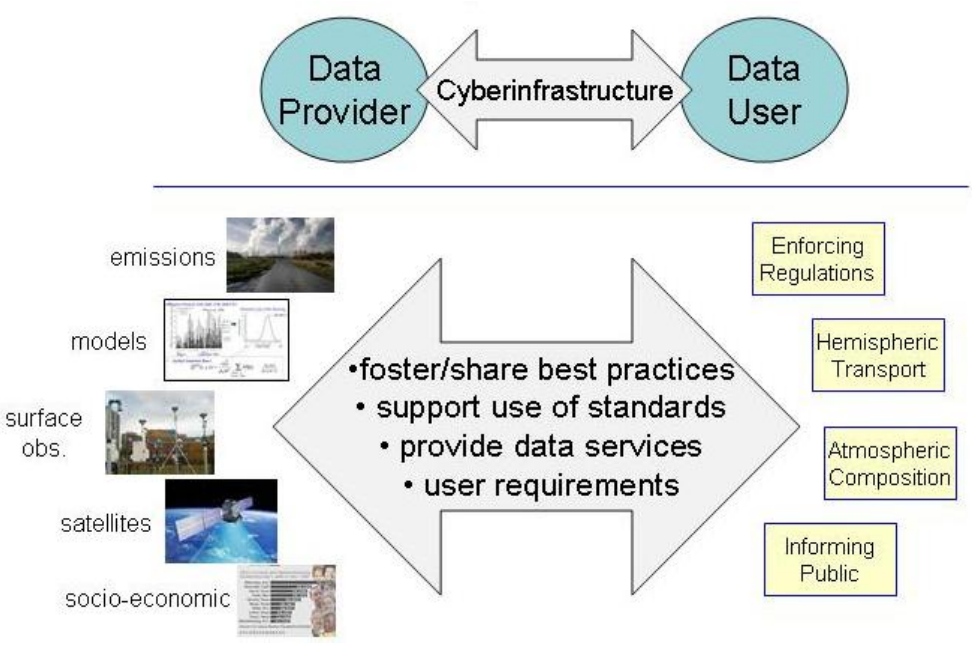

Figure 1. Cyberinfrastructure connecting data providers and users.

Measurement and modeled data are collected, processed, and provided by various organizations and systems. The provided data are used by a variety of organizations and

systems. In most cases, the data providers design their systems with specific end uses in mind. However, secondary end uses may be possible from these same data providers. The community-wide creation of a cyberinfrastructure that fosters adoption of standards for data formats, metadata, naming conventions, data publication, data access, and shared analyses strengthens the relationships between air quality data providers and data consumers. When the cyberinfrastructure achieves critical mass, the community is expected to gain in resource savings, improved decision support, effective collaboration among distributed teams, and improved coordination with international partners.

The air quality community has faced challenges in developing a cyberinfrastructure; these challenges have been identified in previous workshops and reports (2007 Data Summit and its follow-on ad-hoc committee activities, 2009 Federation of Earth Science Information Partners' (ESIP's) Summer Air Quality Meeting, and the 2009 GEO-VI Air Quality Meeting). More recently, the EPA-sponsored CyAir project identified the number one priority as the documentation of a set of best practices for the community to implement an air quality cyberinfrastructure. Documenting the best practices in current cyberinfrastructure efforts creates a shared understanding across the air quality community of how organizations can both advance their own cyberinfrastructure capabilities and use others' capabilities.

This document is a first version of a best practice document that covers areas such as naming conventions, data and metadata format standards, data access via Web services, and search and discovery methods. These best practices will evolve as the air quality community begins implementing and refining these practices. This document is intended to be used as a guidance or reference document. Its goal is to describe current practices across the air quality community in interoperability, present the information associated with following these best practices in an understandable manner, and identify and encourage, as appropriate, use of the currently common and/or preferred community practices to maximize the advancement of a cyberinfrastructure for air quality management.

1.2 Interoperability: Foundation of Cyberinfrastructure

1.2.1 Defining Interoperability

Interoperability generally refers to the ability of cyberinfrastructure systems or components to easily exchange information. Definitions of interoperability span a broad range; two of the most commonly cited definitions are

- "the ability of two or more systems or components to exchange information and to use the information that has been exchanged." (Institute of Electrical and Electronics Engineers, 1990)
- "the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units." (ISO/IEC 2382-01, Information Technology Vocabulary, Fundamental Terms)

In relation to the air quality community, interoperability is the ability to find, access, understand, and use data in a way that enhances collaboration across the community and increases the value of data and analyses in air quality research and management. Community-driven

standards for finding (metadata), accessing (Web services), understanding (naming conventions and metadata), and using (format standards, processing and analysis tools) data are the foundation for air quality interoperability.

1.2.2 Benefits of Interoperability

The ultimate goal of interoperability is to get the right information in the right context to the right person at the right place and time. From an information systems perspective, a core benefit to achieving interoperability is that it fosters the "network effect"—data from an increasing number of data providers can be effectively used by increasing numbers of data consumers at decreasing resource costs. These benefits are manifested in

- Easier access to data
    - Faster updates and automation
    - Quicker analysis and results
    - Use of more data in analyses/decisions
- More widespread use of air quality (AQ) data
    - Access to more comprehensive data sets and related data sets
    - Improvements in data quality
    - Improvements in metadata quality
    - Better decisions
- Improvements to data quality
    - Easier ways to synchronize data systems
    - More consistency in data across different data systems
- Better understanding of the capabilities of an AQ cyberinfrastructure
    - Confidence in the "system"
    - Willingness to build applications
- Simplified application development
    - Reuse and sharing of code and interfaces
    - Others develop user interfaces and lower development cost

Overall, the expectation is that decisions made from the data and derived analyses would improve because the processing leading to decisions would have better, more appropriate, and more complete data.

In a low-resource environment, interoperability lowers development costs because it facilitates sharing and reuse of data, analytical algorithms, tools, and code because they are developed based on common standards and frameworks.

1.2.3 Example of Interoperability Impacts

The potential effect of interoperability on data exchange and analytical processes is illustrated through the following use cases focused on the analysis of measured and modeled black carbon data. The current process for accessing and processing data is entirely manual, with a person interacting directly with emissions modeling and ambient data (from Air Explorer) systems in order to obtain the right data (Figure 2). The necessary processing on the acquired data is also conducted by individuals. In sum, analysts spend 80% of their time in getting and processing the data, leaving only 20% of their time for generating their value-added contribution in the integrated analysis of the data.

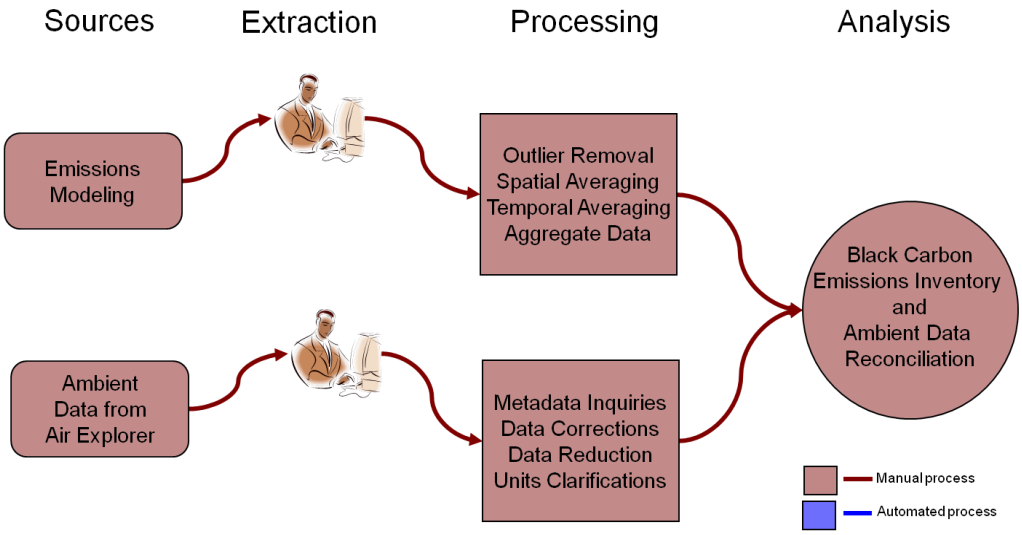

Figure 2. Current analysis process.

Figure 3 summarizes the envisioned streamlined process within a standards-based cyberinfrastructure. Automated services are set up to provide the necessary emissions and ambient measurement data along with their required processing. Humans enter the loop in order to validate and verify the quality of the resulting data and to spend the majority of their time on the unique analyses needed to advance air quality research and management.

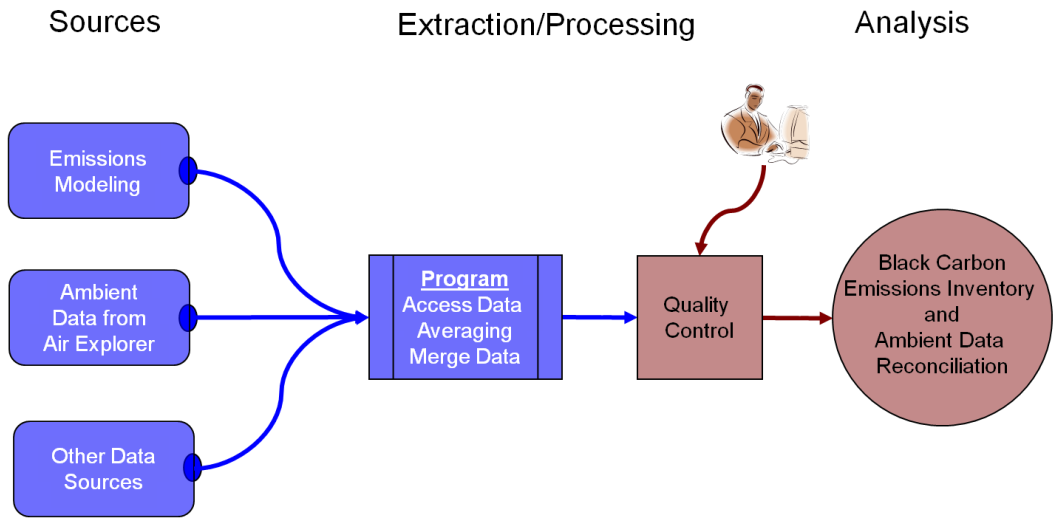

Figure 3. Future analysis process improved with interoperable components.

1.3 Components of an Interoperable Cyberinfrastructure

Air quality cyberinfrastructure components include data and information systems, systems that add value to those data, and tools for discovering and analyzing the data. Developing the

cyberinfrastructure does not necessitate drastic changes to the existing design and operations of data systems; the emphasis is on supplementing those existing systems to make data more accessible. In particular, the following component types are valuable in enhancing the interoperability of an existing information system:

- Data Formats – Standards and conventions for data structures and data files
- Naming Conventions – A common, standardized way of describing data
- Data Web Services – Standards and conventions for data access Web services
- Metadata – Standards and conventions for describing data sets
- Data Publication and Discovery – Standardized methods for publishing metadata to facilitate discovery, including catalog services
- Documentation – Best practices for system processes and use
- Information Page on Website – Common format for key standard information on an existing website

The availability of these components for an air quality data set makes it possible to integrate the data set with the rest of the air quality cyberinfrastructure, thereby increasing its visibility and value to potential users.

1.4 Incremental Interoperability

At first glance, achieving interoperability can seem to be overwhelming and a tremendous burden and effort. However, it is important to keep in mind that interoperability can be achieved incrementally. While the standards and nomenclature taken as a whole are complex and daunting, the process can become more manageable by focusing on an organization's current interoperability capabilities and identifying the easiest next step for advancing their interoperability. Depending on what an organization has already implemented in its data management practices and the resources, the organization can dedicate itself to becoming more fully involved in an air quality cyberinfrastructure. The next step can either be ambitious or conservative. For example, key markers on the interoperability continuum for data are

- Make data openly available
- Make data openly available in a standard data format
- Make data openly available in a standard data format using community naming conventions
- Make data openly available via a standardized Web service in a standard data format using community naming conventions.

By being provided publically (and through the Internet) in a standard format, data can be integrated into an air quality cyberinfrastructure. Achieving this level of interoperability could take the form of files served in an FTP directory or a web-accessible folder. This would not be considered a community best practice, but it provides a starting point that would get the organization engaged with the community and in a position to achieve a standardized Web service interface to their data.

1.5 How to Use this Document

This best practices document is intended to be used as a guidance or reference document. Its goal is to describe current practices across the air quality community in interoperability, present

the information associated with following these best practices in an understandable manner, and identify and encourage, as appropriate, use of the currently common and/or preferred community practices to maximize the advancement of a cyberinfrastructure for air quality management.

This document is intended to be a “living document,” to be updated by the community over time as more information becomes available to better capture the state of the art. The effort of capturing best practices is being coordinated through collaborative community spaces such as the ESIP Air Quality Workgroup and the Group on Earth Observations (GEO) Air Quality Community of Practice (AQ CoP). Practices will change over time. A companion web version of this document is planned; as a more dynamic, “live” document, that version will be open for comment and input by the community to refine the best practices and capture the latest advancements across the air quality community.

Different readers will find different sections of this document more useful than others. For example, a data provider with limited experience in Web services will find the overview sections of Web services more useful than a software developer who might be more interested in the specific implementation best practice for a particular data access Web service. A data analyst may find the best practices in data search and discovery helpful in identifying new mechanisms for finding and accessing data sets. This document is intended to facilitate easy navigation to sections of interest to these broad user types.

1.5.1 Document Structure

This document is primarily organized by best practice types or categories, such as naming conventions, data formats, and Web services. Each category of best practice is described with a background section to provide definitions and context. One or more best practices are presented in each category. For each best practice, a summary and rationale are provided prior to the description of the best practice itself. Finally, examples of the best practice and additional information, such as web resources, are provided for further reading and detail. The outline of the best practice category sections is summarized below.

- Category A
    - Background
    - Best Practice #1
        - Summary
        - Rationale
        - Best Practice (*a more detailed description*)
        - Examples
        - Additional Information
    - Best Practice #2
        - Summary
        - Rationale
        - Best Practice (*a more detailed description*)
        - Examples
    - Best Practice #3
        - …
- Summary Category B
    - …

2. Overarching Best Practices

2.1 Background

Several themes recur in these best practices for interoperability. These overarching best practices are fundamental to the goals of saving effort, reducing costs, and making data more accessible and usable:

2.2.1. Participate in community efforts

2.2.2. Use existing standards and software

2.2.3 Plan for long-term maintenance and operation

2.2.4. Document the approach

These themes are described briefly here, and discussed with more specifics in other sections of these Best Practices.

2.2 Best Practices

2.2.1 Participate in Community Efforts

Summary

Participate in and leverage the air quality community's efforts to improve data interoperability.

Rationale

The air quality community has put significant effort in recent years into developing standards, documenting advances, and adopting interoperable methods. The community's results should be leveraged to save effort and avoid duplication of effort.

Best Practice

When addressing data management questions, save time and cost by using information available from other organizations. The ESIP community, ESIP Air Quality Workgroup, and the GEO AQ CoP are excellent sources of information on interoperability for air quality data. The activities of these groups are open to participation and are usually documented on the Internet.

Examples

None.

Additional Information

None.

2.2.2 Use Existing Standards and Software

Summary

Always try to use standards and other software that are already being successfully used by the air quality community and that contribute to a common approach toward interoperability.

Rationale

No matter how unique a project may seem, developing new systems or using incompatible methods will be more costly in the long term than adapting to existing standards and software. Using existing standards and software not only saves time and effort in the near term, but also allows easier interoperability and possibly saves the effort of retrofitting a system in the long term.

Best Practice

Use existing standards, naming conventions, file formats, Web services, and other software components that are already being effectively applied by the air quality community. While existing software and standards may not fit every need, their components can be used and adapted to enable interoperability.

To select existing methods for a new system, look to similar air quality products, or to products with which the new system should be interoperable. If the new system requires unique features, use and conform to as many existing components as possible.

Examples

None.

Additional Information

None.

2.2.3 Plan for Long-Term Maintenance and Operation

Summary

When beginning a project, collecting data, or designing a data management system or service, consider the resources and effort that the task will require in the long term.

Rationale

Initial consideration for continuity, updates, maintenance, and operation can help ensure that a system will be more useful, have a longer life, and ultimately cost less to maintain.

Best Practice

Every data set, data management system, or data service will have different long-term requirements. For example, the following (non-exhaustive) list should be considered during planning:

- Points of contact, and substitutes or replacements

- Institutional support for the project
- Structure of the system (e.g., modular or hard-coded)
- Requirements/possibilities for manual or automated updates
- Maintenance of web links
- Equipment costs
- Time and personnel costs

Examples

None.

Additional Information

None.

2.2.4 Document the Approach

Summary

In developing a data set or service, always document the methods, sources, and processes used, as well as the outcomes.

Rationale

Data products are research results, and must be supported by documented methods and analyses. A well-documented data set can be more easily understood and applied, and documenting the approach during creation of a data set saves future effort by (1) reducing the time that a user would need to become familiar with the data, and (2) eliminating the time that a provider would waste in re-discovering or re-creating unrecorded information. In addition, documentation of data sets facilitates advancement along the interoperability continuum.

Best Practice

Record all information that would be necessary to understand the process used to create a data set. Document the process of creating the data set or service with the details that a new user (or the data creator after a lapse of time) would need to verify the accuracy of the data. Document any published Web services, including location, interface details, and usage examples.

Examples

None.

Additional Information

None.

3. Data Format Standards

3.1 Background

Achieving interoperability among air quality information systems ultimately depends on the data served by the providing organization being usable by the consuming organization. Central to usability is the structure and format of the data. No single format will serve every organization's needs. However, efforts are ongoing within the air quality community to achieve best practices on the use of platform-independent, interoperable data formats.

From a data provider perspective, compliance with commonly used data format standards will result in increased readability by many of the visualization and analysis tools used across the air quality community. Increased readability will raise the awareness of and exposure to a particular data set across research and air quality management groups.

A trend within the air quality community is the increased adoption of network Common Data Form with the Climate and Forecast conventions (netCDF-CF). NetCDF is a data format standard useful for storing and exchanging data that are provided on a grid as well as data from point monitoring networks. NetCDF is supported by a software library of tools for creating, editing and accessing netCDF-formatted data. The Climate and Forecast (CF) conventions define metadata that can be embedded in netCDF files to provide descriptions of each data-set variable and the spatial and temporal properties of those variables. With data formatted using netCDF-CF, users of data from different sources can decide which quantities are comparable. This data format also facilitates building applications that have powerful extraction, regridding, and display capabilities.

A challenge to advancing the adoption of netCDF-CF in the air quality community is a dearth of visualization and analysis tools that can read netCDF-CF. This diminishes the incentives and realized benefits of creating netCDF-CF–compliant data. However, the GEO AQ CoP has adopted netCDF-CF as its primary data format and is helping to advance the guidance and tools needed to promote its use.

3.2 Best Practices

3.2.1 Employ Data Formats Commonly Used or Recommended in the Air Quality Community

Summary

When making data available for use across the air quality community, provide the data in one or more commonly used data formats and conventions.

Rationale

As a whole, the air quality community has substantial experience formatting data files. Particular formats have become accepted and commonly used across the community due to their ease of implementation; portability; facilitation of self-describing data; and compatibility with processing, analysis, and visualization tools. Adopting already commonly used data format standards will make it easier for data consumers to use the data.

Best Practice

For **gridded data** (e.g., satellite observations, model results), the best practice is to use the netCDF data format with the CF conventions (netCDF-CF). The participants in the GEO AQ CoP have agreed to use the netCDF-CF convention in moving toward interoperability of their information systems.

For **point data** (e.g., monitoring network observations), the best practice cannot be defined simply at this point. The GEO AQ CoP has agreed on using the netCDF-CF convention, but the adoption of that convention for point data lags behind the adoption of that convention for gridded data. Data providers are encouraged to work with the AQ CoP in implementing netCDF-CF for their point monitoring data sets. Other widely implemented data formats that advance interoperability of point monitoring data include air quality designed formats, such as AQCSV, and other standards that have been adopted for air quality related data, such as KML. Tools for converting other data formats to netCDF-CF are available and can help data providers and consumers converge toward common formats. However, these tools are still in their infancy for point data.

For **vector-based data** (e.g., locations of monitors, area of interest), a best practice is yet to be determined. Data formatted in shapefiles can be read by most geospatial information system software. When serving vector data via Web services, the Open Geospatial Consortium (OGC) Web Feature Service (WFS) requires the use of the Geography Mark-up Language (GML) format.

For **map images** (e.g., static maps of monitoring locations or model results), a best practice is to provide map images in one or more of the accepted image file formats that are used in the OGC Web Mapping Service (WMS) standard. These file formats include PNG, TIFF, GIF, BMP, and GeoTiff.

Examples

For understanding the structure of netCDF-CF, the following examples direct you to live netCDF-CF files served on the Internet. A useful tool for visualizing and browsing gridded netCDF files is Panoply (http://www.giss.nasa.gov/tools/panoply/).

1. netCDF-CF files from Unidata:
   http://www.unidata.ucar.edu/software/netcdf/examples/files.html

2. A sample gridded netCDF-CF accessed from the Juelich WCS for HTAP:
   http://htap.icg.kfa-juelich.de:58080/HTAP_monthly?service=WCS&version=1.1.2&Request=GetCoverage&identifier=UM-CAM-v01_SR6SA_tracerm_2001&BoundingBox=0,-90,360,90,urn:ogc:def:crs:OGC::84&format=image/netcdf

3. Instructions and examples for working with point monitoring data using netCDF-CF:
   http://wiki.esipfed.org/index.php/WCS_Wrapper_Configuration_for_Point_Data

For another data format example, see a sample AQCSV formatted file: http://cyair.net/content/aqcsv-format-example-file. AQCSV format specification: http://cyair.net/content/aqcsv-format-specification

Additional Information

The following tools can assist in achieving this best practice:

- netCDF-CF compliance checker tool that verifies whether a netCDF file adheres to the CF metadata convention: http://cf-pcmdi.llnl.gov/conformance/compliance-checker/
- Instructions for multiple methods for creating netCDF-CF files: http://wiki.esipfed.org/index.php/Creating_NetCDF_CF_Files
- Geospatial Data Abstraction Library (GDAL) – translator code library for converting among multiple gridded data formats, including netCDF-CF: http://www.gdal.org
- Configuring a database for serving point monitoring data in netCDF-CF format: http://wiki.esipfed.org/index.php/WCS_Wrapper_Configuration_for_Point_Data
- Converting between I/O API netCDF (common format in the air quality modeling community) and netCDF-CF files: http://resources.arcgis.com/gallery/file/Geoprocessing-Model-and-Script-Tool-Gallery/details?entryID=2FB9C8EE-1422-2418-7FEE-DF5F31D93646

For background information on this best practice, consult the following resources:

- Background and description of the CF standard: http://cf-pcmdi.llnl.gov/
- Excellent summary of netCDF-CF convention: http://www.unidata.ucar.edu/projects/THREDDS/GALEON/netCDFprofile-short.htm
- Lists of standard names for use in CF metadata: http://cf-pcmdi.llnl.gov/documents/cf-standard-names/
- Description of netCDF-CF implementation for the AQ Community WCS: http://wiki.esipfed.org/index.php/WCS_Server_Software#WCS_Server_for_CF-netCDF_Grid_Data_Type

3.2.2 Describe Your Data Format

Summary

If your data do not conform to a commonly used format in the air quality community, provide a detailed description of the format so that the data can be translated into formats needed in data processing, analysis, and visualization tools.

Rationale

It can be a burden for data providers to alter the format they normally use to share their data or to add another data format option to their information system. In those cases, the usability of the data can be increased by providing a description of the data format so that others in the air quality community can translate the data into formats they need to work with the data.

Best Practice

For the structure of the file, indicate the data format name, version of the format (if relevant), encoding (e.g., text or binary), and the structure of the data file (e.g., names of each dimension or column/row in the data set, type of value in each dimension or column/row, size of each dimension or column/row).

For the content of the data, follow the best practice in name spaces (see section 4) and include spatial information and physical/chemical units of measure for the data values.

Examples

Sensor technologies are allowing simple sensors to be built for operation by citizens (sometimes also referred to as “citizen scientists”). These types of sensors can be used across a community to collect data on a variety of environmental parameters. Air quality measurements are being made using sensors of these types; for example, the air quality egg (http://blog.pachube.com/2012/01/airqualityegg-people-participating-in.html) being used by the Pachube community. Pachube is an effort to harmonize Web service interfaces among devices being connected on the “Internet of Things,” which is the notion that many types of everyday items will soon be able to generate data and communicate that data over the Internet.

Pachube has established its own set of internal standards for sharing monitoring data. The data must be formatted as comma-separated values (CSV), JavaScript Object Notation (JSON), or eXtensible Markup Language (XML) based on a set of required and optional data attributes. The table of data attributes can be found on the Pachube website: https://pachube.com/docs/v2/datastream/.

Additional Information

None.

4. Naming Conventions

4.1 Background

All communications require a common language for shared understanding. Exchanging data is no exception. The use of standardized naming conventions (. that is, a common way of describing data) avoids confusion and makes it easier for people and organizations to use your data for research, analysis, and building applications.

Naming conventions apply to several aspects of describing your data, including names of parameters, data formats and conventions, units, and data quality. These best practices provide the currently accepted approaches used by the air quality community and provide practical information needed to publish and share data.

4.2 Best Practices

4.2.1 Use an Existing Naming Convention

Summary

When creating and distributing data products, use an existing and established naming convention to describe the characteristics of the data (parameter names, units, time, etc.).

Rationale

A naming convention helps establish a common, documented, and consistent language that makes it possible for users of different data sources to decide which quantities are comparable; it also facilitates building applications that integrate data from different sources. Using an existing, community accepted, naming convention reduces the resources needed to publish your data, allows for unambiguous use of your data, and increases the use of your data.

Best Practice

Use one of the following naming conventions when publishing and distributing data. These naming conventions are widely used and accepted by the air quality community.

1. Climate Forecast (CF) Metadata Convention (http://cf-pcmdi.llnl.gov/documents/cf-standard-names/). This convention promotes the processing and sharing of files created with the netCDF API (http://www.unidata.ucar.edu/packages/netcdf/index.html). The convention provides a definitive description of what the data in each parameter represent and the spatial and temporal properties of the data. It covers air quality and meteorological data at the surface and aloft from instruments and models, and derived products.

   The CF standard names are constructed using guidelines described on the CF resource page (http://cf-pcmdi.llnl.gov/documents/cf-standard-names/guidelines). Generally, the names are self-describing and built by combining characteristics that describe the parameter. For example, ozone concentration is named mole_fraction_of_ozone_in_air. Ozone concentration at the top of the boundary layer is named mole_fraction_of_ozone_at_top_of_atmosphere_boundary_layer.

2. Air Quality System (AQS) (http://www.epa.gov/ttn/airs/airsaqs/). AQS is the U.S. Environmental Protection Agency's (EPA's) repository of ambient air quality and surface meteorological observations. AQS uses a series of reference tables that relate numerical codes to common names (http://www.epa.gov/ttn/airs/airsaqs/manuals/codedescs.htm). For example, ozone has a parameter code of 44201. Reference tables include parameter names, units, averaging periods, sampling frequency, etc.

Use naming conventions for all components that describe the data, such as

- Field names (reference names used as headers in files or Web services) that identify the contents of the data file.
- Units of reported parameters, which need to be explicitly stated in the data file and in the documentation. Use SI units, but recognize that each discipline may have its own commonly used units of measure.
- Time zone (i.e., time coordinate) for the data. For the air quality community, the time zone is typically either Coordinated Universal Time (UTC) or local standard time (LST).
- Date and time representation, which must include year, month, day, hour, minute, and seconds in a standard format.
- Time reporting convention (i.e., time stamp) indicating whether the reported time represents the beginning, middle, or end of the averaging period of the data. The convention within the air quality community is to report time as the beginning of the sampling period, whereas the meteorological community typically reports time at the end

of the sampling period. For example, a 60-minute average of ozone 1-minute concentrations from 13:00 to 13:59 is reported at 13:00 (begin time) by the air quality community and 14:00 (end time) by the meteorological community.

- Spatial coordinates (latitude and longitude) and altitude should follow the requirements and standards in the AQS or CF conventions.

If you choose not to adopt these naming conventions, then you should consider the following factors when selecting a naming convention:

- Is the convention supported by a large community and does the community actively update and improve the convention?
- Have key air quality organizations adopted the convention?
- What type of technical support is provided? Are detailed documents and tools available?
- What is the process for adding new names to the convention?
- How easy is it to adopt the convention and integrate it into your existing system?
- How will your organization update the naming convention when changes to it occur?

Examples

1. The National Oceanic and Atmospheric Administration's (NOAA's) Weather and Climate Toolkit viewer is an application that provides simple visualization and data export of weather and climatological data archived via Web services provided from the National Climatic Data Center (NCDC) and other organizations. http://www.ncdc.noaa.gov/oa/wct/

   The viewer provides tools for displaying custom data overlays, Web Mapping Services (WMS), animations, and basic filters in standardized format. The viewer supports reading data from netCDF-CF files.

2. AQS and the AIRNow system (www.airnow.gov) are exchanging data using a CSV-formatted file called AQCSV, which contains air quality data encoded with AQS parameters names and unit specifications. This allows for real-time data from the AIRNow program to be synchronized with the historical data contained in the AQS database. More information about the AQCSV format can be found at http://cyair.net/content/aqcsv-format-specification.

Additional Information

Support tools include the following:

1. CF Conventions
   - List of standard CF names: http://cf-pcmdi.llnl.gov/documents/cf-standard-names/.
   - Compliance-checker utility that evaluates a netCDF file to ensure it complies with the CF conformance requirements: http://cf-pcmdi.llnl.gov/conformance/compliance-checker/.

2. AQS

    - Data naming manual: http://www.epa.gov/ttn/airs/airsaqs/manuals/AQS%20Data%20Coding%20Manual.pdf.

Other naming conventions include the following:

- The Substance Registry Services (http://iaspub.epa.gov/sor_internet/registry/substreg/home/overview/home.do), which is the EPA's registry for information about regulated and monitored substances. It provides a common basis for identifying chemicals, biological organisms, and other substances.

Other information resources:

- DataFed has discussion and resources for air quality naming conventions (http://wiki.esipfed.org/index.php/Air_Quality/Chemistry_Naming_Conventions).
- EBAS is a database for hosting observation data of atmospheric chemical composition and physical properties for Europe (http://www.nilu.no/projects/ccc/emepdata.html).
- GEO AQ CoP discusses interoperability issues related to GEO/GEOSS (http://wiki.esipfed.org/index.php/GEO_AQ_CoP).

4.2.2 Document Your Naming Convention

Summary

Document and publish your naming conventions in a web-accessible format.

Rationale

Detailed documents that describe your naming conventions make it possible for end users to properly access, decode, and interpret your data. Using an existing naming convention helps save effort because you can leverage existing documentation.

Best Practice

Provide the following types of information to describe your naming convention:

- Naming convention and version
- Description of terms
- Narrative of how the naming convention works
- Tables that define all parameters, units, and other terms
- Assumptions used in the naming convention, and deviations from accepted standards
- Links to the naming standards

Examples

1. The CF conventions has extensive documentation describing all aspects of the naming conventions: http://cf-pcmdi.llnl.gov/documents/cf-conventions/latest-cf-conventions-document-1.
2. The AQS naming convention has a data dictionary describing each field name used by this standard. (http://www.epa.gov/ttn/airs/airsaqs/manuals/AQS%20Data%20Dictionary.pdf).

Additional Information

None.

4.2.3 Provide a "Crosswalk" to Other Conventions if Using a Non-Standard Naming Convention

Summary

If you plan to publish data using a non-standard naming convention, relate your unique convention to standard naming conventions.

Rationale

Creating a linkage between the parameter and field names to an accepted naming standard helps users better interpret and understand the contents of your data files. Without a crosswalk or method of translating parameter and field names, users and application developers may be reluctant to use your data or may misinterpret your data.

Best Practice

Create a simple, easy-to-use table in a web-accessible format that provides a map between your unique parameter names and community accepted parameter names.

Examples

Table 1 is a simple table showing the mapping of system-specific parameter names with community-accepted parameter names.

Table 1. An example of a table that shows the mapping between system-specific parameter names and community-accepted parameter names.

My Parameter Name	AQS Parameter Name (Numeric Code)
O3	Ozone (44201)
CO	Carbon Monoxide (42101)
SO2	Sulfur Dioxide (42401)
NO	Nitric Oxide (42601)

For a crosswalk table for air quality parameters, see http://vista.cira.colostate.edu/tss/help/parameterkey.aspx.

Additional Information

None.

5. Web Services

5.1 Background

The World Wide Web Consortium (W3C) defines a "Web service" as "a software system designed to support interoperable machine-to-machine interaction over a network" (http://www.w3.org/TR/ws-gloss/). Web services provide an interface for accessing data or information from remote systems and are useful for data exchange and interoperability, research, and software application development. For example, software applications have been developed for mobile devices (e.g., smartphones) that provide users with current air quality forecasts or observations. In response to user requests, software on a mobile device may access Web services (via the Internet) to retrieve desired data. Another example would be a sharing of data between two systems. Each system may publish a Web service to expose its own data (acting as a server), and retrieve data (acting as a client) via the other system's Web service.

Two main architectural styles exist for the development of Web services: (1) Simple Object Access Protocol (SOAP), and (2) Representational State Transfer (REST).

- SOAP: A protocol specification for exchanging structured data and information between computer systems. SOAP uses Extensible Markup Language (XML) as its message format. An individual SOAP interface is described in a machine-readable XML file written in Web Services Description Language (WSDL). A SOAP Web service may expose an arbitrary set of operations.
- REST: An architecture that works in a similar way as the World Wide Web itself. Clients navigate to web resources using the web address (e.g., the Uniform Resource Indicator [URI]) and invoke standard Hypertext Transfer Protocol (HTTP) actions on these resources (GET, PUT, POST, DELETE).

The SOAP approach to Web services has been popular for a number of years, and there are a variety of tools available for developers, including integrated development environments (IDEs), software libraries, debuggers, and other tools. SOAP fits well with object-oriented programming, and this likely contributed to its adoption. However, the SOAP approach may tend towards non-standard interfaces requiring tightly coupled (customized) client software. In addition, SOAP Web services are often very complex, requiring developer support.

REST-type architecture is recommended as a Best Practice for developing Web services for air quality data. REST is simpler to understand and implement, using the standard HTTP interface and actions. Client software accessing a more standardized REST-type Web service may have

limited knowledge of and customization for the server system. A standardized REST model better lends itself towards a service-oriented architecture (SOA), where generic clients access various standardized services, as opposed to custom client software being developed to access a particular server instance.

5.2 Best Practices

5.2.1 Strive for Compliance with Open Standards

Summary

When developing Web services to publish air quality data, strive for compliance with existing or proposed standards.

Rationale

Developing Web services in harmony with standards facilitates usage by exposing an interface that is well understood and consistent with other efforts in the air quality community. By adhering to this practice, data providers will avoid developing services that differ substantially from others being published or used in the air quality community. Striving for commonality and standardization will position your data system for future interoperability with other data providers; increased use of data by decision makers, analysts, and application developers; and an overall increased likelihood of system viability over time.

Best Practice

Design the Web service interface to conform to open standards published by the OGC (http://www.opengeospatial.org). The OGC defines itself as “an international industry consortium of 442 companies, government agencies, and universities participating in a consensus process to develop publicly available interface standards.” The OGC consensus process has produced standard definitions for a variety of geospatial Web services. The following OGC Web services are recommended for publishing air quality data:

- Web Coverage Service (WCS). A WCS is a Web service that provides access to data within a specified geographical area for a selected date/time range. Requested data are returned as “coverages,” gridded or point data values (e.g., in NetCDF-CF or CSV format) found within the spatiotemporal range specified. A WCS provides the following required operations:
    - GetCapabilities – Returns information about the server’s capabilities and the coverages available.
    - DescribeCoverage – Returns metadata for selected coverages provided by a server.
    - GetCoverage – Returns a coverage (data) within a specified geographical area and date/time range.
- Web Map Service (WMS). A WMS is a Web service that returns map images (e.g., JPEG, PNG, TIFF, GeoTIFF, etc.) for a specified geographical area, elevation, and date/time. The client request may include specific data parameters and geographic layers. The returned map image may include transparency so that multiple layers can be combined. A WMS provides the following required operations:

    - GetCapabilities – Returns a list of available data and valid WMS operations and parameters.
    - GetMap – Returns the map image requested by the client.
- Web Feature Service (WFS). A WFS is a Web service that returns geographical information (e.g., monitoring site locations) for a specified area. Data may be returned in XML format. A WFS provides the following required operations:
    - GetCapabilities – Returns a list of available feature collections.
    - DescribeFeatureType – Returns information about available attributes for features.
    - GetFeature – Returns the subset of feature data requested by the client.

When implementing a WCS for air quality data, build upon, reference, or model software after the Community WCS, which is being developed and implemented in the air quality community and complies with OGC standards for WCS version 1.1.0–1.1.2. The Community WCS is a REST-type Web service that is freely available at SourceForge (http://aq-ogc-services.sourceforge.net), with versions available for Windows and Linux (see Section 2.2.2 in this document).

When implementing a WMS for air quality data, build upon, reference, or model software after open source Web service frameworks that strive for compliance with OGC standards. Recommended examples include MapServer (http://mapserver.org/ogc) and GeoServer (http://geoserver.org).

When implementing a WFS for air quality data, build upon, reference, or model software after MapServer (http://mapserver.org/ogc) and GeoServer (http://geoserver.org) examples.

Publish metadata (see also Section 6) by including required operations and outputs described by OGC standards for Web services (see above). Operations that return metadata include

- GetCapabilities and DescribeCoverage for WCS
- GetCapabilities for WMS
- GetCapabilities and DescribeFeatureType for WFS

Examples

The Institute for Climate and Energy Research – Troposphere at Forschungszentrum Jülich (FZJ), Germany, implements the Community WCS. The FZJ WCS includes data from the Hemispheric Transport of Air Pollution (HTAP) Multi-Model Data Archive and the European Monitoring of Atmospheric Composition and Climate (MACC) project. FZJ Web service access is available at http://ogc-interface.icg.kfa-juelich.de:58080/.

The Community Initiative for Emissions Research and Applications (CIERA) implements the Community WCS to publish data from HTAP and the Emission Database for Global Atmospheric Research (EDGAR) and the National Emissions Inventory (NEI): http://pocus.wustl.edu:8080/NEI-EDGAR_Area, http://pocus.wustl.edu:8080/NEI-EDGAR_Point.

Demonstrations of MapServer Web services are available at http://demo.mapserver.org/.

Demonstrations GeoServer Web services are available at http://apps.who.int/tools/geoserver/demo.do.

Additional Information

OGC standards, definitions, and references: http://www.opengeospatial.org/standards/.

W3C information including Web services architecture: http://www.w3.org/, http://www.w3.org/TR/ws-arch/.

Community WCS project description, downloadable software, instructions, and links: http://ag-ogc-services.sourceforge.net/, http://wiki.esipfed.org/index.php/WCS_Wrapper_Support.

GeoServer open source software project and documentation: http://geoserver.org/display/GEOS/What+is+Geoserver, http://docs.geoserver.org.

MapServer open source software project and documentation: http://mapserver.org/, http://mapserver.org/documentation.html.

5.2.2 Re-Use/Build Upon Existing Software Packages

Summary

Re-use/build upon existing Web service software, particularly software developed for and used by the air quality community.

Rationale

Implementing open source or otherwise freely available software in use by the air quality community (and striving to comply with open standards—see Section 5.2.1) will reduce development costs and lead to a commonality of approach by various data providers. In addition, contributing to a common codebase will assist the community to move toward a standardized approach.

Best Practice

Review available Web service software packages used in the air quality community to determine usability on current or future projects. To publish gridded or point data, obtain and build upon the Community WCS (see Section 2.1). The Community WCS is being maintained as an open-source software project, developed in a public, collaborative manner. The Community WCS software is freely available at SourceForge (http://ag-ogc-services.sourceforge.net). The Darcs distributed revision control system is being used to manage software versions; instructions for using Darcs to download the Community WCS and to contribute to the codebase are at http://ag-ogc-services.sourceforge.net/development.html. Documentation, step-by-step instructions, and other resources for the Community WCS are available on the ESIP wiki at http://wiki.esipfed.org/index.php/WCS_Access_to_netCDF_Files.

If the Community WCS is not appropriate for project requirements (e.g., a WMS or WFS is required), use a common Web service software framework that complies with open standards. MapServer (http://mapserver.org/ogc) and GeoServer (http://geoserver.org) are recommended. When implementing third-party Web service software, reference air quality community software projects (e.g., the Community WCS) and strive to model the interface, output formats, and naming conventions to be consistent with community efforts.

Examples

The FZJ WCS (see examples in Section 5.2.1) is available at http://ogc-interface.icg.kfa-juelich.de:58080/.

The CIERA WCS (see examples in Section 5.2.1) is available at http://pocus.wustl.edu:8080/NEI-EDGAR_Area, http://pocus.wustl.edu:8080/NEI-EDGAR_Point.

Demonstrations of MapServer Web services are available at http://demo.mapserver.org/.

Demonstrations GeoServer Web services are available at http://apps.who.int/tools/geoserver/demo.do.

Additional Information

For more information, see following resources and tools:

- Community WCS project description, downloadable software, instructions, and links: http://aq-ogc-services.sourceforge.net/, http://wiki.esipfed.org/index.php/WCS_Wrapper_Support
- GeoServer open source software project and documentation: http://geoserver.org/display/GEOS/What+is+Geoserver, http://docs.geoserver.org
- MapServer open source software project and documentation: http://mapserver.org/, http://mapserver.org/documentation.html

5.2.3 Determine Approach Based on User Needs

Summary

Develop Web services using an approach or technology that meets specific user needs.

Rationale

Various Web service implementations return differing payload types and formats; therefore, considering user requirements during initial planning will lead to the publishing of Web services that can better meet user needs. Web services that effectively meet user needs are employed by researchers, software application developers, decision makers, and data providers.

Best Practice

Analyze your user base and their specific data requirements to identify use cases.

Determine appropriate Web services to implement based on user requirements. For example,

- If users require gridded or point data values, implement the Community WCS (see Section 5.2.1). The Community WCS currently returns gridded data values in NetCDF-CF format and point data values in CSV format.
- If users require data integrated into map images (e.g., JPEG, PNG, GIF), such as for use in a GIS viewer, implement a WMS. Recommended examples include MapServer (http://mapserver.org/ogc) and GeoServer (http://geoserver.org) (see Section 5.2.1).
- If users require geographical information (e.g., monitoring site locations), implement a WFS (see Section 5.2.1) following the examples at http://mapserver.org/ogc (MapServer) and http://geoserver.org (GeoServer).

Examples

See examples in Section 5.2.1.

Additional Information

For more information, resources, and tools, see the “Additional Information” section in Section 5.2.1.

5.2.4 Document the Approach Used

Summary

Develop and publish documentation describing the approach used, Web service type(s), specific framework employed, standards compliance, custom software modifications or extensions to existing open source software, and code samples.

Rationale

Documentation of the approach used to implement Web services is required for wider understanding and use of the published interfaces by the air quality community and others. Clear documentation of published Web services will facilitate usage of the interface. Sample code will make it easier for developers to understand how to invoke the published service. In addition, documentation of extensions to existing software, particularly those that move standards and practices towards better accommodation of air quality data, will benefit future developers of Web services within the air quality community.

Best Practice

Plan for documentation of Web services from the beginning of the software development project. Include an overview of the available Web services, how you expect the services to be used, and a summary of the documentation itself.

Include specific information about the Web services:

- Provide descriptions and locations for the Web services. Give details about the types of services available, standards compliance and version, and framework used (e.g., when using open-source codebases such as the Community WCS, MapServer, or GeoServer).
- Document the operations available for each Web service:
    - Describe what each operation does.
    - List and explain the query parameters/arguments required for the operation.
    - Describe the outputs produced by the operation.
- Provide detailed, functional code samples for using the Web services.
- Provide descriptions and locations of informational operations and metadata files.
- Provide detailed descriptions of payload options and data formats.
- Describe error handling and error messages returned by various operations.
- Describe any authentication or security credentials required for accessing resources and how these are implemented.

- Plan for maintenance of the documentation to reflect any updates or changes. Preserve a history or revisions.

As described in Section 5.2.2, when possible, re-use or build upon existing software used by the air quality community. When existing software is used, include references or links to documentation provided by any open source codebase being implemented.

Examples

None.

Additional Information

None.

5.2.5 Consider Long-Term Operation and Maintenance

Summary

Before deploying a Web service to publish air quality data, consider operational requirements such as performance and system support. Plan for maintenance of the Web service in an operational state over time.

Rationale

Users may depend on operational Web services in a variety of ways. For example, users may routinely access Web services to update and synchronize other data systems, or the Web services may be accessed on demand by custom software. Planning for an appropriate level of operational support will ensure that the Web service meets user needs over time.

Best Practice

Determine operational requirements by analyzing typical or historical data usage patterns and by soliciting input from prospective users of Web services. For example, consider the following:

- Specific performance requirements, including:
    - Response times – how quickly the system handles individual user requests
    - Throughput – the number of requests the system can handle
    - Concurrency – the number of simultaneous user requests the system can handle
- System tolerance for downtime or maintenance windows.
- The average size of data returns (based on an average of requested spatial and temporal extents).
- Whether the Web services will primarily be used for occasional retrieval of archival data or will include on-demand requests for real-time data (e.g., by software applications).
- Whether the Web service will be available to a limited number of users, or will be available to the public.
- Whether there will be an expectation of 24/7/365 operation.

On the basis of these considerations, plan for

- Server and network requirements, including number of processors, processor speed, memory, disk space and throughput speeds, and network bandwidth.
- Data caching methods.
- Hardware and software redundancy.
- Data center staffing.
- User support (e.g., a Help Desk).
- User training and/or training materials.

Develop and document a plan for eventual Web service removal (decommission) or migration. For example, determine

- How users of Web services will be notified of changes.
- How much advance notice will be provided.
- How long the legacy Web services will be maintained in parallel with new versions when Web services are migrated to newer or modified versions.

Examples

None.

Additional information:

None.

6. Metadata

6.1 Background

Metadata is a description that accompanies a data set or service; it provides quantitative and qualitative information about the data set or service's attributes and provenance. Metadata can be thought of as the information needed in order to understand your data or use your service in the future. This information helps a potential user identify the data or service that will meet their needs, based on location, time, data quality, resolution, processing, or other characteristics; understand the methods that were used to create it; and apply it correctly. Considering metadata early in the data management process will save significant time when sharing and receiving data in the future.

Even at a small scale, metadata is required for sharing data and services; metadata standards allow interoperability on a much larger scale. Metadata standards go beyond descriptions to provide a common set of terms and definitions in a structured format. The standard terms and structured format allow a diverse set of users (both humans and machines) to understand, read, and correctly use data and services. Because metadata standards allow for a common understanding, they are vital to reducing ambiguity and allowing easy data sharing.

6.2 Best Practices

6.2.1 Always Provide Metadata in Some Format

Summary

For every data set or service, provide a description that would allow someone who is unfamiliar with the work to understand the contents.

Rationale

Potential data users must fully understand the parameters in the data set, including the parameter name, unit of measure, and format. By providing appropriate metadata, we ensure that data can be used correctly and appropriately now and in the future. Including adequate descriptions of data, products, and services increases their value and impact by allowing products to be discovered by more potential users, included in more analyses, used appropriately, and cited more widely.

Best Practice

Every data set and service should have accompanying metadata in some format, but ideally in a common or standard format (see the best practices below on metadata standards).

For a small or simple data set, the metadata content and format may be comparably small or simple, as long as it describes the product adequately. More complex products and services require more detailed descriptions. Metadata can be provided in a separate file that is identified within the product, or as a header within the data file. At a minimum, metadata should include information to help a user understand the following:

- What the data describe (parameters, units, etc.)
- When and where the data were collected or created, and when modified
- Temporal and spatial extent and resolution, and how to interpret their formats
- Potential quality issues, and how are they flagged
- Missing value definitions
- Meaning of any codes or abbreviations
- Who created the data (project, group, contact)

For data discovery and data sets provided on the web, metadata should include the following fields to help a user find and obtain the data:

- Web service and service type
- Fees or costs for the data
- Access or usage constraints
- Publication date and version information
- Data host or provider, if different from the entity that created the data

Examples

Link to examples of XML metadata files for emissions inventories, model output, AERONET, and other data sets: http://capita.wustl.edu/dataSpaceMetadata_ISO/.

Additional Information

NOAA has developed a description of how to provide good documentation for data sets and services: https://geo-ide.noaa.gov/wiki/index.php?title=Creating_Good_Documentation (there may be a security warning, but the site is safe).

DataOne has a series of best practices on metadata that is a useful reference: http://www.dataone.org/best-practices/metadata.

The European Commission has a tool for creating metadata that is consistent with the INSPIRE guidelines. It is an excellent resource for a list of the fields that should be provided for every data set in a range of categories (contact person, identification, classification, keyword, geographic, temporal, quality, conformity, constraints, and responsible party) http://www.inspire-geoportal.eu/EUOSME/ (if this link is broken, search the web for “inspire editor”).

6.2.2 Always Include Contact and Citation

Summary

Include a contact person and instructions for citing the product in the metadata.

Rationale

Users need a contact person to answer potential questions about the data or service. Including instructions or guidelines for citing the data or service ensures that users can cite it properly and that publications can be linked.

Publications traditionally provide a point of contact for clarifications and further information. Data providers should provide the same information to aid in correct usage and interpretation of their products. Similarly, citation of data has the same benefits and purposes as the well-established practice of citing previous publications, such as giving credit to authors, tracking uses and publications, and permitting reproducibility of results. Ultimately, providing citation information allows direct reference to the data that are used in a publication.

Best Practice

In the metadata, include the name, organization, and email address of the person who can answer questions on the data or service. A generic email address for a team may be useful if the specific point of contact for the future is uncertain (e.g., datacontact@organization.org).

Follow the guidelines for data providers on citation information developed by ESIP (http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines). Provide as much information as would be needed for the reader of the citation to locate and obtain the data set. Citations for data are similar to citations for publications, so parallel information should be provided. The minimum information for a citation includes

- The name(s) of the author(s)

- Release date and version
- Title of the data set
- Archive or distributor (i.e., publisher, or the agency responsible for the data)
- Locator, identifier, or media (i.e., information about how the data was accessed, such as Digital Object Identifier (DOI) or CD-ROM)
- Access date for online resources

An organization that provides multiple data sets may also list a citation policy for when and how the data should be cited.

Examples

The ESIP guidelines for data providers include citation examples: http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines.

NASA's Giovanni provides a citation policy and example: http://disc.sci.gsfc.nasa.gov/giovanni/overview/how-to-acknowledge-giovanni.

Additional Information

The Digital Curation Centre provides a guide on citations for data and linking them to publications: http://www.dcc.ac.uk/resources/how-guides/cite-datasets.

DataCite provides international Metadata Schema for the Publication and Citation of Research Data: http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf.

6.2.3 Comply With at Least One Major Standard

Summary

Provide metadata that is consistent with an existing major metadata standard, preferably ISO 19115 or 19115-2.

Rationale

Metadata standards allow data to be fully understood by a wider range of users, and more easily read by machines. The ISO 19115 metadata standard is emerging as the most widely used metadata standard in the air quality community. As DataOne notes, "the metadata standards that you use determine the communities that can easily use your data."

Providing metadata in standard formats saves time and effort when comparing, combining, or using multiple data types from multiple data providers. Following metadata standards used by the air quality community, particularly ISO 19115, makes it easier for data users to understand the products and apply them.

Best Practice

Provide metadata in the standard ISO 19115 format, which is widely used by the air quality community. ISO 19115 is the metadata standard for geospatial information which is accepted by GEOSS. ISO 19115 also includes ISO 19119 metadata for describing geospatial services (WMS, WCS) and AQ-specific metadata for finding data sets. ISO 19115-2 includes information

about the instruments used to collect the data. Crosswalks and translators exist to make translation between existing metadata standards possible, so metadata that are compliant with one standard can more easily be made compatible with another.

Ideally, the air quality community would use one common metadata standard, and ISO 19115 is emerging as the most widely accepted standard. There are a number of standards currently in use, and some translators exist for the different standards.

Examples

Examples of ISO 19115/19119 metadata provided by NOAA: http://www.ngdc.noaa.gov/metadata/published/NOAA/IOOS/iso/.

Examples of ISO 19115 metadata compiled by capita.wustl.edu: http://capita.wustl.edu/dataSpaceMetadata_ISO/.

Additional Information

NOAA provides a wiki page on many aspects of metadata and ISO 19115 standards: https://geo-ide.noaa.gov/wiki/index.php?title=Main_Page (there may be a security warning, but the site is safe), including a FAQ (https://geo-ide.noaa.gov/wiki/index.php?title=ISO_FAQ) and a "building blocks" approach to becoming familiar with the standards (https://geo-ide.noaa.gov/wiki/index.php?title=Category:ISO_Building_Blocks).

Some parts of metadata may be identical for the various data sets provided by a group. NOAA provides guidance on shortcuts for repeatedly using boilerplate information (such as contact person) with ISO standards: https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Boilerplate (there may be a security warning, but the site is safe).

The GEO AQ CoP provides a list of core ISO 19115/19119 metadata fields for discovery of air quality data: https://sites.google.com/site/geosspilot2/air-quality-and-health-working-group/aq-community-catalog---publish-register-harvest/iso-metadata.

The universal language of metadata is XML, and many powerful XML management tools exist that can be used to create and validate metadata. The oXygen XML editor (http://www.oxygenxml.com/) and Altova's XMLSpy (http://www.altova.com/xmlspy.html) can be used to edit XML metadata files.

The European Commission has a tool for creating metadata compliant with ISO 19115 and ISO 19119: http://www.inspire-geoportal.eu/EUOSME/.

Datafed.net provides a form to create and download ISO 19115 metadata for WCS and WMS applications: http://webapps.datafed.net/AQ_register.aspx.

The ISO 19115/19119 XML schemas are available at http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/.

6.2.4 If Unique Aspects Are Required, Provide Crosswalks or Translators

Summary

If a data or service requires unique descriptors that are not part of an existing metadata standard, provide information that allows the unique descriptors to be clearly understood or translated into other standards.

Rationale

It is possible that existing metadata standards, such as ISO 19115, may not fit all the needs of a particular data set or service. In some cases, the standard can be extended to accommodate those needs, or non-standard fields must be described to ensure that they are understood.

Best Practice

If non-standard fields are required, look at existing metadata for similar products and leverage similar fields that have already been used by the air quality community. If no similar examples can be identified, consult an expert in the air quality community. If you determine that the field truly does not fit an existing standard, provide the following information:

- A description of the field;
- How the field should be interpreted or read; and
- What the field is called in other metadata standards, if applicable.

Examples

The Federal Geographic Data Committee (FGDC) provides a crosswalk between FDGC CSDGM and ISO 19115: http://www.fgdc.gov/metadata/documents/FGDC_Sections_v40.xls/view?searchterm=ISO%20Crosswalk.

DIF to ISO mapping is provided by NASA GSFC: http://gcmd.gsfc.nasa.gov/Aboutus/standards/difiso.html.

Additional Information

See links to examples of metadata and information in other sections.

6.2.5 Provide as Much Information as Possible

Summary

Always provide as much detail in metadata records as is possible, without obscuring the most important metadata.

Rationale

Metadata is not intended as shorthand for users that are already familiar with the product, but as a complete description for new users. Fully descriptive metadata will help users to discover and appropriately apply data sets and services. In some cases, a “minimum” set of metadata fields is defined. However, practice has shown that providing only the minimum set of metadata fields

will limit interoperability efforts. For example, the set of minimum metadata fields may make it possible for a user to find the data set, but not offer sufficient information for using the data set.

Best Practice

Provide as much detail as possible in all metadata records. Rather than populating the minimum fields required by a metadata standard, provide as much information as would be needed for a new user to fully understand the product.

Some metadata standards currently in use by the community require a minimum set of information. Because it is unlikely that one single standard could require a set of information that is fully descriptive for all types of data and services, the best practice is to provide as much information as possible.

Examples

None.

Additional Information

None.

6.2.6 Share the Metadata You Create

Summary

Once metadata is developed, make it available both for describing a product and for making it possible for users to find or “discover” the product.

Rationale

Metadata is intended to help users discover and apply data products and services, but it must be published or otherwise made available to users. The metadata become part of the overall air quality cyberinfrastructure and can be used by search tools, analysis tools, and air quality researchers.

Best Practice

Metadata availability and publishing should be considered as part of the data management plan. There are a number of ways to make metadata available and useful. The best practice is to select a method that is appropriate for the product and data system. Among the possible methods are the following:

- Make metadata available in a web-accessible folder.
- Design the data management system so that metadata can be accessed with every data download.
- Publish metadata by including required operations and outputs described by OGC standards for Web services (see Section 5.2.1 on best practices for Web services).
- Add your metadata to a catalog, such as GEOSS or the Air Quality Community Catalog.

Examples

The GEO AQ CoP provides a comparative list of elements for several metadata discovery standards: https://docs.google.com/spreadsheet/ccc?key=0AqFCPt1U1kS4cEcwY0QzNVNCX0E4Vk1iY29VR3BrVWc#gid=1.

Additional Information

GEOSS provides some information on metadata publishing for services: https://sites.google.com/site/geosspilot2/Home/clearinghouse-catalogue-registry-metadata/metadata-for-services-2/metadata-for-services.

6.2.7 Indicate Data Quality

Summary

Provide information about the data quality validation level and data quality flags for all data. Data validation level is a method of indicating the degree of confidence in the overall data set. Data quality flags indicate the quality of an individual data point. Publish and share only data that you consider valid.

Rationale

Information about the quality of your data helps users develop appropriate applications. Without data quality information, users may assume that all data can be used for any application (e.g., regulatory analysis).

Best Practice

Indicate data validation level using the convention in Table 2.

Indicate data quality using quality control codes shown in Table 3 and only publish data that are valid, estimated, or missing. Avoid publishing suspect or invalid data.

Table 2. EPA's data validation levels for air quality and meteorological data.

Level	Definition
Level 0	Raw data for internal use only. Not for distribution. Essentially, Level 0 data are raw data obtained directly from the data acquisition systems in the field. Level 0 data have been reduced and possibly reformatted, but are unedited and unreviewed. These data have not received any adjustments for known biases or problems that may have been identified during preventive maintenance checks or audits. These data should be used to monitor the instrument operations on a frequent basis (e.g., daily), but should not be used for regulatory purposes until they receive at least Level 1 validation.
Level 1	Data have been quantitatively and qualitatively reviewed for accuracy, completeness, and internal consistency. Quantitative checks are performed by software screening programs and qualitative checks are performed by trained personnel who manually review the data for outliers and problems. Quality control flags (consisting of numbers or letters) are assigned to each datum to indicate its quality.

Level	Definition
Level 2	Involves comparisons with other independent data sets. This includes, for example, intercomparing collocated measurements or making comparisons with other measurement systems.
Level 3	Involves a more detailed analysis when inconsistencies in analysis and modeling results are found to be caused by measurement errors.

Table 3. Suggested quality control (QC) codes for meteorological and air quality data.

Code	Meaning	Description
0	**Valid**	Observations that were judged accurate within the performance limits of the instrument.
1	**Estimated**	Observations that required additional processing because the original values were suspect, invalid, or missing. Estimated data may be computed from patterns or trends in the data (e.g., via interpolation), or they may be based on the judgment of the reviewer.
7	**Suspect**	Observations that, in the judgment of the reviewer, were in error because their values violated reasonable physical criteria or did not exhibit reasonable consistency, but a specific cause of the problem was not identified. Additional review using other, independent data sets (Level 2 validation) should be performed to determine the final validity of suspect observations.
8	**Invalid**	Observations that were judged inaccurate or in error, and the cause of the inaccuracy or error was known. Besides the QC flag signifying invalid data, the data values themselves should be assigned invalid indicators.
9	**Missing**	Observations that were not collected.

Examples

None.

Additional Information

EPA provides data guidance on environmental data verification and data validation (http://www.epa.gov/quality/qs-docs/g8-final.pdf).

EPA's Data Quality Objectives Process provides information on how to apply systematic planning to generate performance and acceptance criteria for collecting environmental data (http://www.epa.gov/quality/qs-docs/g4-final.pdf).

7. Data Publication and Discovery

7.1 Background

A key challenge in attaining interoperability among the components of distributed information systems is the ability for data providers to make their data easy to discover by consumers in need of those particular data. Current approaches in publishing data Web services to foster their

discovery by the broader community include standards-based metadata web catalogs as centralized search points of registered metadata, and “service casting” for more localized publication of data services.

OGC’s Catalog Services for the Web (CSW) standard provides a framework for creating an online catalog that data providers can use to register their data services using standard metadata. The catalog maintains the metadata and standardized interfaces for searching the metadata and identifying data services of interest to the user. Service casting is based on using syndication feeds for advertising or broadcasting Web services. The standardized feed makes it possible for users to discover the service and provides information for either directly accessing the service or integrating it into data search tools.

In the air quality community, a centralized AQ Community Catalog approach is currently used to organize metadata for data access services. During the GEOSS Architecture Implementation Pilot (AIP), a method for using Web Accessible Folders (WAFs) was defined where standards-based metadata files using the ISO 19115 metadata standard are organized in a common folder accessible through the Internet. The location of the folder is used by centralized catalogs (such as GEOSS) or search tools (such as uFind).

7.2 Best Practices

7.2.1 Expose Your Metadata to a Common Catalog in the Air Quality Community

Summary

Sharing data access service metadata with the Air Quality Community Catalog makes that metadata available for discovery across the air quality and broader earth science community.

Rationale

The air quality community has been working on approaches for improving the sharing and discovery of data access service metadata through the use of metadata standards and catalogs. While current approaches are not yet mature enough to be considered best practices, they represent the best approach for air quality data providers and data consumers to become engaged with others in advancing to a common approach to search and discovery of air quality data.

Best Practice

1. Create standard metadata for your data service (see Section 6).
2. Make your metadata accessible through the Web by publishing ISO 19115-compliant metadata files in Web Accessible Folders.
3. Work with the AQ community in enhancing and formalizing the process for publicizing and finding air quality related data services.

Examples

The Air Quality Community Catalog Web Accessible Folder contains the metadata files for data services accessible to the GEOSS Catalog and Air Quality Community Catalog: http://capita.wustl.edu/DataspaceMetadata_ISO/.

Additional Information

- Air Quality Community Catalog (uFind): http://webapps.datafed.net/CORE.uFIND
- Group on Earth Observations Catalog: http://geoportal.org (you may be prompted to install the Google Earth plugin)
- Global Change Master Directory: http://gcmd.nasa.gov
- Background information on service casting:
  - Overview from NASA JPL: http://sciflo.jpl.nasa.gov/scast/
  - ESIP Discovery Cluster efforts: http://wiki.esipfed.org/index.php/Discovery_Cast_Atom_Response_Format_v1.1