

3: Information Storage and Retrieval – Prt I: Database Systems

Thomas Haigh
L&IS 110 – February 7, 2005

Information Storage and Retrieval

- Will be considering three main kinds of system
 - Database Management System (DBMS)
 - Online Information Retrieval System
 - Search Engines
- Used for different purposes
- Store information in different ways

Brief Demo of Digital Library

- Will return to next week

Brief Demo of Search Engine

- Will return to next week

Data Base Management System

- Single most important kind of corporate IT
 - Foundation of almost every
 - Advanced web site
 - Administrative application
 - Increasing use in science
 - Experimental results, clinical trials, etc.
 - Data acquired from equipment
 - Gene sequences, etc.

Background: Org. Info Systems

- Big organizations have hundreds of different systems to do different things
 - Some written specially
 - Some adapted from standard packages
 - Some totally standard
- Each system needs to access data
 - Most systems also update some data
 - Systems are mostly to DO things, not just answer questions.

File Based Systems

- Computer data is stored in "files"
 - Very old idea – predates computers
 - Can be on tape, hard disk, CD, etc.
- Files usually in unique format specific to application creating
 - E.g. .doc files for Microsoft Word
- Problem: other applications need to access the same data
 - Much more of a problem for specialized, custom applications, because business processes are integrated

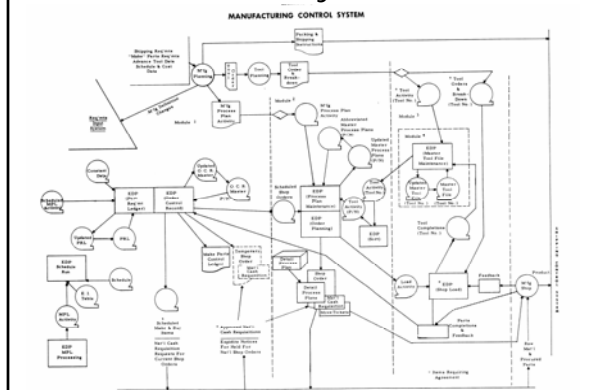
File-Based Processing



Sales Files
PropertyForRent (propertyNo, street, city, postcode, type, rooms, rent, ownerNo)
PrivateOwner (ownerNo, fName, lName, address, telNo)
Client (clientNo, fName, lName, address, telNo, prefType, maxRent)

Contracts Files
Lease (leaseNo, propertyNo, clientNo, rent, paymentMethod, deposit, paid, rentStart, rentFinish, duration)
PropertyForRent (propertyNo, street, city, postcode, rent)
Client (clientNo, fName, lName, address, telNo)

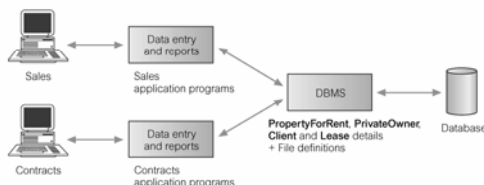
A File Based System, 1962



Data Base Management System

- Standard piece of system software
 - Oracle, SQL Server, DB2, Access are most widely used
- Supports multiple data bases
 - Creation, modification of data structures (i.e. tables)
 - Via Data Definition Language (DDL)
 - Retrieval, insertion, deletion, updating of data (i.e. rows)
 - Via Data Manipulation Language (DML)

Database Management System (DBMS)

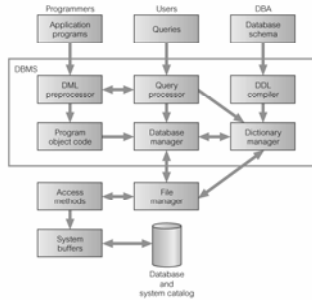


PropertyForRent (propertyNo, street, city, postcode, type, rooms, rent, ownerNo)
PrivateOwner (ownerNo, fName, lName, address, telNo)
Client (clientNo, fName, lName, address, telNo, prefType, maxRent)
Lease (leaseNo, propertyNo, clientNo, paymentMethod, deposit, paid, rentStart, rentFinish)

Advantages of DBMS

- Sits between users/applications and data
- Once database replaces many files
 - Can be used by multiple applications
 - Gives "program/data independence"
 - Changing a program doesn't mean changing database
 - Can alter database format without changing all programs that use the data
- Adds many advantages

Components of a DBMS



Gives Different Views on Data

- Different users have different permissions
 - View, change, delete
 - On various parts of the data base
- “Views” are used to present
 - Data joined, grouped or filtered in particular ways
 - Can include results of calculations or functions
- This allows
 - Avoidance of duplicated data
 - Store it once; present in many different ways
 - Called “normalization”

Current DBMS Systems

- Are “relational”
 - Follow relational model
- Use SQL (Structured Query Language)
 - to define and manipulate data
- Support multiple simultaneous users
 - Can be ad-hoc individuals
 - Can be batch jobs or programs
- Have many advanced capabilities
 - E.g. code “triggered” when data changes
 - “Transactions” to protect against data loss
 - Backup and mirroring of data to protect against loss

Not One Big Database

- Big central database doesn’t work in practice
 - Though still better than file systems
- Finish up with dozens/hundreds of little data bases
 - Physically separate
 - All incomplete
 - Different data formats
 - Different concepts of data
- Dominant model is “relational” (eg Oracle)
 - Good for updating
 - Flexible
 - Can be slow & complex to extract data for reports
 - Special “data warehouses” often built for this purpose

DBMS as Information Technology

- Compared to original (1960s) hope of universal information retrieval technology
 - New concept of database is narrower
 - More general information retrieval problems are excluded
- DBMS is not well suited for
 - Irregular records
 - Full text or even keyword searching
 - Ad-hoc linkages between records
 - Context, relevance (in IS terms)
- Only with search engines of 90s
 - Is much attention given to unstructured data

DBMS versus Search Engine

- DBMS data is very highly structured
 - Each table has defined “fields”
 - Each field has specific format, length, etc.
 - Each record in the table follows the same format
 - Each record has a “key” to uniquely locate
 - E.g. supplier ID number, SSN, campus ID number
- Advantages
 - Can reliably access specific information on a record
 - Makes it possible to update
 - E.g. find specific employee record and increase salary \$500
 - Can produce totals, statistics, etc.
 - Can automatically join together information held in multiple tables

DBMS versus Search Engine

- But, structure means that DBMS is
 - Rigid in terms of information stored
 - All storage or searching in specific field
 - Can't easily add extra information on just one record to reflect special features
 - DBMS data structure still closely related to needs of specific applications using