Section 1. Introduction to Databases and Spreadsheets#

Although, strictly speaking, databases and spreadsheets have very different functions, it can be argued that in many archaeological applications both are used to collect and store data in a similar way (defined in terms of records/rows and fields/columns). From an archival perspective this similarity becomes more apparent when the significant properties are taken into account. When looking to preserve data in these formats, the key significant properties of both databases and spreadsheets are both the data values themselves and the structure (tables or sheets) in which this data is held. From this perspective both types of object can be treated (and archived) in a similar way.

This guide aims to provide an overview and guide to preserving the most common features of databases and spreadsheets created as a part of archaeological research. The guide will highlight their similarities - and how they can be dealt with in a similar fashion - together with areas in which they differ and in which additional elements, characteristics and processes must be documented. This guide does not, however, aim to provide detailed guidance on the design of databases or spreadsheets beyond considerations that influence their preservation.

1.1 What are Databases and Spreadsheets#

Although, at the simplest level databases and spreadsheets are similar in that they contain tabular data with values organised in columns and rows, a distinction between the two applications can be made based on their intended functions. Spreadsheet applications, based originally on paper accounting worksheets, specifically aim to deal with mathematical data (e.g. accounts) and to perform on-the-fly calculations and processing. Database applications, however, are designed to store data of a wide variety of types and to provide complex searching and reporting functions upon this data.

Spreadsheets

Spreadsheets are arguably a simpler format than databases and generally consist of single or multiple 'sheets' containing tabular data. The data itself can be used within a spreadsheet to create additional values (e.g. column totals) via formulas and can also be used to generate any number of graphs and charts which in turn can be placed within a sheet or exist as a sheet themselves. Formatting, of either cells or of the values within them, can also be an important element of a spreadsheet and can be often used to convey meaning or highlight specific elements. Data entry and use of a spreadsheet can also be controlled to limited degree through the use of protected or locked cells and cell-specific formatting (e.g. rounding values to limited decimal places or displaying them in currency formats).

Databases

In contrast to spreadsheets, which largely share a similar fundamental design and approach, databases can be divided into a number of types (known as models or architectures). The two of these most commonly found in use in archaeology are *Flat File* and *Relational* databases although there is a slowly growing movement towards the use of *Object-oriented* database models. The flat file model is broadly similar to that of a spreadsheet in that tabular data is organised into horizontal rows, representing records, and vertical columns or fields representing a type of value or attribute to be recorded. In flat file databases there can be an inherent looseness in the way that data is defined and recorded along with a significant duplication of sets of information from record to record. The relational model addresses these and other issues by requiring a data structure to be pre-defined and by splitting related groups of attributes into separate tables which are then linked together through

key fields (*Primary* or *Foreign* keys). In contrast to spreadsheets and many flat file databases, most database applications allow (and in fact require) the strict specification - in terms of field length, data type (numeric, etc.) of the data types to be recorded

As with charts generated from spreadsheet data, databases can potentially consist of more than just data values. Forms, used for data entry or for running queries, are often the only way in which many users interact with databases and can be viewed as part of the database but separate from the data itself. Likewise, the queries and results or reports that result from user interaction may also be considered as 'non-data' components of a database.

1.2 Current Issues and Concerns#

Embedded Objects

As with word processing software, many database and spreadsheet applications allow users to embed other media (especially images) within files. Whereas databases are more likely to store links to external files rather than the files themselves, spreadsheet applications such as Microsoft Excel and OpenOffice Calc allow users to embed graphs and charts generated from data along with other images. Again, as with text files, it is advisable that such content is stored and archived separately thereby retaining the original qualities of the content (e.g. image resolution) and allowing it to follow a separate archival strategy to the textual content.

Data Consistency and Documentation

Coded or inconsistently entered data presents a problem for both databases and spreadsheets when it comes to data reuse. Coded fields and data should be adequately documented and such documentation should be archived alongside the database or spreadsheet so that the meaning of fields or data is not lost over time. Inconsistent data entry (which can be controlled to a greater extent within databases than spreadsheets) can also result in the meaning of data being lost (e.g. is 'A' actually equal to 'a'?) and provide problems with querying the dataset.

Non-data Content

As discussed above (and in more detail later on) both databases and spreadsheets can consist of more than just tabular data, charts and images. Spreadsheet applications in particular allow a great deal of formatting (i.e. font colour and style, cell colour, border styles) to be applied to data and the cells that contain it. In general, such styles are usually employed to highlight certain aspects of the data (such as column totals, negative figures and so on) but often these styles can be used to convey meaning which can then be lost during data migrations (especially to plain text).

Next | Contents