



Course 5

Text-only version

Data management in a national forest inventory

The interactive version of this lesson is available free of charge at <https://elearning.fao.org/>



Some rights reserved. This work is available under a CC BY-NC-SA 3.0 IGO (<https://creativecommons.org/licenses/by-nc-sa/3.0/igo/>) licence.

In this course

Lesson 1: Basic concepts of data management	5
Lesson Introduction.....	Errore. Il segnalibro non è definito.
Basic terminology for database management.....	5
Requirements for forest information systems	13
Summary	14
Lesson 2: Collecting and managing field data	16
Lesson introduction.....	16
NFI data tools: An overview.....	16
Data collection for NFIs	18
Data management for NFIs	20
Databases in NFI.....	20
Building a schema for a database.....	21
Data validation	22
Summary	26
Lesson 3: Online reporting and data archiving.....	27
Lesson introduction.....	27
Data dashboards and portals	27
Online analytical processing (OLAP)	28
Data archiving.....	30
Summary	33

About the course

This course provides an overview of information gathering and data management for national forest inventories (NFIs).

Who is this course for?

This course is primarily intended for people who are involved in NFIs, especially with an interest in the methods and tools for management, processing and sharing of NFI data. Specifically, this course targets:

1. Decision-makers planning investments for implementing NFIs
2. Data managers and NFI team leaders
3. Data technicians responsible for data analysis
4. Students, as curriculum material in forestry schools
5. Youth and new generations of foresters

Course structure

There are three lessons in this course.

Lesson 1: Basic concepts of data management

This lesson outlines the role of Information and Communication Technology (ICT) in forest data management. As this lesson also aims to define some ICT-related forestry terminology, it serves as your primer to the more complex lessons in this course.

Lesson 2: Collecting and managing field data


This lesson focuses on managing field data, and the processes that go into making the data ready for long-term storage and analyses.

Lesson 3: Online reporting and data archiving

This lesson explores OLAP (Online Analytical Processing) technology and describes the importance of metadata and microdata in collecting and storing forest information.

About the series

This course is the fifth in a series of eight self-paced courses covering various aspects of an NFI. Here's a look at the complete series:

Course	You will learn about
Course 1: Why a national forest inventory?	Goals and purpose of an NFI and how NFIs inform policy- and decision-making in the forest sector.
Course 2: Preparing for a national forest inventory	The planning and work required to set up an efficient NFI or a National Forest Monitoring System (NFMS).
Course 3: Introduction to sampling	Introduction to sampling
Course 4: Introduction to fieldwork	Considerations for fieldwork, plot-level variables and tree-level measurements.
 Course 5: Data management in a national forest inventory	(You are currently studying this course)
Course 6: Quality assurance and quality control in a national forest inventory	QA and QC procedures in forest inventory data collection and management.
Course 7: Elements in data analysis	Typical approaches/calculations in data analyses and related topics.
Course 8: National forest inventory results: Reporting and dissemination	NFI reporting and the importance of reporting in the context of REDD+ actions.

Lesson 1: Basic concepts of data management

Lesson Introduction

In this lesson, we will understand the role of Information and Communication Technology (ICT) in forest data management.

In order to do this, we will look at the ICT-related terms, techniques and methods that are used to manage NFI data. We will also look at the requirements of a forest information system.

Learning objectives

At the end of this lesson, you will be able to:

1. Outline the basic terminology and concepts for data management in database technology.
2. Explain the key requirements for an information system on national forest monitoring.

Basic terminology for database management

Understanding the key terminology associated with database management will help you get a hold on more complex concepts as we progress through the course. Let's begin this course by looking at the ICT terminology that is relevant for NFI data gathering and organization.

Information and communication technology (ICT)

Information and communication technology (ICT) is an **umbrella term** that includes the design, development, implementation, support and management of computer-based information systems. In essence, ICT deals with the use of computers and software to convert, store, protect, process, transmit and retrieve information.

Data

A collection of facts, such as numbers, words, measurements, images, observations or just descriptions of things. **Information** refers to the meaningful output obtained after processing data. Our numbers will not make any difference if we do not translate them into meaningful stories.

Data integrity

The term data integrity refers to the accuracy and consistency of data. When creating databases, attention needs to be given to data integrity and how to maintain it. A good database will enforce data integrity whenever possible.

For example, a user could accidentally try to enter a vegetation type code into a date field. If the system enforces data integrity, it will prevent the user from making these mistakes, through a warning or error message when wrong data is entered into a specific field. This is one of the advantages of using online field forms in the field data collection. Of course, not all erroneous input can be identified by this formal check.

Data management

An administrative process that includes acquiring, validating, storing, protecting and processing data to ensure its accessibility, reliability and timeliness for users. A fundamental principle of forest data management is to store all data as they were collected in their original form. This allows flexibility in the way data can be processed and ensures that all calculations are reproduced from the original data. Original data also serves as the 'base data' and can be important for counterchecking errors noticed during data validation and/or data analysis.

Database and Database Management System (DBMS)

A database is a **collection of data held in a computer system in an organized form** for easier access and management. Databases are basically containers for data. In NFIs, databases are used to manage and archive field inventory data, field photographs, maps and remote sensing data, and related documents (such as field manuals, guidelines, inventory reports).

Database Management Systems, commonly referred to as DBMS, are **software that allows us to perform various operations on databases**. DBMS enable users to access databases, as well as manipulate, report, and represent data. They also help control access to the database. Some examples of popular database software or DBMSs include MySQL, Microsoft Access, Microsoft SQL Server, PostgreSQL, FileMaker Pro, Oracle Database, and InterBase.

Database access language

A database access language is used to **access, enter, update, or retrieve data** from the DBMS. It is a key computing tool to organize data, create databases and control data using query languages. There are several such languages that can be used for this purpose, including SQL (Structured Query Language).



Video resources

What is Database & SQL?

<https://www.guru99.com/introduction-to-database-sql.html>

This Database tutorial explains the concept of DBMS (Database Management System). To help beginners, it cites examples of real-life data base management systems. It explains types of DBMS for beginners. It explains how SQL works. This video course is a complete introduction to Database. Click on the time points below to view different sections!

Learn Basic SQL in 10 Minutes

<https://www.youtube.com/watch?v=bEtnYWuo2Bw&t=1s>

Schema

The **structure of the database** is called the schema, and it defines what type of data is being stored. You can think of it as the skeleton structure that represents the logical view of the entire database. It defines how the data is organized and how its components are related. It formulates all the constraints that are to be applied on the data. The key factors to consider about data when designing a database include:

1. **Data types**, which are used in each field and they classify data values that have common properties;
2. **Validation**, which are the rules for accepting data; and
3. **Key field, or the primary key** (you will learn more about keys later in this lesson).

Developers plan a database schema in advance, so they know what components are necessary and how they will connect to each other. This is also important in the NFI context: designing a smart data base schema will facilitate analyses.

Entity and attribute

A data **entity** is an **object in a data repository**. Data entities are the objects of the data model such as 'tree' or 'soil'. Entities do not represent any data by themselves but are containers for attributes and relationships between objects.

A data **attribute** is a **unit of information inside a data entity**, i.e. it is a single-value descriptor for a data point or data object. These are like properties of data entities: for example, an entity ‘tree’ can contain the following attributes: species, height and age.

Data models

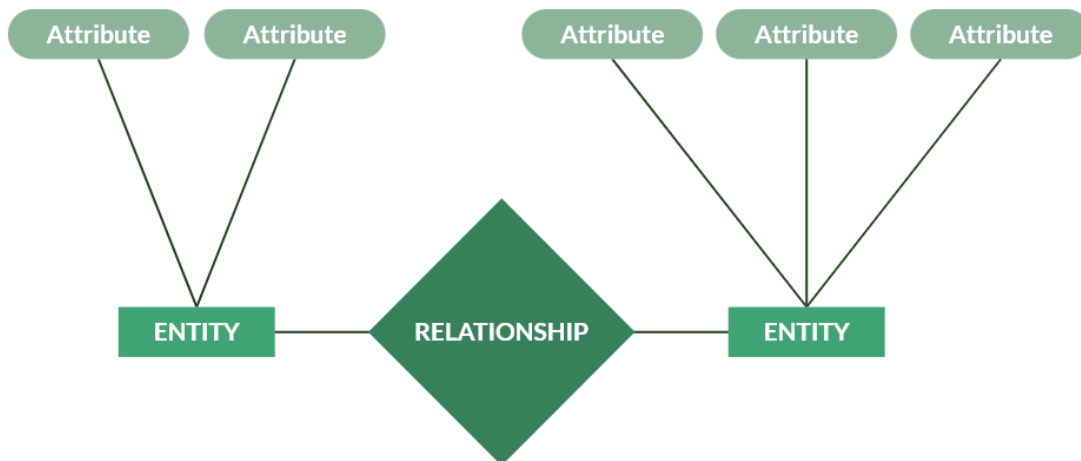
Data models define the **logical structure of a database**. They define how data is connected to each other and how it is processed and stored inside the DBMS.

A data model helps one to perceive, organize and describe data in a conceptual schema that includes both the data and the operations for manipulating the data set. It is worthwhile to now consider the two main data models: the **Entity-Relationship Model**, and the Relational Data Model.

Entity-relationship model

As the name suggests, the Entity-Relationship (ER) Model is **based on the notion of real-world entities and relationships among them**. While formulating real-world scenarios into the database model, the ER Model creates an entity set, relationship set, general attributes and constraints.

The ER Model is best used for the conceptual design of a database. Conceptual design is the first stage in the database design process. The goal at this stage is to design a database that is independent of database software and physical details. The output of this process is a conceptual data model that describes the main data entities, attributes, relationships, and constraints of a given problem domain.



Relational data model

Relational databases, which are the **most widely used data model in DBMS**, have become the leading technology for managing databases, applicable from smartphones and personal computers to

mainframes.

The relational data model **represents data in the form of tables**. A relation is a named, two-dimensional table of data. Each relation (or table) consists of a set of named columns and an arbitrary number of rows. An attribute is a named column of a relation. Each row of a relation corresponds to a record that contains data (attribute) values for a specific entity.

id	species_name	dbh	tree_height_calc	tree_basal_area	tree_volume_stem	tree_volume_bole	tree_biomass_ag	tree_biomass_bg
1001	Trichila prieuriana	28	12	0.061575216	0.038792386	0.024630086	0.041559922	0.011221179
1002	Acacia nilotica	40	18	0.125663706	0.118752202	0.100530965	0.123853538	0.033440455
1003	Acacia nigrescens	22	14	0.038013271	0.027939754	0.025145779	0.030169727	0.008145826
1004	Trichila emetica	24	14	0.045238934	0.033250617	0.029925555	0.035754825	0.009653803
1005	Toona ciliata	28	10	0.061575216	0.032326988	0.024630086	0.034785146	0.009391989
1006	Trichila prieuriana	32	10	0.080424772	0.042223005	0.032169909	0.045143384	0.012188714
1007	Trichila dregeana	32	14	0.080424772	0.059112207	0.048254863	0.062692425	0.016926955
1008	Acacia nilotica	28	14	0.061575216	0.045257784	0.024630086	0.048307525	0.013043032
1009	Trichila emetica	68	18	0.363168111	0.343193865	0.217900866	0.348935144	0.094212489
1010	Trichila emetica	20	16	0.031415927	0.026389378	0.025132741	0.02853468	0.007704364

A view on a relation (table) in a database.

A record is a complete row of data elements in a table. Look at the example here. Tree number 1 005 in the table has a record—highlighted in orange in the table above—where input data is all recorded in one table. This differs from a tuple (that typically arises through a query), which is a set of records that incorporates data from multiple tables. The data in a tuple have a particular reference value in common, e.g. all variable records (volume, biomass, stem value etc.), across different tables, belong to tree number 1 005. Therefore, data, from various tables, that form a tuple, all associate to one tree.



Did you know?

What are the main features of relational database models?

The major features of this model are:

- Data is stored in tables (called relations).
- Relations can be "normalized".
- Each relation in a database must have a distinct or unique name.
- A relation must not have two attributes with the same name. Each attribute must have a distinct name.
- Within one relation, each row must be unique.
- Each column in a relation contains values from the same domain. A domain is a unique set of values permitted for an attribute in a table. For example, a domain of month-of-year can accept January, February, up to December as possible values, and a domain of integers can accept whole numbers that are negative, positive and zero.



Video resources

Watch a video that explains database normalization for beginners.

[Basic Concept of Database Normalization - Simple Explanation for Beginners](#)

[Database Schema: Entity Relationship Diagram](#)

[Converting ER Diagrams to Schemas | SQL | Tutorial 23](#)

Data types

Data typing is a way of **classifying data values that have common properties**.

When you create a database, you need to set data types for each field. For example, in a forest inventory database you might need text fields for 'Recorder Name', but numbers for 'Tree Age'. Fields are usually restricted to one specific data type.

Different kinds of data values also need different amounts of memory to store them and have different operations that can be performed upon them. For example, if the data type is a character string, you may have numbers entered—but you cannot make calculations as the figures are understood by the software as characters. The most supported data types are:

- **Integers** or whole numbers, such as 5, 27, 65575;
- **Floating point numbers** (with decimal points, sometimes called real numbers, or floats) e.g. 5.2, 37.4, 196.247;
- **Characters** such as c, F, 3 \$, #;
- **Character strings** such as abc, deu496, 3erf08!@; and
- **Boolean values**, for instance True/False, or Yes/No.

A forest inventory database typically contains all these data types. Additionally, it can also contain files (photos, videos, audio files, digital documents), dates, times and complex types (i.e. combined attributes).

Examples of complex data types in a forest inventory database fall into two main categories: coordinate data and taxonomic data.

Coordinate data (points) can consist of a horizontal coordinate reference system (CRS), X- coordinate, and Y-coordinate. Vertical CRS contains the third dimension, such as Z or height or altitude. In compound CRSs horizontal and vertical system data can be joined.

Note: Coordinate reference system (CRS) and spatial reference system (SRS) are synonyms and are commonly interchanged. It is critically important not only to specify the coordinates but give also the information which reference system had been used. Without this information, coordinate values are close to meaningless.

Taxonomic data can contain data about family, genus, species, subspecies, variety. Taxonomic data can also contain the author citation that refers to listing the person (or team) who first makes a scientific name of a taxon available.



Quick tips!

In addition, an NFI database can be expanded into a **spatial database** that stores data about mapped objects (in the forms of points, lines, and polygons). These databases allow using spatial queries and spatial analysis to be used with NFI data

Identifier and primary key

Each entity requires a tag that uniquely identifies it. In a relational database, this is called a **primary key**. A primary key is therefore a **special relational database table column** (or combination of columns) **designated to uniquely identify each table record**. Without the primary key and closely related foreign key concepts, relational databases would not work. This is why we need to define a primary key for each entity while building a database schema. Remember that there can be more than one primary key for an entity.

Flat file database and data warehouse

A flat file database **stores data in a plain text file**. Each line of the text file holds one record, with fields separated by delimiters, such as commas or tabs. While it uses a simple structure, a flat file database cannot contain multiple tables like a relational database can. Flat file databases were developed and implemented in the early 1970s by IBM. Data warehousing projects use flat files to import data.

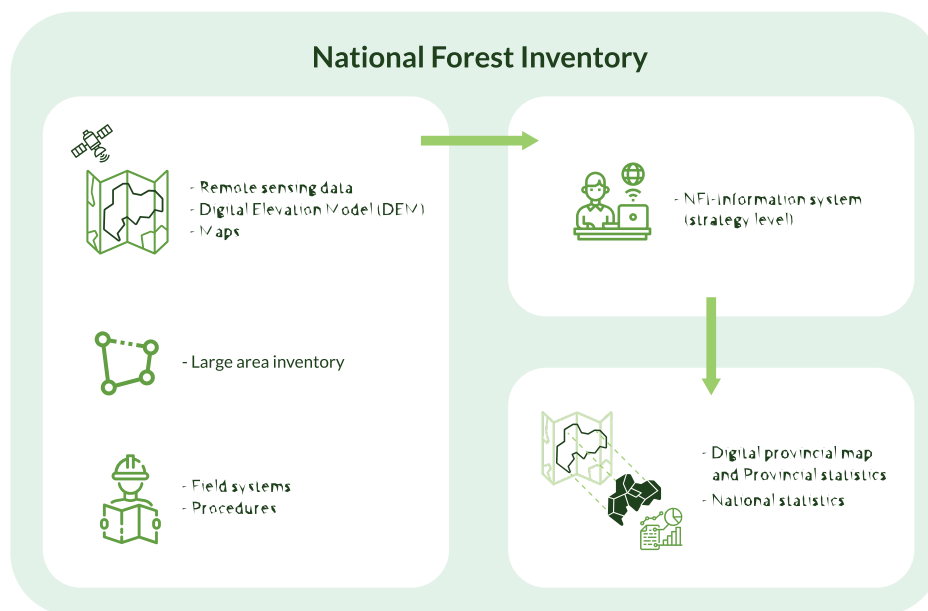
Data warehouses **collect content from various operational databases**, including personal, workgroup, department and enterprise resource planning (ERP) databases. They are systems used for reporting and data analysis and are considered to be a core component of business intelligence.

Country example: Korea Forest Research Institute

This example shows how to design a data warehouse to set up a sharing infrastructure for the spatial and field survey data on a forest information system. As you will see in the example, data warehouse architecture can be complex. Nowadays, larger NFI data management systems in countries with an institutionalized NFMS (e.g. NAFIDAS for the Swiss NFI and TaxWebb in Sweden), can be compared to a data warehouse architecture. None of these systems are better or worse than the other, and the decision of which system is suited for which local situation is based entirely on available recourses and needs.

Requirements for forest information systems

Forest information systems provide a platform to offer decision-makers with high-quality and comprehensive information that has undergone quality control processes. NFIs and other forest inventory data can be seen as essential elements within this concept.



A modern system for managing forest resource information requires the following:

1. **Adaptability to different hardware and software environments**, different conditions, and different geographical scales of inventory.
2. **Ability to use existing data** in all the phases of inventory.
3. **Scalability** of being transferrable to a larger operating system where it can take full advantage of the larger operating system in terms of performance.
4. **Flexibility to accommodate** diverse data models.
5. **Adaptability to host data from repeated NFIs** with possibly changing variable lists and other adaptations.

In addition, the system should be secure, user-friendly, accessible online and offer flexible reporting tools. The documentation of systems and their procedures needs to be transparent and allow accommodating future improvements.

Skills required for data management

Forest data management and processing require personnel with robust knowledge of forest inventories, information regarding characteristics of forest (and nature) and knowledge about the tools and methods that are needed to perform all tasks related to these areas.

Basically, expertise is required in the following three fields:

Knowledge in statistics	Statistical knowledge, particularly statistical sampling and statistical modelling, is necessary to define the procedures and mathematical formulae at the planning stages of data analysis. The expert must understand the overarching objectives of the NFI, be involved in the initial planning stages and have a clear understanding of the variables used.
Database design	In order to design, build and maintain the database, the database designer must not only program and implement the system at the start but is also required to maintain and enhance the database system—this requires job-specific IT qualifications.
Database administration	Database administration requires receiving data from the field, supervising its input to the system, carrying out error checking and correction, and handling <i>ad hoc</i> queries from users. The database administrator needs to have IT and database expertise, and it is also important that the expert has participated in field work and is familiar with the procedures used, as well as the overarching inventory objectives.

Summary

Before we conclude, here are the key learning points of this lesson.

- A Database Management System (DBMS) is a collection of programs that enables its users to access databases, manipulate, report and represent data. DBMS is the mainstreaming technology for all data management.
- The Entity-Relationship Model is best used for the conceptual design of a database. Relational databases have become the dominant technology for database management.
- When planning a database, we need to design a schema that contains entities, attributes,

relationships, key fields and validation rules.

- Forest data management and processing require people with good knowledge about forest inventories, characteristics of forest information, and knowledge about tools and methods. Roles associated with necessary skills are those such as forest biometrician or statistician, database designer and maintainer, and data analyst/database administrator.

Lesson 2: Collecting and managing field data

Lesson introduction

This lesson focuses on managing field data, and describes the processes required to make this data ready for long-term storage and analyses. It also explains how to build a working database for hosting forest inventory data. Because the quality of data is always a priority, this lesson also reviews data validation processes.

Learning objectives

At the end of this lesson, you will be able to:

- Describe how field data is collected, stored and managed.
- Identify the specific requirements of ICT systems for forest inventory and monitoring.
- Explain how to build a working database for hosting forest inventory data.
- Identify methods to improve the overall quality of an NFI database.

NFI data tools: An overview

Choosing ICT tools for forest inventory data management

From a technological point of view, it is now easy to procure necessary office IT equipment, because standard office desktops and laptops are capable of managing NFI data.

At the corporate level, local servers or cloud services—hosted by a cloud service provider—are needed. In most cases, the biggest impetus for getting a server is an increase in the number of workstations for staff that extensively use the network.

Suppose you anticipate growing to more than seven workers at an office, each using a computer. In that case, you might want to consider a server to better manage your workforce and the data they produce.

Forest inventory field teams perform most of their tasks outdoors, exposing their equipment to variable weather conditions and rough handling. Because of this, **rugged computers are an advantage to work with**. However, rugged devices are usually more expensive than standard devices.

Using touchscreen technologies

There are two varieties of touchscreen technologies that are currently in demand—**resistive touch**, that uses the **pressure of the human body as an input**, and **capacitive touch**, that uses the **electrical properties of the human body as inputs**. While choosing between these two technologies, consider the environment your device will be in. **If your device will be used in more rugged and rainy conditions**, that may require wiping the screen repeatedly, **a resistive touch panel may work better**, as you can use it with both gloved hands and a stylus.

If on the other hand, **you are using the device for more sophisticated applications, go for a capacitive touch panel**. With capacitive touch panels, multi-touch is not a problem, you can scroll with ease and have excellent sensitivity.

Mobile data collection application

There are various applications for mobile data collection, including:

[Open Data Kit \(ODK\)](#)

[Survey Solutions](#)

Open Foris (OF) [Collect Mobile](#)

[ArcGIS Survey123](#)

Some applications are better suited for interviews and socio-economic surveys (such as, ODK and Survey Solutions), while others work best for land surveying (e.g. [ArcGIS AppStudio](#) and [Q Field](#))—thus, you need to evaluate applications based on your need.

Showcase: Open Foris Collect and Collect Mobile

[Open Foris \(OF\) Collect](#) is a mobile-based Android app for data collection for field-based surveys. It works as a field data management solution, and allows full customization of the data inventory structure, variables and data validation rules. The data entry User Interface (UI) of is generated automatically.

The system can be used in a standalone environment—with no need for an Internet connection— or can be installed into a server where multiple users can work simultaneously with access to the same survey data. In a larger project, as with NFIs, a server installation is needed, the server providing the storage space for hosting data. OF also allows on-the-fly validation to improve data quality and integrates with OF Collect for data management and analysis.

Data collection for NFIs

Field data can be collected and recorded using electronic devices or printed field forms. Mobile data loggers have been used in NFIs since the late 1980s, and nowadays, rugged tablets or smartphones are commonly used in collecting field data. Data input—into a computer after the field assessment—may be done manually (with a keyboard), via a cable or wirelessly from an electronic device.

If mobile communication is available in the field, data can be directly transferred into the central database or into the cloud. In any case, **it is advantageous to perform data entry as soon as possible (in terms of time and space) to where the data is generated.** Methods for checking data directly in the field allow field crews to correct errors and inconsistencies immediately, which eliminates sources of errors and improves data quality.

Mobile telephone communication increases safety for field crews because they can communicate more easily in emergency situations, and it also allows online data entry to a central database. However, coverage for mobile phone networks is poor in many regions, particularly in sparsely populated rural areas where field plots may be located, making direct mobile communication from the field technically impossible.

In such situations, it is crucial that the team leader checks the collected data and backs it up using portable storage devices, and later when the team arrives to the areas with mobile connection, the data can be uploaded or sent to the central database.



Quick tips!

Forest inventories usually require the collection of attribute data containing taxonomic and spatial (point) location information. In the application's database, it can be a beneficial feature to consider the hierarchy of data and relationships (e.g., cluster-plot-subplot-entity) instead of collecting a flat-file database consisting of just one file/table. SQLite is a popular cross-platform database in mobile devices.

Let's now see the process of data flow from the field into a clean database ready for analysis.

Field form data

While processing field form data, it is recommended to keep photocopies of the field forms at a local office, if possible. The field forms are transferred to the main office, and data is entered manually into the database.

Data in PDA/Smartphone

There are two options in which data can be transported to the main database.

Option 1: After a working day, data is exported from the tablet and copied onto a computer. A 'safety copy' of the data is stored. The data file is then sent to the main office via email, cloud storage or carried by USB memory.

Option 2: Data can also be sent to the server directly from the tablet.

Main data

The main data comprises:

Entry data: This is validated, the 'error list' is removed, and all wrong entries are checked.

Validation data: All mistakes are flexed, and the data is cleansed. It is then ready to move to the next stage.

Analysis data: Clean data is ready for analysis.

Data transfer and input

Data transfer from tablets or smartphones can be organized in many ways.

1. In cloud applications, data is automatically stored in the cloud server.
2. Suppose the offline data is collected first and the device has Internet data connection. In that case, the data can be stored into the cloud server (Google Drive, OneDrive, DropBox etc.) or transferred by email to the office.
3. In some systems, collected data can also be exported first into the tablet's storage and then copied or sent wirelessly (e.g., using Bluetooth) to a laptop and transferred later to the central database.
4. In some systems, the data can also be sent directly to a network server via a mobile phone

network.

5. The conventional use of paper field forms requires that the field officer delivers the forms to the office, where they are inputted manually into the main database.

Data management for NFIs

Specific requirements of ICT systems for forest inventory and monitoring

Here are some typical considerations for a workable information system for NFI:

1. The requirements for data, analysis, and reporting are determined by the varying information needs that exist for different situations. Forests are complex, so the DBMS can contain data about trees, other vegetation, fauna, soil characteristics, water flows, etc. Also, spatial information needs to be collected (such as coordinates of attributes).
2. The results of a data requirements analysis can usually include topics such as objects, data, relationships, processes, access paths, data integrity, information design, data sharing and data security.
3. Governmental regulations, standards, and reporting commitments need to be considered while building these systems.
4. These systems usually contain data on changes in the environment and multi-temporal datasets.

Databases in NFI

Databases are efficient tools for analyzing, storing and managing large NFI datasets, while a spreadsheet is more useful for *ad hoc* analysis than for NFI data hosting and analysis. However, building a database for a survey requires more time than using spreadsheets.

Because there are many ways to store data collected from NFIs, there is no simple conclusion on which way is optimal. Efficiency and quality requirements demand a flexible, sustainable and low-maintenance system. Typically, NFI data is stored in a centralized database that works as the main entry point for all collected data.

An NFI database is hosted by a DBMS for creating and managing several databases. A DBMS makes it possible to create, read, update, and delete data in a database. The DBMS serves as a secure interface between databases and end-users or application programs, ensuring that data is consistently organized and remains easily accessible. There are numerous database management software in the market, both

commercial and open source.

Categorical data

DBMSs can host many different data types, one of the most popular among them being categorical data.

A categorical variable is a variable that contains a fixed number of possible values divided into groups or categories. An NFI database typically contains several categorical variables (such as land use class, vegetation type or soil color) or other variables that have been converted into that form, for example as grouped data (e.g. species grouping and canopy cover class).

In data management, categorical data can be stored as a list, or more commonly as a lookup table.

Moreover, categorical data can be organized as a **flat** or a **hierarchical** table(s).

Building a schema for a database

When you start building a schema for a forest inventory database, you need to define every object (entity) and describe the variables (attributes) that should be measured. Before starting to work with the NFI schema, it is necessary to have a clear idea of the logical structure of the survey, a detailed list of the variables to be measured during field work and to decide on the optimal way of collecting data about each variable.

While building a schema, you need to examine the sampling and plot design (document) and the field manual. The corresponding documents can tell you what entities are observed, how they are defined and how they are in a relationship with each other and what properties (i.e., attributes) they contain. Usually, the database designers sit together with the designer of the sampling and plot design and go through the list of variables that the database needs to accommodate.



Quick tips!

It is recommended to make this conceptual data model design in parallel when creating the field forms in order to avoid difficulties at the data processing stage. Too often, the database design is made after creating the field forms, and the data structure that works well for the field (forms) may not work efficiently in a data management system.

It is recommended to study how databases are built in other forest inventories or use ready-made templates for DB schemas. In order to be sure that the database schema allows deriving the desired estimates or results, it is recommended to test the whole workflow with some test (mock) data before the final schema is implemented and used by the field teams.

Forest assessment data is often hierarchical. So, the hierarchical database model looks like an organizational chart or a family tree- it has a single root segment (Level 1) connected to lower level segments.

Sometimes, the hierarchy consists of multiple levels.



In a real forest inventory database, however, the schema structure is often much more complex than the simple diagrams we just looked at. This is because large amounts of diverse data require more entities. For example, if data about forest disturbance in sample plots are collected with multiple variables, a new entity is required.

In addition, the grouping of attributes can make a new entity into the database. The grouping is needed to make User Interfaces (UIs) more user-friendly when using mobile devices with smaller screens. In this way, the data collector just can see a smaller number of attributes on the screen at one time.

Data validation

Data validation means checking the accuracy and quality of entered data before using, importing, or otherwise processing them. Diverse types of validation can be performed. Data validation is a form of data cleansing. The purpose of data validation is to ensure that the recorded data adhere to the definitions and accuracy requirements fixed in the inventory protocol. The goal is to create a consistent, accurate and complete database. We can add multiple types of validation rules into a DBMS in order to help improve the quality of the data. In some systems there are two severity classes and messages: **warnings** and **errors**. The warnings are for cases in which the system user just needs to check (for

example abnormal dbh-height ratio, or a very tall tree), and errors are entries which must be corrected.

The typical validation types are as follows:

1. No null values, e.g. no missing data allowed;
2. Data type (e.g. integer, string);
3. Uniqueness, e.g. duplicates are not allowed;
4. Comparison validation, where a data attribute is compared with another data attribute or a literal by comparing operators (i.e., <, <=, ==, !=, >=, >) and attribute values; and
5. Consistent expressions, e.g. a person's email must be spelled correctly.

Validation of plant species lists

Each forest inventory needs a tree species list and more lists for bamboos, other plants, or even for other taxa such as animals (e.g. in Papua New Guinea NFI). When building a species list for an NFI or comparing or merging any list of tree (or plant) species, taxonomic names must be checked to avoid typos or artificial inflation due to synonyms. Standardization of plant names is a critical step in an NFI and in all ecological survey research.

There are **multiple global databases on the Internet that encompass millions of species**. Most of these databases can be freely accessed online or by downloading the database into the local machine.

However, such databases were often compiled from heterogeneous data sources varying in time of publication and place of origin. To date, the most used reference list of vascular plant names is [The Plant List \(TPL\)](#), hosted by the Royal Botanic Gardens, Kew. However, TPL has not been updated for a decade and originated in a time when new phylogenetic information on many genera did not exist.

In the case of long species lists, an efficient method is to use R to retrieve and process taxonomic data for validating plant names. R is the leading tool for statistics, data analysis, and machine learning. It is more than just a statistical package, it is a programming language. One advantage of using R compared to spreadsheets is that R scripts can work as a document of the data processing chain.

There are various R packages (e.g., taxonstand, taxize, RBIEN, rentrez) or online tools (e.g. [Global Name Resolver](#) or the [Taxonomic Name Resolution Service](#)) supporting researchers to check their taxonomic information. However, some of these tools rely on TPL as a reference list, and one should check whether the backstopping database is up to date.

The validation of species names can be based on direct and fuzzy matching. Fuzzy search matches words even if there are typos or misspellings, and fuzzy matching finds information based on similarities. Some R methods can present both the results and success rates for selecting the expected best single matches. With the help of R, the search algorithms can be chained so that a plant list is validated with the help of multiple repositories. It should be noted that some of these repositories require an Application Programming Interface ([API](#)) key which identifies the user. Similarly, the service provider can set limits to the API users to not exceed a certain number of requests per second (e.g., see [Terms of Use for Kew](#)).

Species coding

Species coding is strongly recommended for storing and processing taxon data in a DBMS. The reason for using codes is to create a unique identifier for each entry in a species table and simplify data processing and aggregation of data in the reporting phase. In addition, if there is a typo in the species name (in the taxon table), that error can be easily fixed anytime. So, all species must have a unique code, and genera can be coded if needed. If there are several different plant (or animal types) in the inventory, all these can have their own taxon table, as shown in the screen below. Let's now look at different methods of coding species. First, we look at species coded as text strings.

Code	Rank	Scientific name	Synonyms
ACACI/AURIC	Species	Acacia auriculiformis	
ACACI/CRASS	Species	Acacia crassicarpa	
ACACI/FARNE	Species	Acacia farnesiana	
ACACI/HOLOS	Species	Acacia holosericea	
ACACI/LEPTO	Species	Acacia leptocarpa	
ACACI/MANGI	Species	Acacia mangium	
ACACI/MEARN	Species	Acacia mearnsii	
ACACI/SIMSI	Species	Acacia simsii	
ACACI	Species	Acacia sp.	

Next, we look at species coded as numbers

Code	Rank	Scientific name	Synonyms
33296	Species	Abutilon pictum	
25213	Species	Abutilon Sinense	
50004	Species	Abutilon sp.	
25214	Species	Abutilon theophrasti	
38000	Species	Abutilon andamanica	
10001	Species	Abutilon auriculiformis	
23003	Species	Acacia catechu	
10002	Species	Acacia comosa	
23004	Species	Acacia concinna	



Quick tips!

Suppose the country does not have any species coding system in place. In that case, an applicable method is to use a system where the code is formed by combining two text strings: the first string identifies the genus, and the second identifies the species. Longer codes can be added for forms, subspecies, variants etc. as needed.

Advantages of using text codes

There are several advantages using text codes instead of numbers, including:

1. Species coding can follow the same alphabetic order as species names.
2. New species can be easily added to or removed from the list.
3. If new species are added, they can more readily be linked, integrated into the codes of related

species in the same genus. In a numbering system, this connection would be lost.

4. Efficient species coding enables easy joining operations (as with lookup tables or external equations).
5. Aggregation by genus is fast in the analysis and reporting phases (for example when computing biodiversity indexes and numbers of tree genera per plot).
6. Text codes are more readable than numeric codes.

Once the species list is validated, approved, and the data collection has started, you can only edit the names or add more synonyms. If there are missing species, these can be also added (with new codes and names) into the list during the forest inventory cycle.

For more about species coding with the help of R scripts, please see the following [Open Foris' materials](#).

Summary

Before we conclude, here are the key learning points of this lesson.

- Forestry inventory teams perform their tasks outdoors, meaning their tools get exposed to variable weather conditions and rough handling. Hence, the use of rugged devices is recommended.
- Standard office desktops and laptops are often good enough for daily office work with forestry data. Local servers or cloud services hosted by a cloud service provider are needed at the corporate level.
- Mobile data collection improves data quality because we can use methods for checking data directly in the field.
- Before starting to work with the NFI schema, it is necessary to have a clear idea of the survey's logical structure and a detailed list of the variables to be measured. Consulting the inventory protocol is paramount.
- Data validation means checking the accuracy and quality of source data before using, importing, or otherwise processing data.
- A validated species list and proper species coding will make the processing and reporting of results more efficient. It will also improve the quality of result information.

Lesson 3: Online reporting and data archiving

Lesson introduction

This lesson showcases some of the current methods, techniques and tools for NFI data reporting. It also describes the concept of metadata, and provides guidance on the importance of data archiving and storage for future use.

Learning objectives

At the end of this lesson, you will be able to:

- Describe the role of data dashboard technology in reporting forest inventory information.
- Explain the 'Online analytical processing' (OLAP) technology and how it works.
- Define the concept and basic requirements for metadata and data archiving.

Data dashboards and portals

What is a data dashboard?

Dashboards and portals have the same functionality but different uses. Portals provide a centralized repository for key information for an organization or the public, and they typically contain rich text, shortcuts, interactive images and maps. Some portals can provide real time analysis of the underlying data.

Dashboards, on the other hand, provide quick visibility in order to facilitate understanding, with easy access to the most frequently needed charts, graphs and reports. Furthermore, as a rule, portals produce static representations of results in pre-defined tables and maps, while all dashboards provide more dynamic content by using data models and real time data analysis. Here are some examples of NFI dashboards and portals for you to explore:

[Global Forest Resources Assessment](#)

[Satellite Land Monitoring Systems \(SLMS\), empowered by UN- R EDD/FAO](#)

[Bangladesh Forest Inventory Results](#)

[Germany data dashboard \(various forest inventories\)](#)

[Crance data dashboard](#)

[Canada NFI portal](#)

[Papua New Guinea Climate Change and Forest Monitoring Web portal](#)

[National Forest Monitoring System Portal \(Democratic Republic of Congo \)](#)

[National Land Monitoring System of Suriname](#)

Online analytical processing (OLAP)

What is OLAP?

Online analytical processing (OLAP) is a computer processing technology that allows the rapid execution of complex analytical queries. It allows the user to slice and filter data, and produce specific results without having to implement separate calculations. For these reasons, OLAP databases are popular for reporting.

OLAP is part of the broader category of business intelligence, encompassing relational databases, report writing and data mining. OLAP applications were first applied in business reporting, but nowadays this technique is applied in multiple sectors, including forestry and agriculture. Forest inventory reports and NFI results can be processed and shown with the help of OLAP.

However, there are **challenges associated with OLAP—it uses a terminology that is different from standard statistical and database terminology**. However, we will not further explore these differences, but instead focus on the main advantages and applications of OLAP techniques.

Next, we will look at three examples of OLAP use in NFI data processing.

Swedish NFI reporting and analysis system

A relational database makes it possible for the Swedish NFI to use the OLAP technique. This database allows fast reporting and analysis of data for standard products, used for national and international reporting. The Swedish NFI stores the most frequently used data in a Microsoft (MS) Analysis Services, which is an OLAP and data mining tool. Tools used for reporting and analyzing data are MS Excel, MS PowerPivot and different MS Reporting Services tools. For ad-hoc reporting and specific assignments, the Swedish NFI uses SAS Institute and Microsoft Power Pivot tools to extract data directly from the database and perform analysis. However, most other statistical packages can extract and handle data

from the MS SQL Server database.

[TaxWebb](#) is the Swedish NFI's interactive analysis tool that aims to provide quick and easy access to the statistics. It allows the users to build their own reports.

SWISS NFI

The National Forest Inventory and Analysis System (NAFIDAS) was developed for the Swiss NFI. NAFIDAS uses a similar design applied in the Swedish NFI—it consists of an operational data store with interfaces to source data systems (internal and external data sources), a [data and metadata storage area](#) and end-user presentation tools.

NAFIDAS produces tables and maps using the Swiss NFI data or data of regional inventories, and it consists of three major components:

1. a web application for management, documentation and reporting,
2. databases for storage, and
3. the data analysis application for analysis.

Open FORIS calc with SAIKU analytics

The demo video of Open Foris shows an example of OLAP in forest inventory results reporting in Open Foris Calc with the help of Saiku Analytics software.

Saiku Server is a web-based open source software that facilitates data visualization and data querying. Although a version of the software is freely available on the Saiku website, a special version has been customized for greater compatibility with Collect Earth. Saiku Server is included in the Collect Earth installer.

Note: Saiku story starts at the time point 3:44.

Data archiving

Every organization is responsible for preserving data. In general, well organized and stored paper form sheets are quite durable but they are hard to reuse, or even back up.

When using digital media, hard drives crash, files are misplaced, data formats may get outdated, passwords forgotten, specific details required to use data are forgotten, the media on which they are stored become obsolete (e.g. floppy disks) and even the persons responsible may move, retire, or pass away. Fortunately, technological advances and the advent of publicly available archives have made long-term data preservation easier and more reliable.

A **data archive** is a **collection of datasets with accompanying metadata stored** so that a variety of users can locate, acquire, understand and use the data. Archived data is secure against natural and human-made disasters and are conserved in a form that will continue to be accessible as technology changes.

Advantages of data archiving

There are many reasons why data archiving is valuable, with some of the most compelling being to:

1. reduce data loss,
2. use for new analyses and studies. Old data can also serve as invaluable baseline data for analyzing long-term trends,
3. use as training material for students because they are a cost-efficient means of increasing a country's scientific productivity. Archived data can help develop a cadre of highly trained and productive postdoctoral scientists,
4. assuage concerns regarding the export of intellectual property, and the failure to include local scientists in data collection efforts that often plague foreign scientists working in tropical countries,
5. verify results and correct mistakes, and
6. meet donor and government mandates and requirements.

Metadata

Metadata is **structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource**. Metadata is often called **data about data**, or information about information.

Metadata provides information that enables users to fully understand **data** (e.g. documents, images, datasets), **concepts** (e.g. classification schemes) and **real-world entities** (e.g. organizations, places, sampling protocols). A typical forest inventory example would be a detailed field protocol describing how the measurements are done and which categories (codes) are used.

In the context of an NFI, metadata can be used to answer such questions as what data were collected, how they were collected, why they were collected, how reliable they are and what issues should be accounted for when working with them. Metadata can also describe how to get data, what tools are needed to work with that data and other related matters.

The objective of documenting data is to provide enough information about the data set to allow someone to readily work with the data 20 years from now and make results reproducible. The more metadata properties are attached to data, the more valuable and smart the data becomes. Metadata is important for any information repository because it provides the ability to:

- determine availability, location, age of creation, data owner and accessibility of data;
- understand and correctly use data.
- manage data more efficiently; and
- guarantee interoperability of data.

NFIs and metadata

The NFI information system for data management needs to be well documented with metadata. All descriptions need to be compatible with the field manuals that describe the data.

For an NFI database, metadata can contain detailed documentation of the fields comprising the datasets, including the definition, the type of measurement, units where applicable and any controlled vocabularies or code lists present in the data.

A metadata standard is a common set of terms and definitions that describe data, outlining the

characteristic properties to be recorded and the values that the properties should have.

Metadata standards are especially well developed for geospatial data.



Quick tips!

When building metadata for forest information, we can cite examples implemented in forestry and other sectors. For NFI, we can look at [database documentations](#) of the US Forest Inventory and Analysis (FIA) National Program. The [Vermont Forest Inventory and Analysis Program](#) and the [National Forest Inventory on Woodlands in England in 2018](#) provide more examples of metadata.

Statistical microdata

Microdata is a concept meaning small data sets or aggregated data on specific levels (spatially or context). It is particularly applied in some international organization, as in the World Bank, Eurostat and some UN organizations, including FAO.

Microdata can include characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment. As such, the concept is applied more in socio-economic surveys rather than in (biophysical) forest inventories, but this is a new concept and it can be beneficial to understand it in the context of data management.

Microdata catalogues are datasets for storing and sharing collected datasets. The data in the dataset may stem from primary data collection (see [e example of microdata](#) from the Swedish NFI sample plots) or secondary data via aggregation or synthesis. The results are more commonly published as aggregates both for privacy reasons and because of the large quantities of data involved. Typically, all personal identifiers are removed to ensure privacy and confidentiality. Microdata catalogues can well accommodate forest information as forms of processed statistics.

Aggregated microdata can also be rich in policy analysis, research and highly disaggregated statistics (e.g. by locations, populations, age group). They allow to get a clear picture of issues by studying relationships and interactions among phenomena. Microdata are thus key to designing projects and formulating policies, targeting interventions and monitoring, and measuring the impact and results of interventions.

FAO's Food and Agriculture Microdata (FAM) Catalogue provides an inventory of datasets collected through farm and household surveys which contain information related to agriculture, forestry, food security and nutrition.

Summary

Before we conclude, here are the key learning points of this lesson, followed by a list of references from which this lesson was compiled.

- Portals typically provide key information to an organization's team or the public.
- Dashboards provide quick visibility into the status of the business, thus providing easy access to view the main metrics, charts, graphs and reports.
- Online analytical processing (OLAP) is a computer processing technology that allows the rapid execution of complex analytical queries.
- Preserving data is every organisation's responsibility. Data archives have made long-term data preservation easier and more reliable.
- Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.
- Microdata catalogues are datasets for storing and sharing collected small datasets.