Environmental Data Management Best Practices

Geospatial Data Management
Data Standards

Developing geospatial data standards for your organization is critical to ensure data are collected and managed properly so that they can easily be used for analysis, visualization, and dissemination of information. This document describes best practices for geospatial environmental data standards to improve interoperability. This document will help program and project managers understand the importance of data standards and provide bulleted lists and additional resources for GIS professionals to implement standards.

Overview

Geospatial data management plays a connecting role between data acquisition, data modeling, data visualization, and data analysis (Evans et al. 2017). The massive use of geo-referenced data sets in many fields of science, including earth observation, environmental sciences, city planning, building information modeling, real-time processing, and analytics for geospatial data makes geospatial data management increasingly a central task in the workflow of geospatial data processing. Big data practitioners are experiencing a huge number of data quality problems, which can be time-consuming to solve, or can even lead to incorrect data analytics (Breunig et al. 2020). All organizations should have geographic data best management practice guidance or standard operating procedures (SOPs). Organization standards are discussed in a separate subtopic sheet (Organization Standards for Geospatial Environmental Data Management) and are different from the data standards described here.

Developing geospatial data standards for your organization is critical to ensure data is collected and managed properly so that it can easily be used for analysis, visualization, and dissemination of information. When working in a regulatory program, for example, it is important to follow all requirements of that organization. This document contains general best practices that can be included in an organization’s guidance and SOPs for geographic environmental data standards to improve interoperability. Examples of regulatory requirements are also provided.

Project Planning and QAPP

Knowing what your output will be at the end of your project determines how the geographic data will be collected, stored, and managed. Defining performance criteria for new data gathered during a project and acceptance criteria for existing or historical data sets that may be incorporated into a project is an important step in the project planning process. Performance criteria and acceptance criteria are usually defined in the quality assurance project plan (QAPP).

- Quality of geographic data begins at project inception when the QAPP specifies the quality assurance measures needed for geographic data to make sound environmental decisions. Complete discussion of developing a geospatial QAPP is beyond the scope of this ITRC Team; USEPA guidance QA/G-5G (USEPA 2003) provides instructions for preparation of geospatial data QAPPs. An important aspect to consider, however, is the acceptance measures for values. These may include a range of acceptable values for identifying correct location of sample collection points or the use of recreational-grade GPS receivers versus professionally surveyed locations of sample points, for example.
- Identify geospatial quality measures needed for project goals and objectives.
- At minimum, QAPPs should include guidelines for:
    - Accuracy.
    - Consistency.
    - Validation of data or quality assurance and quality control (QA/QC) processes.
    - Other important aspects of a geospatial QAPP are discussed in Guidelines for Creating a Geospatial Quality Assurance Project Plan: EPA QA/G-5G (USEPA 2003) and ITRC's Environmental Data Management (EDM) Data Quality fact sheets.

Data Storage

A key component to storing environmental data in GIS is consistency. Consistent file structure, geospatial database structure, feature class names, and attribute data should be used within an organization. Being consistent within an organization allows multiple users to quickly understand and work with various environmental data. Maintaining consistent environmental data also allows data from multiple projects to be combined for visualization and analysis on a larger scale. Here are suggested strategies to create consistency in an organization:

- **Drop-Down Lists for Data Entry**
    - Create lists for users to select from to ensure consistency between users and data sets and ensure use of valid values.
    - Spreadsheets are a common data source for GIS. Drop-down lists can easily be built into spreadsheets to create consistent data sets before the data are brought into GIS.
    - Drop-down lists can be used in GPS and field collection applications.

- **Field-Type Definitions**
    - While GIS is seldom used as the sole means to organize environmental data management (EDM), if field types are defined, GIS can be used to query, display, and analyze environmental data from the environmental data management system. Refer to the Environmental Data Management Systems White Paper for further discussion. Any software data management program, including GIS, has specified field types. In GIS, each attribute field should have a defined field type. Field types in GIS include numbers, text, dates, binary large objects, object identifiers or unique key, and geometry.
    - Field types ensure text is stored as text, numbers are stored as numbers, and dates are stored as dates.
        - Example—If dates are stored in text fields rather than date fields, time analysis tools for trend analysis will not be able to be run in GIS.

    - Organizations should have defined field types for all data stored. This will allow the organization to combine data from multiple sources.
    - You can set ranges of acceptable values in GIS so if data are imported into a GIS those values make sense.
        - Example—Set pH range to 0–14 to ensure all numbers are acceptable pH values.

- **Valid Values**
    - See detailed information on valid values for location information in the Data Sharing and Transfer section below. Refer to the Valid Values Fact Sheet for a detailed discussion of valid values for EDM.

- **Standard Naming Conventions for GIS File Names**
    - GIS data should be organized by category and data type.
    - Feature data sets can be used to store related categories of data together.
    - The first part of a file name should be the largest category, such as a geographic area, project name, or type of work. Each subsequent part of a file name should be increasingly more specific. This approach to file naming groups like data sets together when looking at a list of files or searching for a particular type of data.
        - Hydrologic Unit Maps are a good example of this file naming structure. The United States is divided and subdivided into successively smaller hydrologic units, which are classified into four levels: regions, subregions, accounting units, and cataloging units (more commonly referred to as watersheds). The hydrologic units are arranged (or nested) within each other, from the largest geographic area (regions) to the smallest geographic area (subwatersheds). Each hydrologic unit is identified by a unique hydrologic unit code (HUC) consisting of between two and sixteen digits based on the classification in the hydrologic unit system. Visit the U.S. Geological Survey (USGS) website https://water.usgs.gov/GIS/huc.html for detailed information.

- **Standard Naming Conventions for Environmental Locations**
    - Ensure that a unique identifier is used for each location.
    - A location name should correspond to a single set of coordinates for point features. Ideally, the GIS manager will work with the data manager to ensure consistent identifiers for locations. Investigation-specific identifiers can also be developed. It is important that field data collection and data entry personnel are trained to use those identifiers during data collection and entry.
        - Example: A monitoring well named MW-1 is located at a unique (X, Y) longitude and latitude coordinate pair. If procedures are implemented to ensure MW-1 (and not MW1, MW01, MW 1, or MW_1) is the identifier used for all sampling events, data from multiple events for this location can more easily be stored and accessed to evaluate trends.

- **Data Dictionaries**

All the above strategies for creating consistency can be built into a data dictionary. For groups of people working with similar data, having a shared data dictionary facilitates standardization by documenting common data structures and providing the precise vocabulary needed for discussing specific data elements. Shared dictionaries ensure that the meaning, relevance, and quality of data elements are the same for all users (see USGS Data Dictionaries at https://www.usgs.gov/data-management/data-dictionaries). Data dictionaries explain what the data are, but do not specifically go into detail needed for data mapping necessary in data exchange or electronic data deliverables. See also the Electronic Data Deliverables and Data Exchange Fact Sheet for more information.

- Data dictionaries store and communicate metadata about data in a database, system, or in data used by applications. Data dictionary contents typically include:
    - A listing of data objects (names and definitions)
    - Detailed properties of data elements (data type, size, nullability, optionality, indexes)

        - Entity-relationship and other system-level diagrams
        - Reference data (classification and descriptive domains)
        - Missing data and quality-indicator codes
        - Business rules, such as for validation of a schema or data quality

- Data dictionaries can be used for the following:
    - Documentation—provide data structure details for users, developers, and other stakeholders
    - Communication—equip users with a common vocabulary and definitions for shared data, data standards, data flow and exchange, and help developers gauge impacts of schema changes
    - Application design—help application developers create forms and reports with proper data types and controls, and ensure that navigation is consistent with data relationships
    - Systems analysis—enable analysts to understand overall system design and data flow, and to find where data interact with various processes or components
    - Data integration—clear definitions of data elements provide the contextual understanding needed when deciding how to map one data system to another, or whether to subset, merge, stack, or transform data for a specific use
    - Decision-making—assist in planning data collection, project development, and other collaborative efforts

- **Version Control**
    - Processes should be developed for version control for field data management to ensure users have the most recent data and there are not multiple versions of the same data.
    - One final version should be maintained.
    - Interim versions may be useful for repeating an analysis. Organizational policy or project standards may specify keeping or deleting interim versions. If interim versions are kept, they should be archived with documentation to identify the date created and date archived.

Data Sharing and Transfer

Almost all environmental data have a spatial component. Developing standards across an organization and between organizations allows spatial data to be linked to attribute data. Data are more useful when they can be displayed spatially, over time and linked or viewed with other data sets to provide a better overall picture of the data. GIS can be used to link spatial and nonspatial data. For example, point data representing monitoring well locations can be linked with nonspatial attribute data such as well construction information or analytical data from samples collected at the monitoring well.

Many organizations have standards for data submittals. See the Resources section below for a list of some organization requirements for environmental geospatial data submittal and the organization standards subtopic (Organization Standards for Management of Geospatial Data subtopic sheet). The Resources section also contains guidance used in the geospatial and environmental industries for developing data standards and QAPPs. Here are some best practices to develop standards to improve data sharing and transfer:

- **Valid Values**
    - Many regulatory organizations require coordinate information and geographic data for submissions. Organizations should develop a set of valid values for spatial information. Valid values should be used whenever possible for all spatial data so users can easily compile multiple data sources into a single data set.
    - Environmental data managers should have a plan for how to handle data that are collected at the same location over an extended period. For example, are only the most recent results stored? Are data archived or deleted when new results are obtained at the same location?
    - If possible, organizations should adopt valid values already developed by other organizations to increase interoperability and data sharing.

- If possible, the same standards should be adopted across administrative boundaries (that is, water boards, towns/cities, states, federal, and various regulatory organizations).
- Different programs within each state organization will access, consume, and analyze geospatial data differently. It is best practice to consider all potential uses of data for your organization, other organizations, or individual users.
- Using the same valid values may not always be possible, but having similar attribute data in GIS is helpful for analysis across administrative boundaries. The Hydrologic Unit Map codes discussed in the data storage section are a good example of consistency across the United States.

- See the Valid Values Fact Sheet for a detailed discussion of valid values. For geospatial environmental data, here are some common location fields that should be included in attribute information and have predefined valid values:
    - Location name.
    - Location type or classification. The location type or classification can be used to group like samples together for analysis or modeling purposes. It may also be used to identify where the sample was collected. An example is USEPA's Facility Registry Service GEO_SUB_TYPE_LK that contains codes of operable units or subunits of facility sites such as wellhead, stack, or point of record.
    - Coordinate information.
        - X (longitude) and Y (latitude) values.
        - Coordinate codes to identify spatial reference—The organization should determine acceptable coordinate systems and unique codes to identify that coordinate system.
        - Spatial reference domains can be set to ensure the data fall within the intended area. A spatial reference domain is the allowable set of values for coordinate and elevations values (x, y, z) in a feature class. See Spatialreference.org for a list of spatial references.
        - Method for coordinate collection. Accuracy of the location coordinates is identified based on the collection method.
            - There are various methods to collect coordinates (approximation based on aerial, digitization from field sketches or paper maps, approximation based on orthoimagery, captured using mobile devices such as phones or tablets, GPS, or surveyed). Organizations should determine the methods that are acceptable based on intended use and quality desired of the data.
            - The method of coordinate collection and approximate accuracy should be stored as attribute information in separate fields.

    - Depth/elevation.
        - Depths and elevations are often components of environmental data. They should be stored in a consistent manner for each organization.
        - Source of elevation, date elevation data collected, and approximate accuracy should be stored as attribute information.
        - Depth and elevation can be recorded from various surfaces at a location on a given date, so the reference point for measurement needs to be recorded. For example, elevation can be collected at the top of a well casing, at ground surface, or both. If using depth below ground surface, a ground surface elevation, including datum from which ground surface is measured, should also be provided for modeling purposes.

- Date collected and datum elevation are important because of potential epoch differences (see the Geospatial Data Collection Consistency subtopic sheet) and possible freeze/thaw cycles causing shift of the monitoring well or other location of sample collection. Because of these potential differences of location values over time, organizations must determine and document in the project planning stage if they will keep all coordinate information for the life of the project, or if coordinate locations will be determined every time the site is visited and samples collected. If a site is on the east or west coast of the United States, the geospatial data manager will need to consider and manage for coordinate shifts.

- **Geospatial Data Submittals**
    - Organizations should define format requirements for submittal.
    - Receiving data in the same format allows the organization to combine data from all sources.
    - Government organizations should encourage electronic data submittals so data are easy to extract.
    - Approaches to electronic data submittal can include:
        - Online portal for submittals
        - Email
        - Specific email address for each submittal type can be used to sort data
        - Specific format with email subject line
        - Required information included in email body
        - Unique identifier for site for tracking purposes and ensure uniqueness of data sets
    - The Electronic Data Deliverables and Data Exchange Fact Sheet provides further information on best practices for electronic data submittals and data exchange.

Resources

- Organization requirements for geospatial environmental data standards and submittal:
    - Delaware EQuIS Data Submittal: https://dnrec.alpha.delaware.gov/waste-hazardous/equis/
    - Indiana Spatial Data Collection Standards: https://www.in.gov/idem/resources/publications/spatial-data-collection-standards/
    - North Dakota GIS Standards: https://www.gis.nd.gov/standards
    - NYEDC data submittal manual: https://www.dec.ny.gov/docs/remediation_hudson_pdf/eddmanual.pdf
        - Location information requirements Section 4.1.4
    - NJDEP: https://www.nj.gov/dep/gis/standards.html
        - Submittal Standards for Surveyed Map Data—July 2017
        - NJDEP Mapping and Digital Standards—October 2013
    - NJDEP Administrative Requirements for GIS Deliverables: http://www.nj.gov/dep/srp/gis/adminstrative_requirements_for_gis_deliverables.pdf
    - Washington Department of Ecology https://apps.ecology.wa.gov/eim/help/
    - Wyoming Department of Transportation Geographic Information Systems Standards: https://gis.wyoroad.info/docs/standards.html
- Industry guidance:
    - Guidance for Geospatial Data Quality Assurance Project Plans: EPA QA/G-5G (USEPA 2003): https://www.epa.gov/fedfac/guidance-geospatial-data-quality-assurance-project-plans
    - Federal Geographic Data Committee data standards: https://www.fgdc.gov/standards
    - Open Data standards: https://standards.theodi.org/

        - ESRI spatial reference https://spatialreference.org/ref/esri/
        - ITRC GRO-1: https://gro-1.itrcweb.org/
        - ASTM: https://www.astm.org/search/fullsite-search.html?query=Geographic%20information%20system
        - NSGIC: https://www.nsgic.org/
        - ISO: https://www.iso.org/committee/54904/x/catalogue/

- Software guidance:
        - Create drop-down lists in excel: https://support.microsoft.com/en-us/office/create-a-drop-down-list-7693307a-59ef-400a-b769-c5402dce407b
        - ArcGIS online Define attribute lists and ranges: https://doc.arcgis.com/en/arcgis-online/manage-data/define-attribute-lists-and-ranges.htm