[DM-01-004] Geodatabases

Abstract

Geodatabases are specialized databases designed to store, manage, query, and manipulate geographic data. Geographic data is spatial data using a mathematical model of the Earth as a reference system. Given the increasing importance of geographic data in everyday life, for example for navigation and emergency response, it is no surprise that more and more data is collected and larger and larger datasets become available. Geodatabases are the system of choice for hosting, manipulating, querying, and storing these data. There are three main types of geodatabases provided by ESRI, namely file, mobile, and enterprise geodatabases. Geodatabases have also become more important for sharing data, following the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles. For this, the concept of Spatial Data Infrastructure is of crucial importance. Interoperability standards, for example those defined by the Open Geospatial Consortium, as well as metadata standards such as the ISO 191** series, organize the exchange and reusability of geospatial data. Geodatabases can use SQL queries including spatial queries and using topological relationships to carry out location-based analyses. Geodatabases are furthermore used to integrate interdisciplinary data on geographic research topics that explore questions of space and place, often across multiple spatial and temporal scales. Geodatabase integration with data servers, following a spatial data infrastructure, is especially useful for data distribution and making data publicly available, e.g., by entities and institutions with large repositories of geographic data, including state agencies or other organizations. To summarize, geodatabases are useful systems for working with geographic data and for supporting location- and topology- based analyses.

Author & citation

Koch, J., Ruiz Mendoza, F., Masiliunas, D., and Brecheisen, Z. (2025). Geodatabases. The Geographic Information Science & Technology Body of Knowledge (Issue 1, 2025 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2025.1.2.

Explanation

1. Definitions
2. Introduction
3. Geodatabase Types
4. Interoperability and Versioning
5. Geospatial Analysis and Querying
6. Conclusion

1. Definitions

Attribute: An attribute is a record of a non-spatial characteristic of a geographic feature.

Geographic Feature: A geographic feature is a part of or an object on the Earth’s surface that can be represented on a map.

Geospatial Data: Information on the location of a geographic feature on the Earth's surface.

Field (database): A field is a column entry in the attribute table that holds the value for a geographic feature.

Field (data model): A field is a surface that represents the value of a geographic phenomenon, such as elevation or temperature.

Object (data model): An object is a distinct geographic feature, such as a building or a tree.

Query: A logical expression applied to a database to retrieve, manipulate, or delete data stored in the database.

2. Introduction

The terms geographic database (or geodatabase) and spatial database are often used synonymously. However, geographic databases and spatial databases differ in the type of space they were developed for, rendering this synonymous use of terms imprecise. A spatial database "…is a general-purpose database […] that has been enhanced to include spatial data that represents objects defined in a geometric space, along with tools for querying and analyzing such data" (Spatial Database 2024). In comparison, a geodatabase describes a spatial database specifically developed for working with geographic data (i.e., data referenced to geographic space). This makes a geodatabase a more specialized iteration of a spatial database. The use of geographic data, describing geographic features (e.g., road networks) and their characteristics, is becoming increasingly important with prominent applications such as emergency response (Hatzis et al. 2024). Geodatabases are a way to host, manipulate, query, and store these geographic data.

A geographic feature describes an entity located on or near the surface of the Earth. Geographic Coordinate Systems (GCS), reference systems based on mathematical models of the Earth, are used to describe the location of said features using the angular measures latitude and longitude (Chang 2026). Building on GCS, Projected Coordinate Systems (PCS, also referred to as Planar Coordinate Systems) use map projection methods to convert latitude and longitude coordinates to a 2D space and provide locations on the Earth as x and y coordinates in a Cartesian system. Knowledge of the different types of GCS and PCS is included in Geodatabases, among other things, via assigning a (geo-)spatial reference to each dataset in a database. To learn more about coordinate systems, refer to the entries on Geographic Coordinate Systems and Planar Coordinate Systems.

While storing data on the location and characteristics of geographic features in a geodatabase is important, there is often also an interest in making the data publicly available. For this purpose, a Spatial Data Infrastructure (SDI) formalizes the technology, standards, tools and resources for sharing geospatial data. These SDIs often embed geodatabases for sharing geospatial data with the public. SDIs frequently implement the "FAIR Guiding Principles for scientific data management and stewardship" (Wilkinson et al. 2016) by providing an infrastructure that makes geoinformation Findable, Accessible, Interoperable, and Reusable (FAIR). Recently, the Collective benefit, Authority to control, Responsibility, and Ethics (CARE) principles for Indigenous data governance (Carroll et al. 2020) have also received more attention.

The term geodatabase is also used to describe an Esri proprietary file format developed in the late 1990s. While Esri file formats are proprietary, the Geospatial Data Abstraction Library (GDAL) provides tools for working with geodatabases, making this file format also available (with limitations) in free and open-source GIS applications such as QGIS or GRASS GIS. Esri's geodatabase follows an object-relational database model, as do other database management systems (DBMS) such as SQLite and PostgreSQL. The object-relational model combines a relational database management system (RDBMS) with the strength of object-oriented programming. The PostGIS extension for PostgreSQL and GeoPackage/SpatiaLite extensions for SQLite allows them to work with geospatial data, i.e., providing a geodatabase application. To learn more about RDBMS, refer to the entry on Relational DBMS and their Spatial Extensions.

3. Geodatabase Types

While different geodatabase applications exist, there are also different types of geodatabases. The following section refers to the geodatabase types by ESRI. The three types are (a) file geodatabases, (b) mobile geodatabases, and (c) enterprise geodatabases. A fourth type of ESRI geodatabase, the personal geodatabase type, has been phased out with the transition from ArcMap to ArcGIS Pro.

3.1 File Geodatabases

One important advantage of a file geodatabase is that it is easy to create and use. Being a collection of different files (organized in a directory with the .gdb extension) combining a variety of geospatial and non-geospatial datasets in one database makes these databases well-suited for thematic applications. File geodatabases also provide the benefit of making a geodatabase straightforward to copy to other systems. While the lack of a complex organizational structure provides these advantages in addition to the benefit of fast read and write operations for smaller datasets, a file geodatabase also has several disadvantages. These include a limited fit for executing complex querying operations or shortcomings when handling multiple read and write operations at the same time. Editing operations are limited to one user at a time. While the default maximum size for files in a file geodatabase is one terabyte (TB), it can be extended up to 256 TB.

3.2 Enterprise Geodatabases

Enterprise geodatabases are stored in DBMS, and supported by PostgreSQL, Oracle, Microsoft SQL Server, and others. Some cloud implementations are also supported, e.g., Google Cloud SQL. This type of geodatabase is intended to allow multiple users to access and edit a geodatabase at the same time. If multi-user editing is not a requirement, it is advised to use a file geodatabase instead. It is also not recommended to store raster or imagery data in an enterprise geodatabase but instead use, e.g., cloud storage options for rasters/imagery and then link those into the geodatabase (Esri 2023). The limit on the number and size of datasets included in the enterprise geodatabase depends on the limitations of the underlying DBMS. Unlike file or mobile geodatabases, enterprise geodatabases do not come in a specific file format, but rather follow the format of the underlying DBMS. Permissions and security management are also organized through the underlying DBMS (Esri ArcGIS Pro Resources, n.d).

3.3 Mobile Geodatabases

The mobile geodatabase type was developed for use by a single app or user at a time. This geodatabase type is stored in SQLite and consists of a collection of datasets stored in a single file. The file comes with a .geodatabase file extension. There is a size limit of 2 TB for mobile geodatabases; the underlying operating system manages permissions and security of this geodatabase type (Esri ArcGIS Pro Resources, n.d).

4. Interoperability and Versioning

4.1 Interoperability

In the context of geodatabases, interoperability refers to the ability of different geodatabase systems to use and exchange geospatial data. The more people and organizations collaborate on a project with a geodatabase the more important interoperability becomes. Geodatabases can be integrated with non-GIS applications, allowing for a streamlined project workflow. Geodatabase servers can be accessed remotely, making them useful for collaboration and data-reuse.

Similarly important is the coordinate system. Geographic coordinate systems (GCS) use angles of rotation – latitude and longitude – to specify a geographic feature's location on the Earth, whereas projected coordinate systems (PCS) use a Cartesian coordinate system to give the location of a geographic feature on a plane (i.e., a "flat" map, physical or digital). Being aware of which datasets use which coordinate systems and how the datasets have to be processed to be suitable for a combined analysis is crucial. Moreover, all datasets in a geodatabase should be in a consistent GCS or PCS to allow for a correct geospatial analysis combining information from different datasets.

According to the Open Geospatial Consortium (OGC), "…a standard is an agreed specification of rules and guidelines about how to implement software interfaces and data encodings" (OGC 2017). The OGC defines such standards, which are freely and publicly available, not requiring license fees when used. The OGC defines encoding standards (e.g., for metadata or data models) and interface standards (e.g., for catalog or data services). Following these standards consistently ensures compatibility of different products, and hence, is important for interoperability.

The importance of metadata (data about data) for creators and users of geospatial data cannot be overstated. Documentation of the who, what, why, where, and when of a dataset is crucial for the sharing and reuse of geospatial data. A metadata schema provides a description of the elements to be included in the metadata (e.g., title, date, creator) and their structure. Different schemata exist for different types of data. The current standard for geographic information metadata is the International Organization for Standardization (ISO) 19115-1:2014. ISO 19115-3 provides the definition of the XML schema implementation for the Geographic Information metadata implementation. In the USA, the Federal Geographic Data Committee (FGDC) has authored and endorsed the Content Standard for Digital Geospatial Metadata (CSDGM). However, in 2010 the FGDC endorsed the ISO 19115 and recommends migrating from the CSDGM to the ISO geospatial metadata standard (FGDC, n.d.). The European Commission's Infrastructure for Spatial Information in Europe (INSPIRE) also endorses ISO 19115.

4.2 Versioning

In the geodatabase context, versioning can be defined as a mechanism allowing editing of a geodatabase by multiple users at the same time without interfering with each other's work. This is especially important for good data management when working in a team, but there is also value to using versioning when working individually. One key advantage of using versioning with geodatabases is the creation of a historical record of changes made to the database. Furthermore, versioning helps to organize multiple workflows on the same geodatabase.

For workflow organization, a specific terminology is used. The term "default version" describes the original version of the geodatabase, also referred to as the "parent version." For working on edits, a copy of the default version of the geodatabase is created, and the edits are implemented in this copied version, called the "child" version. With multiple users making changes to the same version, there is the potential for a conflict. The term conflict describes a situation where multiple users made changes to the same content or features of the geodatabase, which is per se not an issue, but it has to be addressed through conflict resolution, i.e., deciding which changes to keep and which ones to dismiss. Bringing back together different versions of a geodatabase requires comparing the differences and changes made to those versions and is referred to reconciling versions. Applying changes implemented in a child version of the geodatabase to the parent database is called posting.

For managing different versions and permissions, it is important to develop a shared understanding of how to manage different versions of a geodatabase and, of course, how to manage user permissions for different versions. Assigning appropriate permissions is important as it reduces the risk of unintentional changes to a database. Typical permissions include reading, adding, modifying, and removing data.

5. Geospatial Analysis and Querying

Geodatabases provide functionality similar to non-geospatial databases, allowing for working with and querying the non-spatial attribute data of a database. The special feature of geodatabases, however, is the availability of location information, which supports querying of data using geographic information and geospatial functions to answer questions. The language used in geodatabases and other GIS software applications is the Structured Query Language (SQL) – a language specifically designed to construct queries and used widely in RDBMS. An overview and short introduction to SQL can be found under the entry for Structured Query Language (SQL) and attribute queries.

For example, SQL can be used to create a query selecting a subset of entries from the geodatabase where a certain (geospatial) condition is met. Queries can range from simple to complex, and geodatabases provide the option to use topological relationships in queries. Topology is an important concept for spatial analyses, as it describes the spatial relationships between adjacent features. In general, geodatabases focus on the vector data model and work with points, linestrings, polygons, multipoints, multilinestrings, and multipolygons. The latter three are referred to as multi-geometry types, which combine similar features in one dataset. For example, a multipoint dataset combines multiple point datasets into one. By doing so, these multi-geometry types help to efficiently manage spatial data, making sure that features are handled as one entity, which is important for data consistency. A special feature of geodatabases is that location information and

information on topological relationships are available. This information can be used to create geospatial queries or used in the form of functions to answer geospatial questions. The relationships fall into different categories such as point-point, point-line, point-polygon, etc.

For example, one might want to find out which oak tree(s), represented by points, are located in a certain park named Park A, represented by a polygon. For this, we assume that the geospatial database has two feature classes. The first is called my_parks and contains information on the park boundaries (represented as polygons) and non-spatial park characteristics (e.g., the name of the parks). The second feature class is called my_trees and contains information on the location of individual trees (represented as points) in the area of interest and non-spatial tree characteristics (e.g., the genus of the individual trees). For this example, we can use a topological relationship to identify the tree(s). Figure 1 shows a schematic view of this example.

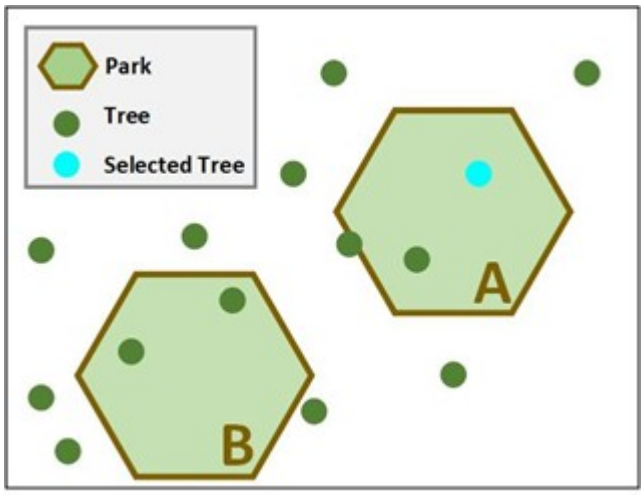


Figure 1. One oak tree (represented by a point highlighted in blue) is located within Park A (represented by a green polygon with a brown boundary), while other trees (represented by dark green points) are either not oaks or not located within Park A. Source: author.

The example describes a “within” topological relationship, a situation where one geographic feature (tree, represented by a point) is completely located within another feature (park, represented by a polygon). In a geodatabase, we can conduct this selection using an SQL query. This example query is designed to work in a PostGIS geodatabase application:

SELECT t.*

FROM my_trees t

WHERE t.genus = 'Oak'

AND ST_Within(t.geom, (SELECT geom FROM my_parks WHERE name = 'Park A'));

The query is a Boolean expression that can be evaluated into TRUE or FALSE. If an entry (i.e., a row in a spatial table representing a geographic feature) in the geodatabase returns

TRUE, it will be included in the selection and if it returns FALSE, it will not be included. We want to select all trees of the genus Oak which are located in Park A.

The first line of the query starts the selection with **SELECT** and uses the alias **t** for the spatial table **my_trees**, and the wildcard * indicates that all attributes (or columns) will be included in the selection. The second line of the query identifies the spatial table to select from, i.e., **my_trees**, and assigns the alias **t**. The third line specifies selecting the rows where the entry in the genus column of t equals “Oak”. The fourth line of the query starts with the logical operator **AND** to combine another condition with the selection. The part following the **AND** is where the geospatial relationship comes into play. The **ST_Within** function takes two geometries, geometry X and geometry Y, and returns TRUE if X is within Y. In this example, X is t.geom, and Y is the geometry resulting from the query in parentheses, i.e., **SELECT geom FROM my_parks WHERE name = 'Park A'**. The result of this sub-query is the geometry of Park A since it selects the geometry of the entry from the my_parks spatial table where the name equals “Park A”. Hence, the result of executing the entire query is the selection of the tree(s) of the genus Oak located within the boundaries of Park A (Figure 1).

The key to bringing location into PostGIS analyses are the geometry and the geography types. These types are PostGIS’ way of encoding location information as point, linestring, polygon, etc. The geometry type that operates on a Cartesian plane and is best suited for working on the local to regional scale. In terms of coordinate systems, PCS are the counterpart to geometry. When working at the global or continental scale, the type to work with would be geography, and GCS would be the counterpart to the geography type. Computations for the geography type are carried out on a spheroid, which makes them computationally more demanding than calculations on a plane. This can make operations with the geography type slower than calculations with the geometry type.

The “within” relationship is one of many relationships that can be tested for in geodatabases. Other relationships include “contains,” “shares,” “intersects,” “crosses,” “disjoints,” or “touches,” to only name a few. Since some of these relationships sound similar--but work differently--it is important to always consult the documentation for the respective function or operation. Figure 2 displays examples of some of the topological relationships that can be used in geodatabases for making selections.

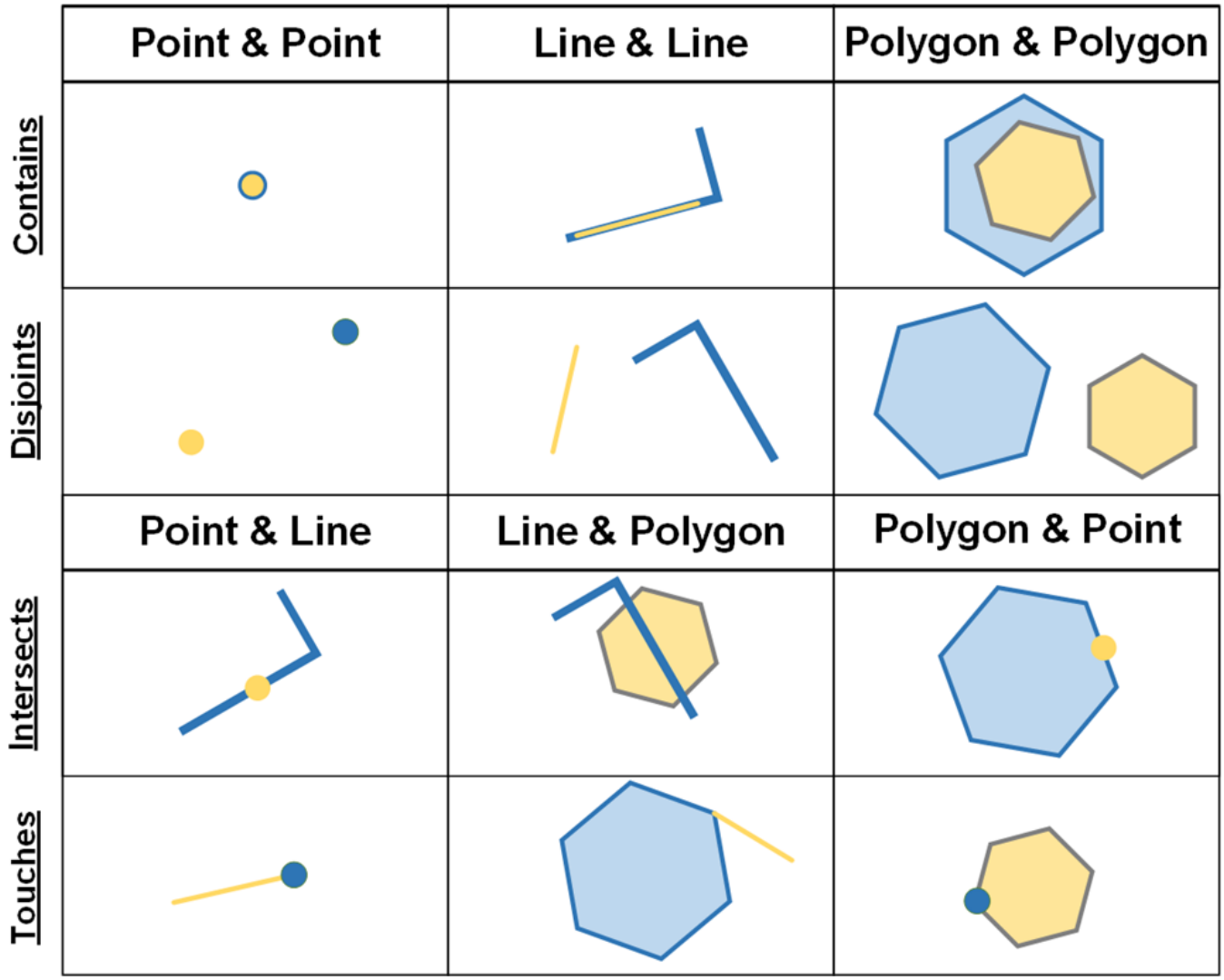


Figure 2. Examples of four different spatial relationships for data selection applied to a combination of different geometries. Source: author.

In combination with the other functionalities of geodatabases, these operations allow to answer geospatial questions. Selection is powerful but there is more functionality available in geodatabases that is very useful to answer geospatial questions. For example, spatial joins are important operations when working with geodatabases. A spatial join “combines represented geographic objects and their associated attributes based on a spatial relationship text (or predicate)” (Morgan 2023). To learn more about spatial joins, refer to the entry for Spatial Joins.

6. Conclusion

The term geodatabase describes a spatial database operating in geographic space, specified by either a geographic or projected coordinate system. The term geodatabase is also used to refer to proprietary ESRI file format. Many organizations use geodatabases to make their data freely and publicly available, and they also provide an opportunity for researchers, e.g., in socio-environmental systems research, to follow the FAIR data principles and accompany publishing their research findings by making geospatial data collections available. One example is the use of geodatabases for transboundary and international study regions since, in these situations, data collection and integration often takes a lot of time and effort as research is often confronted with different geodata

standards and inconsistent data, especially regarding data sources and attribute information (e.g., units given in metric versus imperial systems) (Plassin et al. 2020). With geodatabase types that have this option available, multi-user functionality provides an important advantage for collaborative work settings. To summarize, geodatabases are valuable systems that can help foster integrated research efforts and have the potential to also support participatory and community-engaged research and activities.

References

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., … Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal,19,1:43.

Chang, K.-T. (2019). Introduction to Geographic Information Systems, 9th Edition. McGraw-Hill Education, New York, NY.

Esri. (2023). Enterprise Geodatabase: An Introduction. Presentation by Chet Dobbins and Morgan Robinson, given at the 2023 ESRI User Conference.

Esri. (n.d.) Types of Geodatabases. ArcGIS Pro Resources. Accessed January 17, 2025.

Federal Geographic Data Committee (FGDC). (n.d.) Benefits of ISO Metadata. Accessed January 16, 2025.

Hatzis, J.J.; Kim, J.; and Klockow-McClain, E. (2024). An Agent-Based Modeling Approach to Protective Action Decision-Related Travel during Tornado Warnings, Natural Hazards Review, 25, 1.

Morgan, J. D. (2023). Spatial Joins. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2023 Edition), John P. Wilson (Ed.). DOI: 10.22224/gistbok/2023.1.1.

Open Geospatial Consortium (OGC). (2017). Introduction to OGC Standards. Accessed September 11, 2024.

Plassin, S.; Koch, J.; Paladino, S. et al. (2020). A socio-environmental geodatabase for integrative research in the transboundary Rio Grande/Río Bravo basin. Sci Data 7, 80.

Wikipedia contributors. (2024, December 19). Spatial database. In Wikipedia, The Free Encyclopedia. Retrieved 19:43, February 17, 2025.

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1):1-9.