**COMP 1630 – Assignment 1**

Group 1 – Renin Catungal, Xiaochen Dai, Wing Chu, Gael Attal

1. **Define each of the following terms:**
   1. **Data - raw facts, (ie. Phone numbers, date of birth, names,) have little meaning unless organized in a logical manner**
   2. **Field - character or group of characters that has a specific meaning, a field is used to define and store data**
   3. **Record - logically connected set of one or more fields that describes a person, place or thing (ie. Customer record consists of name, address, phone number, date of birth fields**
   4. **File - collection of related records, (files may contain data about employees currently employed at big corporations)**
2. **What is data redundancy, and which characteristics of the file system can lead to it?**

* **Data redundancy occurs when the same data are stored unnecessarily at different locations. File system structure make it difficult to combine data from multiple sources, (ie. Different departments have access to specific locations, employees within that department access to the same files and can have different versions of the same data.)**

1. **What is data independence, and why is it lacking in file systems**

* **Data independence exists when you can change the data storage characteristics without affecting the program’s ability to access the data**
* **In file systems, all data access programs are subject to change when any of the file’s data storage characteristics change**

1. **What is a DBMS, and what are its functions?**

* **DBMS – Database Management System**
* **Functions include: Data dictionary management, data storage management, data transformation and presentation, security management, multiuser access control, backup and recovery management, data integrity management, database access languages and application programming interfaces, database communication interfaces**

1. **What is structural independence, and why is it important?**

* **Structural independence means the file structure can be changed without affecting the application’s ability to access the data. From the point of view of a programmer and database manager, data independence makes the file system extremely cumbersome, and data independence might provide solutions to overcome these limitations.**

1. **Explain the differences among data, information and database**

* **Data are raw facts which have not yet been processed to reveal their meaning. It constitute the building blocks of information;**
* **Information is the result of processing raw data to reveal its meaning. It is produced by processing data;**
* **Database is a shared, integrated computer structure that stores a collection of end-user data and metadata.**

1. **What is the role of a DBMS, and what are its advantages? What are its disadvantages?**

* **The DBMS serves as the intermediary between the user and the database.**

**Advantages:**

* **Improved data sharing**
* **Improved data security**
* **Be data integration**
* **Minimized data inconsistency**
* **Improved data access**
* **Improved decision making**
* **Increased end-user productivity**

**Disadvantages:**

* **cost of hardware and software**
* **cost of data conversion**
* **cost of staff training**
* **appointing technical staff**
* **database damage**

1. **List and describe the different types of databases.**

**- Databases can be categorized according to a series of factors.**

**The number of users**

**(a) single-user database: supports only one user at a time**

**(b) multi-user database: supports multiple users at the same time**

**Location**

**(a) centralized database: supports data located at a single site**

**(b) distributed database: supports data distributed across several different sites**

**Type of Data stored**

**(a) general-purpose database: contains a wide variety of data used in multiple disciplines**

**(b) discipline-specific database: contains data focused on specific subject areas**

**The intended data usage**

**a) operational database: supports a company’s day-to-day operations**

**b) analytical database: focuses on storing historical data and business metrics used exclusively for tactical or strategic decision making**

**The degree to which the data are structured**

**a) unstructured data: data that exist in raw state**

**b) structured data: result of formatting unstructured data to facilitate storage, use, and the generation of information**

**b) semistructured data: data that have already been processed to some extent**

**9. What are the main components of a database system?**

**- A database system is composed of five main parts: hardware, software, people, procedures, and data**

**10. What are metadata?**

**- Metadata (data about data) describes the data characteristics and the set of relationships that links the data found within the database.**

**- It stores information such as the name of each data element, the type of values (numeric, dates, or text) stored on each data element, and whether the data element can be left empty.**

**11. Explain why database design is important.**

**- Database design is crucial for efficient and proper data management and to generate accurate and valuable information to make key business decision.**

**- A poorly designed database can lead to erroneous results when data is retrieved, which may lead to the failure of an organization.**

**12. What are the potential costs of implementing a database system?**

**- Requires sophisticated hardware, software and highly skilled personnel**

**- Training, licensing, and regulation expenses**

**- Frequent upgrade /replacement cycles**

**- Vendor dependency since companies might be reluctant to change vendor once a database system is implemented**

**- Adoption of database system must be properly managed. Security issues must be assessed constantly**

**13.  Use examples to compare and contrast unstructured and structured data. Which type is more prevalent in a typical business environment?**

**Suppose a free form was submitted to users which asked them to input data such as their date of birth. Let us use the following examples of sample data that we would retrieve from what should be structurally inconsistent input:**

**January 12th, 1969**

**04-02-1992**

**Mar-30-1977**

**The 11th of April, 1945**

**November the fifth, nineteen-ninety-nine.**

**What we observe is an example of raw, unstructured data. In other words, it's very difficult to derive specific data from such varying methods of input. If we were to rewrite this data in a structured (and consistent) format, it might look something like this:**

**Jan-12-1969**

**Feb-04-1992**

**Mar-30-1977**

**Apr-11-1945**

**Nov-05-1999**

**Notice that the data is much clearer: The format is consistent in that the month comes first, and then the day, followed by the year. The month is always a three letter abbreviation, the data is separated by dashes, etc. This is called structured data, and is essential (and usually more prevalent) in the running of a successful business as it maintains consistency throughout the formatting of data.**

**14.  What are some basic database functions that a spreadsheet cannot perform?**

**A database is a powerful tool which can be used to define relationships between different tables. It can also impose referential integrity on the data contained within all the tables it relates and as a result ensure that all the data across the board is consistent (which is also done through the use of constraints and enforcement of data types). A database also supports self-documentation using metadata. A spreadsheet is unable to perform any of these basic functions.**

**15.  What common problem(s) does a collection of spreadsheets created by end users share with the typical file system?**

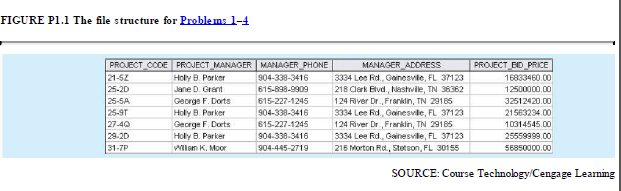
**The most glaring problem a collection of spreadsheets created by end users shares with the typical file system is that of data inconsistency and redundancy. A record that is modified in one location must imperatively be modified in every other location in which it exists and every file that communicates or interacts with it in some way must be aware of this change including parent and child data that might result from it.**

**The second problem that can be encountered is simply the cumbersome nature of retrieving information and completing queries. It's slightly easier using spreadsheets due to their digital nature, however if one were to try and derive a narrow query that demands very specific records from multiple spreadsheets, then retrieving that information can still prove to be a challenging and inefficient endeavour.**

**16.  Explain the significance of the loss of direct, hands-on access to business data that end users experienced with the advent of computerized data repositories.**

**The significance of the loss of direct, hands-on access to business data that end users experienced can be described in terms of the new and improved tools that these users received that aided them in manipulating company data and creating new information. It also resulted in the forming of a rift between the data and the users. The desire to close that rift was in part one of the driving forces behind the development of new computer technologies and system designs, among other things**

**Problems**



**Problem 1: How many records does the file contain? How many fields are there per record?**

* The file contains 7 records with 5 fields

**Problem 2: What problems would you encounter if you wanted to produce a listing by city? How would you solve this problem by altering the file structure?**

* If I produced a listing by city I would get multiples of Holly B. Parker, and George F. Dorts. Update fields (null fields) where Holly and George occur so that their names, address, and phone number only occur once, retain project\_code and project\_bid\_price.
* OR alter the file structure by decomposing original table into two tables (Table 1 – Manager\_Address, manager\_phone and project\_manager, Project\_code Table 2 – Project\_code, project\_bid\_price)

**Problem 3: If you wanted to produce a listing of the file contents by last name, area code, city, state, or zip code, how would you alter the file structure?**

Decompose the data into two related tables as well as separate the name and address fields

Fields of Table One:

MANAGER\_LAST\_NAME

MANAGER\_FIRST\_NAME

PHONE\_AREA\_CODE

PHONE\_LOCAL\_NUMBER

ADDRESS\_STREET\_NUMBER

ADDRESS\_STREET\_NAME

ADDRESS\_CITY

ADDRESS\_STATE

ADDRESS\_ZIP

PROJECT\_CODE

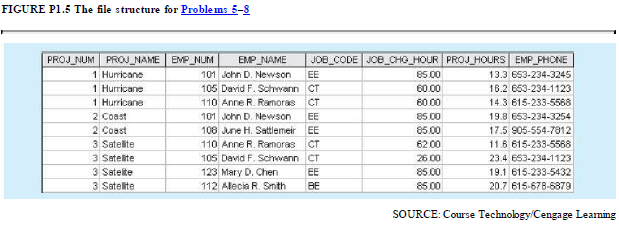
Fields of Table Two:

PROJECT\_CODE

PROJECT\_BID\_PRICE

**Problem 4: What data redundancies do you detect? How could those redundancies lead to anomalies?**

The fields Project\_Manager, Manager\_Phone, Manager\_Address have multiple occurrences with same values. These data redundancies might lead to anomalies. Ideally, a field value change should be made in only a single place. For example, if we want to change the phone number of the project manager named “Holly B. Parker”, we should update the MANAGER\_PHONE field of all records whose PROJECT\_MANAGER field is “Holly B. Parker”. If unfortunately one place is missing, we might encounter a data inconsistency problem.



**5. Identify and discuss the serious data redundancy problems exhibited by the file structure shown in Figure P1.5.**

- There are unnecessary duplication of information such as project name, employee name, job charge per hour, employee phone number causing data redundancy and anomalies

- Employee John D. Newson has 2 different employee numbers: 653-234-3245 and 653-234-3254 causing data inconsistency

- Job code CT for project satellite has 2 different charge hours: 62.00 and 26.00 causing data inconsistency

- Project Name Satellite is spelled wrong (satelite) in one of the records causing data inconsistency

**6. Looking at the EMP\_NAME and EMP\_PHONE contents in Figure P1.5 what change(s) would you recommend?**

- Have EMP\_NAME, and EMP\_PHONE in another table along with EMP\_NUM being the key field. This will eliminate redundancy and inconsistency in the employee information fields.

- For the EMP\_NAME contents: Separate the first name, initials, last name into different fields

- For the EMP\_PHONE contents: separate the area code from the main phone number into different fields

**7.  Identify the various data sources in the file you examined in Problem 5.**

The data sources in Problem 5 can be identified as follows:

**Project** ( Project Number, Project Name, Hours Allocated to Project )

**Employee** ( Employee Number, Employee Name, Employee Phone, Hours Allocated to Project )

**Job** ( Job Code, Job Charge per Hour )

This is to say that the source of all the data from this table can be derived from three different entities: Project, Employee, and Job

**8.  Given your answer to Problem 7, what new files should you create to help eliminate the data redundancies found in the file shown in Figure P1.5?**

It's apparent that the creation of an Employee, Project, and Job table will help eliminate the redundancies that can be found in the file shown in Figure P1.5. However, it seems that the Job Charge per Hour field isn't exclusively dependant on the Job Code. Depending on whether or not it's also dependant on PROJ\_HOURS, PROJ\_NUM, or EMP\_NUM (or any combination of these), the right propitiations must be made to ensure that the data is consistent and that redundancy is kept to a minimum.