Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 5. DATABASE MANAGEMENT SYSTEMS

LESSON 1. WHAT IS A DATABASE?

NOTE

Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Objectives

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

- understand **what a database is**;
- identify the **main benefits** from using a database; and
- understand the **role played by databases** in the information lifecycle.

Introduction

The World Wide Web is one of the major sources of information we have today.

We can look up millions of pages containing pieces of information about any kind of subject through an Internet connection and a web browser (software commonly available on any computer).

Such pages are often not static documents; they are **created dynamically when they are requested**, and the information they contain is often retrieved **from a database**.

Database definition

ENTRY = RECORD

Surname= .....
Name = ...
Address = ...
Telephone no. = ...

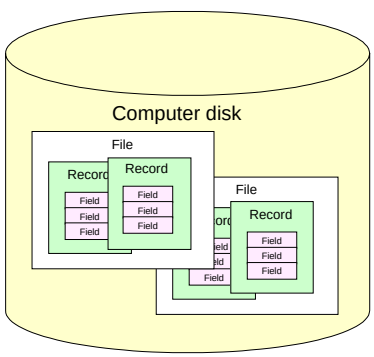
FIELDS

A database can be defined as a **persistent collection of structured data**. It provides a way to store data on computer disk so that the structure of information and the relationships between data items can be preserved.

A database is a sort of **archive** allowing us to manage huge amounts of information of a similar kind, by sorting them in a simpler and faster way compared to printed books or other types of documents (spreadsheets or text files).
The concepts of a database are similar to those of a paper-based telephone directory. The directory is indexed (alphabetically) and is composed of a series of records, representing the entries for each phone number.
The phone number itself is the unqiue identifier of the entries – no two entries have the same number. Each record contains several fields (name, address, postcode) sharing the **same structure**, which presents the information in the same way, although each record differs in the data it contains.

Physical structure of databases

The information within a database is organized in files, records and fields.



FIELD

It is the **lowest unit of storage**. Fields can be of various types, and any stored field in the database is an occurrence (or instance) of one of those field types.

RECORD

It is a **collection of field instances**. Just as with fields, each stored record is an instance of a record type. The record type defines the collection of field types that make up the record.

FILE

It is a **collection of records**. It may be physically stored on a single storage volume (computer disk) or may be spread across several storage volumes.

Features of databases

What are the main differences between a **text file and a database**?

Structure | Length | Data typing

Text is often described as being **unstructured data**.
You can store text in a file or as a field in a database, but inside the text itself there is no concept of data structures.

Actually, that's not quite true, because you can use descriptive mark-up inside text to represent information structures; hence the use of the term **structured text** to describe text with descriptive mark-up codes embedded in it.

Click on the buttons to read the explanation.

Features of databases

Structure | **Length** | Data typing

Fields in a database are of defined data types and very often they are of a **specified size** (or length). This help the database optimize the way data is stored persistently on disk.
Text, on the other hand, **is generally required to be of arbitrary** (even unlimited) length – you want to be able to store text, add to it and remove things from it without worrying about the size of the file or the field it is stored in.
For quiet a long time this caused a problem for databases that were used to store text. Fortunately, most modern databases have solved this problem and do now support data types for variable length text fields. Some even have data types for structured text such as XML.

Structure | Length | **Data typing**

A string of **text has no concept of data typing:** although the text may contain numbers or dates that are instantly recognized as data types by a reader, as far as the computer system is concerned they are just sequences of characters in the text.

DBMS properties

According to the information provided, would you define a database as a particular type of computer program?

○ Yes

○ No

Click on your answer

DBMS properties

A database is managed and organized by a type of software system known as a **Database Management System** or **DBMS**.

Essentially, a DBMS is a software "layer" between the user of the data and the actual data. The user could be a human or a computer application program.

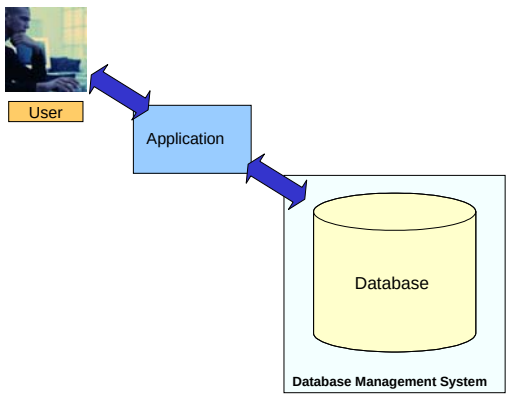
A DBMS has features to:

• **index data** so that it can be quickly searched and accessed;

• support multiple **simultaneous users**;

• provide a **query language** and **application programming interface**.

DBMS properties



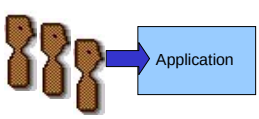
Moreover, a DBMS:

• **diminishes data redundancy**, so that a single data item can be stored once and be used in many different applications and be linked to other data items in multiple relationships;

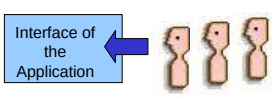
• **handles all requests to access data held in the database**, from any of the applications that use the data; those requests could be to create, delete or amend data in the database.

DBMS properties

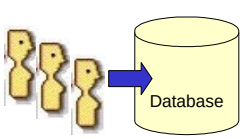
Several actors, with the following roles, are involved in using a Database System:



Programmers: people that write application programs that access data in the database.



End Users: people who use the applications written by the programmers. For them the details of the database implementation are hidden by the user interface of the application.



Administrators: people who do the housekeeping for the database – making sure the DBMS is running properly, backing up the database, monitoring the use of disk storage and memory, etc.

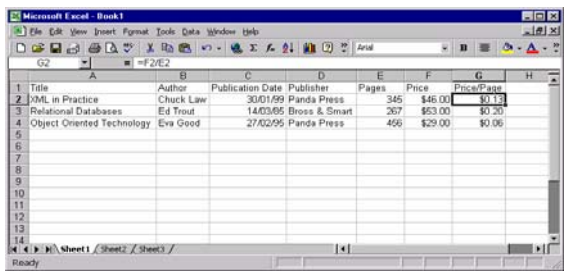
DBMS properties

On the surface **a spreadsheet** looks **a bit like a database system**.

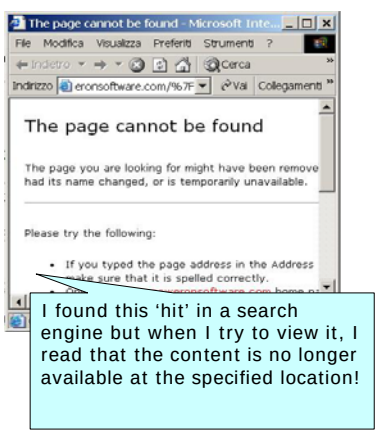
It is stored as a file and is structured as a set of sheets, each with a grid of data fields organized into **rows and columns**. You can specify the data type for each field and you can create **relationships** between fields using formulae. This similarity with databases means that it is quite easy to exchange data between spreadsheets and databases.

However, spreadsheets lack many of the basic features that you could normally associate with a database system:

- they **don't index data** for fast search and retrieval,
- they **don't have application programming interfaces or query languages**, and
- they **aren't designed for simultaneous access by multiple users**.



DBMS properties



Another important feature of DBMS is related to **changes** that can be made to information.

When a set of changes is made to the information in a database, the data are written inside a **database transaction**. That is to say, a transaction starts, the changes are written to the database and the transaction ends.
On the left, you can see a typical problem of inconsistency between the content and the indexed representation of it. Inconsistency errors do not occur **if the content is indexed inside a database**.
Consistency is one of the **ACID properties of a DBMS**.

DBMS properties

Let's look at the ACID properties:

The **ACID** properties ensure:

ATOMICITY

All changes are committed, or none. If the transaction isn't completed properly, then none of the changes are made. You are never left wondering which of your changes were made and which weren't.

CONSISTENCY

The data and the relationships between them remain consistent with the rules of the database before & after a transaction, whatever happens.

ISOLATION

Concurrent transactions are independent. If two or more transactions are trying to read or write the same data, the database implements a locking policy which ensures the transactions don't interfere with each other.

DURABILITY

Once changes have been made, the database is in persistent state, which can be recovered should anything go wrong later on.

DBMS properties

To summarize, what are the benefits of a DBMS compared to those provided by a file system?

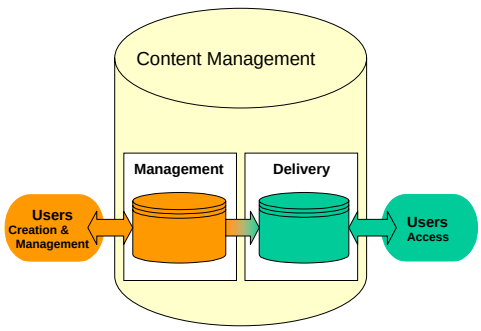
- ☐ It can contain a large amount of data.
- ☐ It allows simultaneous access by multiple users.
- ☐ It can organize information in hierarchical structure.
- ☐ It can index data for fast access and search.
- ☐ It diminishes data redundancy.

Click on your answers.

Role of databases in information lifecycle

What is the role played by a database in the information lifecycle for electronic documents?

Let's look at the diagram of a content management system, with the **creation, management** and **delivery** activities separated and with two different sets of users. One set creates and manages the electronic documents; the other accesses the document content through the delivery system.

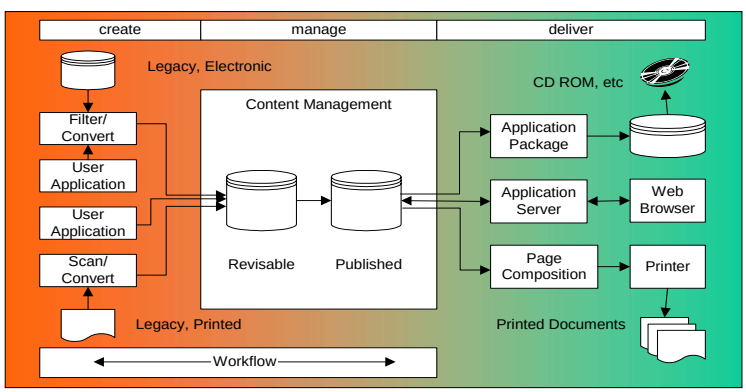


The first question you must ask yourself when thinking about how a database could help you is: **do I have a problem with information management, information delivery or both management and delivery?**

In fact, very often the users and requirements for management and delivery are **different**.

Role of databases in information lifecycle

Let's look at the systems involved in the information lifecycle in a bit more detail.



The Content Management system is shown split into the **revisable**, "work-in-progress" and the **published** content used to drive **delivery**.

Information **creation** is through three main sources – legacy **printed** material which is **scanned and converted**;

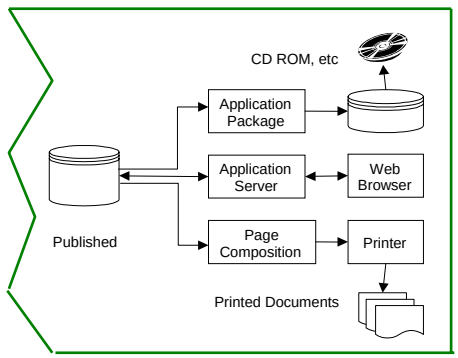
- legacy **electronic** material which is **filtered and converted**;
- **information directly input** through user applications such as word processors, desktop publishing packages, HTML/XML editors or forms interfaces.

The three **delivery** channels shown are for generating content/packages to be distributed on **CD ROM**, for delivery over the **Web** (as HTML, XML or other formats) and for generating composed pages (e.g. as PDF) which are then **printed**.

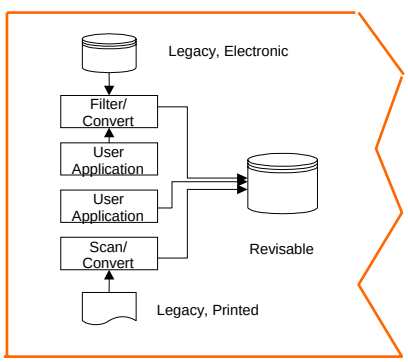
Role of databases in information lifecycle

The main benefits that databases bring for information **delivery** are:

- **fast location** of relevant information through indexed search, and
- immediate access to information at a **fine-grained level**, once it has been located through search (compare this with the file system, where the lowest level of granularity for information storage is a file).



Role of databases in information lifecycle



Databases bring more benefits when we use them for information **management**:

- fast and accurate **information access**;
- integrated content & business process management for **multiple users**, at **distributed locations**;
- management of information **reuse** and re-purposing;
- management of document **changes**;
- **version** control;
- **back up**, archive and restore.

Role of databases in information lyfecicle

In which of the following scenarios does the DBMS play a more complex role?

- ○ When thousands of users are accessing the system to modify and update the information contained in the database.
- ○ When thousands of users are accessing the system to retrieve the information contained in the database.

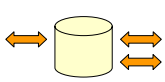
Click on your answer.

Modes of User Access

To better understand the role to be played by the database system, you need to understand **how your users are going to interact with the system**, particularly the way in which they will work using web technology. Here we present three scenarios, starting with the easiest to implement.

READ-ONLY ACCESS

READ / WRITE ACCESS **- controlled group of users viewing and updating**

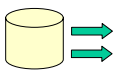


READ / WRITE ACCESS **- thousands of users viewing and updating**

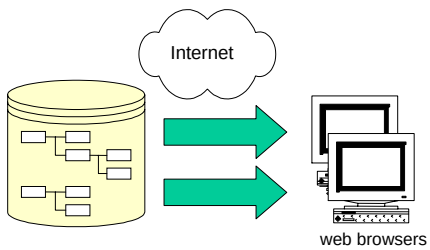
Let's analyze these scenarios in more detail…

Modes of User Access

From the view of a system implementor, the easiest access mode to support is the **database driving delivery of information through a web browser**.



In this configuration the database is for **read-only access** and is built once when the system is installed and is then only updated when the database is rebuilt.

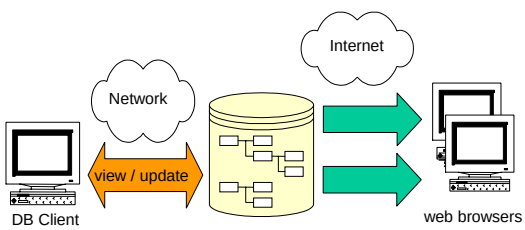


This is a common scenario for **static text bases or websites with indexed content**, which may be **accessed by many thousands of users** over the Internet.

Modes of User Access

A **more complex** mode of user access occurs when the database **is read / write**.

We still have **thousands of users browsing** information on the Internet, with a **smaller number of users viewing and updating** the database on the internal network.

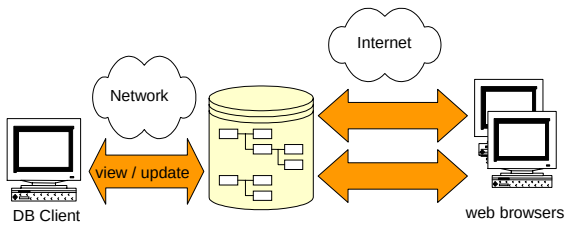


This is a common scenario for text bases or database-driven web applications where there is a **regular update** of the database **by a controlled group of users.** The importance of the database here is that it manages the **locking and integrity** of information which may be read and written by different users simultaneously.

Modes of User Access

The **most complex** mode of user access occurs when the database **is read / write** for **thousands of users**.

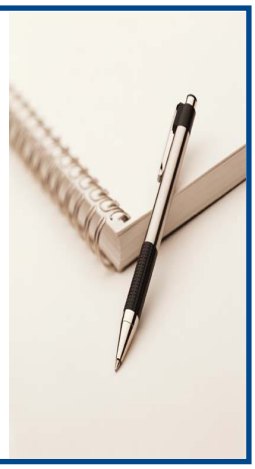
In this case, the **read / write** access to the database is granted not just to the small group of users on the internal network, but also **to the thousands of users** through the web interface.



Here the database plays a crucial role in managing the **potential conflicts** and locks. That occurs when many users simultaneously read and write information.

Summary

- A database can be defined as a **persistent collection of structured data**.
- The information within a database is organized in files, records and fields.
- A database allows users to manage and modify the information it contains without altering the structure.
- A database is accessed through a software layer called a **Database Management System** or **DBMS**, that supports multiple **simultaneous users**, provides a **query language** and an **application programming interface**.
- Databases offer several benefits for the management and delivery phases of the information lifecycle.
- Implementing a data management and delivery system can be quite complex, depending on how users are able to **view and update** the database.



Exercises

The next four exercises will allow you to test your understanding of the concepts described up until now.

Good luck!

Exercise 1

Which of the following definitions can be associated with the concept of database?

- ○ A software application to manage a persistent collection of information.
- ○ A collection of information that is stored in order to preserve its structure.
- ○ A collection of information that is stored in a structured way.

Click on your answer.

Exercise 2

When are the ACID properties of a database involved?

- ○ When a set of changes is made to the information in a database.
- ○ When a user needs a query language to communicate with the database.
- ○ When a large amount of information must be indexed in a database.

Click on your answer.

Exercise 3

Is it possible to use the database in the information management stage only, and not in the information delivery stage?

- ○ Yes.
- ○ No.

Click on your answer.

Exercise 4

In which of the following scenarios would a more complex DBMS configuration be required?

- ○ 3000 users with read-only access to information.
- ○ 2000 users with read-only access to information and 10 users with read/update access to information.
- ○ 2000 users with read/update access to information.

Click on your answer.

If you want to know more...

Date, C.J. An Introduction to Database Systems Addison Wesley; ISBN: 0201787229. The definitive book on database systems.

Date, C.J. & Darwen H. Foundation for Future Database Systems: The Third Manifesto Addison Wesley; ISBN: 0201709287.

Patricia Seybold Group. Industry analysts, including content and information management systems. (www.psgroup.com)

The Gilbane Report (www.gilbane.com) newsletter covers content management, XML, e-catalogs, intranet publishing, content computing architectures, markup languages, information integration, corporate portals, and enterprise search.

Content Management Advisor. An online magazine offering expert advice on managing and publishing digital content. (contentmanagementadvisor.com).