**3.2** **Fundamental concepts of spatial data**

**Reading material**

|  |
| --- |
| Now let us start with the first lesson of the module: Fundamental concepts of spatial data  In this lesson, we are going to define the basic concept of primary and secondary data, and how can we describe them. The data can be displayed as vector or raster. It is used as input for QGIS and many other data analysis tools (refer to Module 1). |
| **Basic concepts:**  **Primary data vs. secondary data**  Spatial data is generally obtained from various sources. It can be collected from scratch, using direct spatial-data acquisition techniques, or indirectly, by making use of existing spatial data collected by others. The first source could include field survey data and remotely sensed images. To the second source belongs printed maps and existing digital data sets.  One way to obtain spatial data is by direct observation of relevant geographic phenomena. This can be done through ground-based field surveys or by using remote sensors on satellites or aircraft. Many Earth science disciplines have developed specific survey techniques as ground-based approaches remain the most important source of reliable data in many cases. |
| Data that are captured directly from the environment are called primary data. With primary data, the core concern in knowing their properties is to know the process by which they were captured, the parameters of any instruments used, and the rigour with which quality requirements were observed.  In contrast to direct methods of data capture, spatial data can also be sourced indirectly. This includes data derived by scanning existing printed maps, data digitized from a satellite image, processed data purchased from data-capture firms or international agencies, and so on. This type of data is known as secondary data. Secondary data are derived from existing sources and have been collected for other purposes, often not connected with the investigation at hand.  THE PROCESS OF DATA COLLECTION IN GIS  Source: <https://uizentrum.de/the-process-of-data-collection-in-gis/?lang=en> |
| Spatial data, or data related to a specific location on Earth, is collected into two formats or types: raster and vector. Both types of spatial data can be paired with attribute data, which describes any additional information that isn’t tied to location. For example, the coordinate's location of a building would be classified as spatial data, while the name of that building would be considered attribute data. |
| **Vector Data**  Data in this format consists of points, lines or polygons to simply the view of real-world as spatial data. At its simplest level, vector data comprises of individual points stored as coordinate pairs that indicate a physical location in the world. These points can be joined, in a particular order, to form lines or joined into closed areas to form polygons. Vector data is extremely useful for storing and representing data that has discrete boundaries, such as borders or building footprints, streets and other transport links, and location points. Ubiquitous online mapping portals, such as Google Maps and Open Street Maps, use data in this format.  **Raster Data**  Raster data provides a representation of the world as a surface divided up into a regular grid array, or cells, where each of these cells has an associated value. In an alternate sense, we can consider a digital photograph as an example of a raster dataset. Here each cell, which in this instance is referred to as a pixel, corresponds to a particular colour value (Digital Number). When transferred into a GIS setting, the cells in a raster grid can potentially represent other data values, such as temperature, rainfall, or elevation. The main point of difference between the digital photograph and the GIS representation is that in the GIS there is accompanying data detailing where the cells can be found on a globe and how big these cells can be. |
| Source: <https://spatialvision.com.au/blog-raster-and-vector-data-in-gis/>  Comparing raster and vector data is like comparing apples and oranges. To analyse and draw conclusions from GIS data, it has to be integrated. Scanned maps and some other raster files can be converted into vectors format – a process referred to as vectorization.  Converting pixels to points can take a great deal of time by hand, but luckily, GIS software can help speed up the process. The reverse process is also possible- and is referred to as rasterization. |
| **Attribute data**  Attribute data are the information linked to the geographic features (spatial data) that describes features, characteristics of geographic features that are quantitative and /or qualitative in nature, Attribute data is information appended in tabular format to spatial features. Attribute data can include text or numeric descriptors: i.e., nominal, ordinal, or interval/ratio data types Usually, a table is used to display attribute data, each row represents a single feature.  Attribute data are also known as non-geographic information data /non- spatial data. Attribute data helps to obtain the meaningful information of a map. Every feature has characteristics that we can describe.    Source: <https://storymaps.arcgis.com/stories/7c4d2d73eaf64263a438bcb9c6ece5a4> |
| **Attribute data models**  **A separate data model is used to store and maintain attribute data for GIS software.** These data models may exist internally within the GIS software or may be reflected in external Database Management Software (DBMS)- either commercial or open source . A variety of different data models exist for the storage and management of attribute data. The most common are:   * Tabular * Hierarchical * Network * Relational * Object Oriented |
| The tabular model is the manner in which most early GIS software packages stored their attribute data. The next three models are those most commonly implemented in database management systems (DBMS). The object oriented is newer but rapidly gaining in popularity for some applications.  The simple tabular model stores attribute data as sequential data files with fixed formats (or comma delimited for ASCII data), for the location of attribute values in a predefined record structure. This type of data model is outdated in the GIS arena. It lacks any method of checking data integrity, as well as being inefficient with respect to data storage, e.g., limitedindexing capability for attributes or records, etc.  **The relational database model is the most widely accepted for managing the attributes of geographic data.** |
| **Relational Model**  The relational database organizes data in **tables**. Each table, is identified by a unique table name, and is organized by **rows** and **columns**. Each column within a table also has a unique name. Columns store the values for a specific attribute, e.g., cover group, tree height. Rows represent one record in the table. In a GIS each row is usually linked to a separate spatial feature, e.g., a forestry stand.  Accordingly, each row would be comprised of several columns, each column containing a specific value for that geographic feature. The following figure presents a sample table for forest inventory features. This table has 4 rows and 5 columns. The forest stand number would be the label for the spatial feature as well as the primary key for the database table. **This serves as the linkage between the spatial definition of the feature and the attribute data for the feature.** |
| Data is often stored in several tables. Tables can be joined or referenced to each other by common columns (relational fields). Usually, the common column is an identification number for a selected geographic feature, e.g., a forestry stand polygon number. This identification number acts as the primary key for the table. The ability to join tables through use of a common column is the essence of the relational model. Such relational joins are usually ad hoc in nature and form the basis of for querying in a relational GIS product. Unlike the other previously discussed database types, relationships are implicit in the character of the data as opposed to explicit characteristics of the database set up. |
| **The primary key** represents the attribute (column) whose value uniquely identifies a particular record (row) in the relation (table). The primary key may not contain missing values as multiple missing values would represent nonunique entities that violate the basic rule of the primary key. The primary key corresponds to an identical attribute in a secondary table (and possibly third, fourth, fifth, etc.) called a **foreign key**.  This results in all the information in the first table being directly related to the information in the second table via the primary and foreign keys, hence the term “relational” DBMS. With these links in place, tables within the database can be kept very simple, resulting in minimal computation time and file complexity. This process can be repeated over many tables as long as each contains a foreign key that corresponds to another table’s primary key.    Source: <https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s09-02-geospatial-database-management.html> |
| There are many different designs of DBMSs, but in GIS the relational design has been the most useful. In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together.  This surprisingly simple design has been so widely used primarily because of its flexibility and very wide deployment in applications both within and without GIS.    Source: <http://wiki.gis.com/wiki/index.php/Database_management_system> |
| In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together.  In fact, most GIS software provides an internal relational data model, as well as support from external relational DBMS' such as MySQL, PostGIS, PostgreSQL etc. This approach supports both users with small data sets, where an internal data model is sufficient, and customers with larger data sets who utilize a DBMS for other corporate data storage requirements. With an external DBMS the GIS software can simply connect to the database, and the user can make use of the inherent capabilities of the DBMS. External DBMS' tend to have much more extensive querying and data integrity capabilities than the GIS' internal relational model. The emergence and use of the external DBMS is a trend that has resulted in the proliferation of GIS technology into more traditional data processing environments. |
| **The relational DBMS is attractive because of its:**   * simplicity in organization and data modelling. * flexibility - data can be manipulated in an ad hoc manner by joining tables. * efficiency of storage - by the proper design of data tables redundant data can be minimized; and * the non-procedural nature - queries on a relational database do not need to take into account the internal organization of the data. |
| The relational DBMS has emerged as the dominant data management tool in GIS implementation and application.  The following diagram illustrates the basic linkage between a vector spatial data (topologic model) and attributes maintained in a relational database file.    Source: <https://studfile.net/preview/2069449/page:3/> |
| Data dimensions in EO  There are three dimensions of spatial data:   1. **Spatial**: spatial dimension of data includes various characters or symbols that communicate to the user information about the location of the feature being observed 2. **Temporal**: The temporal dimension provides a record of when the data were collected (or the record to which data applies). 3. **Thematic/attribute**: The thematic dimension shows the characteristic of a real-world feature to which the data refer. In GIS, thematic data are often referred as non-spatial, or attribute, data. |
| **Spatial data quality**  Data quality is the degree of data excellency that satisfy the given objective. In other words, completeness of attributes in order to achieve the given task can be termed as Data Quality. Data created from different channels with different techniques can have discrepancies in terms of resolution, orientation, and displacements. Data quality is a pillar in any GIS implementation and application as reliable data are indispensable to allow the user obtaining meaningful results. |
| Spatial Data quality can be categorized into **Data completeness, Data Precision, Data accuracy and Data Consistency**;   * Data Completeness: It is basically the measure of totality of features. A data set with minimal number of missing features can be termed as Complete-Data. * Data Precision: Precision can be termed as the degree of details that are displayed on a uniform space. More about precision: GIS Data: A Look at Accuracy, Precision, and Types of Errors * Data Accuracy: This can be termed as the discrepancy between the actual attributes value and coded attribute value. * Data Consistency: Data consistency can be termed as the absence of conflicts in a particular database. |
| But on the other side, Data quality is a relatively abstract construct that is sometimes difficult to interpret. Because the quality depends on the applications or the intended use of data. In short, “Data quality” refers to the **fitness for use of data** for intended applications (Chrisman, 1983). |
| **Criteria for defining a GEOSPATIAL data quality:**   * Reliable and accurate * Current and up to date for the applications * Relevant and timely for the applications * Complete and precise * Format that can be easily maintained, transmitted, distributed, classified, updated etc. * Adequately protected i.e., control access to data integrity * Should have an associated metadata |
| **Four generic measures of data quality**   1. **Accuracy**: degree to which data agree with the values or descriptions of real-world features that they represent. It is usually application specific. Accuracy is one of the most important factors governing the cost of data collection. 2. **Precision**: tells one or indicate the measures how exactly data are measured and stored. High precision does not necessarily mean high accuracy while high accura-cy does not necessarily always require high-precision data representation. Precision has different meaning when applied to categorical data. 3. **Error**: measure of error is relative to measure of accuracy in that highly accurate data are supposed to be free of errors. Three major types of errors during measure-ments and observations:    1. gross error    2. systematic errors    3. random errors (e.g., Total error = Gross+ Systematic + Random) 4. **Uncertainty**: certain degree of doubt about the applicability or validity of the information derived from the data. Simply it is the lack of confidence in the use of the data due to incomplete knowledge of the data. |
| For more details about GIS data quality please refer to the following link <https://www.e-education.psu.edu/geog160/node/1922> |
| **Metadata for EO**  Metadata is a summary document providing content, quality, type, creation, and spatial information about a data set. It is like an instruction manual for data because it describes the who, what, when, where, why, and how of data. It has to be detailed, dependable, and well-documented It can be stored in any format such as a text file, Extensible Markup Language (XML), or database record. Because of its small size compared to the data it describes, metadata is more easily shareable. By creating metadata and sharing it with others, information about existing data becomes readily available to anyone seeking it. Metadata makes data discovery easier and reduces data duplication. |
| Below is the necessary information that have to be incorporated in the metadata file:  GIS Metadata  Source: [https://gisgeography.com/gis-metadata/](https://gisgeography.com/gis-metadata/%20%20%20) |
| **GIS metadata identification** provides a brief narrative of your data. In other words, it summarizes the purpose of your data in a succinct way. For example, identification assigns the following to your metadata:   * Title – Name of the data set * Description – The features in the data set and what they represent. * Keywords – By adding keywords, it helps categorize your data with predefined taxonomy. |
| **GIS metadata contact** includes details on who developed and makes the data available. For example, it includes the following three entities: 1.) the ***Originator, the entity that*** developed the data set, 2.) ***the Publisher, the entity that*** assists in producing, editing, and finalizing the end product. and 3.) the ***Distributor, whose*** main focus is to make the data available.  **GIS metadata quality** explains the accuracy and standards the data set followed. For example, it includes **horizontal and vertical positional accuracy**. Quality also evaluates tests of quality including **completeness, integrity, and inspections** of the data.  **Spatial Reference** information assigns a geographic extent and coordinate system Projection information includes a projection, datum, and units. Geographic extent comes in the form of a bounding box, place keyword, or thumbnail. |
| **Entity and Attribute**: entities refer to the map data type such as points, lines, polygons, or grids. The purpose of this metadata item is to describe how to represent the spatial information in the data. For the entity attributes, it includes a description with a list of valid values and domains.  **Lineage** describes in detail the creation of the data. For example, it lists the processing steps and responsible parties. Each processing step has a date when it took place so users can track changes. It’s like a changelog listing the evolution of the data from start to finish. |
| **Legal** section outlines the constraints for accessing and distributing the data. It describes the liability to assure the protection of privacy and intellectual property. Metadata includes a security classification that handles the restriction over security concerns. For example, confidential, restricted, sensitive, unrestricted, and unclassified are examples of security classification in metadata.   * **Temporal**: Temporal information focuses on when the data were collected or updated and how long it’s valid for. It also states the progress such as when there will be future updates. The frequency of updates can be anywhere from daily, weekly, monthly, or annually. * **Metadata Reference:** the metadata reference section is specific to the metadata. It gives a point of contact when there are uncertainties such as how to cite information when used. The metadata reference has a temporal component for when the user created and will revisit it next. * **Metadata Standard:** For GIS metadata standards, geographic data providers follow guidelines from the Federal Geographic Data Committee (FGDC), ISO 19115, EPA, Esri, Inspire, and MEDIN. Each schema was developed to best suit their particular requirements and needs. More on this later. |
| **Spatial data collection process**  There are mainly five stages of data collection process:   1. **Planning:** this stage includes establishment user requirements, gatherings resources and developing project plan. 2. **Preparation:** this involves obtaining data, redrafting poor quality map sources, editing scanned map images, removing noise, setting up appropriate GIS hardware and soft-ware to accept data 3. **Digitizing and transfer** are the stages where most of the effort will be expended. 4. **Editing and improvement** covers many techniques designed to validate data, as well as correct errors and improve quality. 5. **Evaluation is** the process of identifying project successes and failures.   https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQ3Bm6QTdXRBTXjrZP5_YxtGzdL4Q5UibN6DQ&usqp=CAU  Source: Muvunyi Germain, INES Ruhengeri (lecture notes) |
| Data collection and production are the most cumbersome and time-consuming steps of any kind of GIS and remote sensing techniques. They cover 75% of the total implementation cost for most of GIS project.  The processes of data collection are also variously referred to as data capture, data automation, data conversion, data transfer, data translation, and digitizing. |

**Exercise materials and tasks**

**Quiz questions**

Please answer the following questions to test your understanding so far:

1. Please complete by filling in the blank space with the correct term.

Data that is captured directly from the environment is called \_\_\_\_\_\_\_ data. (**primary**/secondary)

2. Spatial data, or data related to a specific location on earth, is collected in which format?

1. **Raster**
2. **Vector**
3. Metadata

What are three dimensions of spatial data?

1. Accuracy
2. **Spatial**
3. Precision
4. Metadata
5. **Temporal**
6. **Thematic/attribute**
7. Error

What criteria is defining geospatial data quality?

1. **Current and up to date**
2. **Reliable and accurate**
3. No need to have metadata
4. **Complete and precise**
5. Does not have to be protected

Sort in order stages of data collection:

preparation, editing and improvement, evaluation, planning, digitizing and transfer

|  |  |
| --- | --- |
| Step 1 |  |
| Step 2 |  |
| Step 3 |  |
| Step 4 |  |
| Step 5 |  |

Answer:

|  |  |
| --- | --- |
| Step 1 | planning |
| Step 2 | **preparation** |
| Step 3 | **digitizing and transfer** |
| Step 4 | **editing and improvement** |
| Step 5 | **evaluation** |