software carpentry (http://software-carpentry.org)

Selecting Data

In the late 1920s and early 1930s, William Dyer, Frank Pabodie, and Valentina Roerich led expeditions to the Pole of Inaccessibility (http://en.wikipedia.org/wiki/Pole_of_inaccessibility) in the South Pacific, and then onward to Antarctica. Two years ago, their expeditions were found in a storage locker at Miskatonic University. We have scanned and OCR'd the data they contain, and we now want to store that information in a way that will make search and analysis easy.

We basically have three options: text files, a spreadsheet, or a database. Text files are easiest to create, and work well with version control, but then we would then have to build search and analysis tools ourselves. Spreadsheets are good for doing simple analysis, they don't handle large or complex data sets very well. We would therefore like to put this data in a database, and these lessons will show how to do that.

Objectives

- Explain the difference between a table, a record, and a field.
- Explain the difference between a database and a database manager.
- Write a query to select all values for specific fields from a single table.

A Few Definitions

A relational database (../../gloss.html#relational-database) is a way to store and manipulate information that is arranged as tables (../../gloss.html#table-database). Each table has columns (also known as fields (../../gloss.html#field-database)) which describe the data, and rows (also known as records (../../gloss.html#record-database)) which contain the data.

When we are using a spreadsheet, we put formulas into cells to calculate new values based on old ones. When we are using a database, we send commands (usually called queries (../../gloss.html#query)) to a database manager (../../gloss.html#database-manager): a program that manipulates the database for us. The database manager does whatever lookups and calculations the query specifies, returning the results in a tabular form that we can then use as a starting point for further queries.

> Every database manager—Oracle, IBM DB2, PostgreSQL, MySQL, Microsoft Access, and SQLite—stores data in a different way, so a database created with one cannot be used directly by another. However, every database manager can import and export data in a variety of formats, so it *is* possible to move information from one to another.

Queries are written in a language called SQL (../../gloss.html#sql), which stands for "Structured Query Language". SQL provides hundreds of different ways to analyze and recombine data; we will only look at a handful, but that handful accounts for most of what scientists do.

The tables below show the database we will use in our examples:

Person: people who took readings.

ident	personal	family
dyer	William	Dyer
pb	Frank	Pabodie
lake	Anderson	Lake
roe	Valentina	Roerich
danforth	Frank	Danforth

Site: locations where readings were taken.

name	lat	long
DR-1	-49.85	-128.57
DR-3	-47.15	-126.72
MSK-4	-48.87	-123.4

Visited: when readings were taken at specific sites.

ident	site	dated
619	DR-1	1927-02-08
622	DR-1	1927-02-10
734	DR-3	1939-01-07
735	DR-3	1930-01-12
751	DR-3	1930-02-26
752	DR-3	
837	MSK-4	1932-01-14
844	DR-1	1932-03-22

Survey: the actual readings.

taken	person	quant	reading
619	dyer	rad	9.82
619	dyer	sal	0.13
622	dyer	rad	7.8
622	dyer	sal	0.09
734	pb	rad	8.41
734	lake	sal	0.05
734	pb	temp	-21.5
735	pb	rad	7.22
735		sal	0.06
735		temp	-26.0
751	pb	rad	4.35
751	pb	temp	-18.5
751	lake	sal	0.1
752	lake	rad	2.19
752	lake	sal	0.09
752	lake	temp	-16.0
752	roe	sal	41.6
837	lake	rad	1.46
837	lake	sal	0.21
837	roe	sal	22.5
844	roe	rad	11.25

Notice that three entries—one in the `Visited` table, and two in the `Survey` table—are shown in red because they don't contain any actual data: we'll return to these missing values later. For now, let's write an SQL query that displays scientists' names. We do this using the SQL command `select`, giving it the names of the columns we want and the table we want them from. Our query and its output look like this:

```
%load_ext sqlitemagic
```

```
%%sqlite survey.db
select family, personal from Person;
```

```
Dyer     William
Pabodie  Frank
Lake     Anderson
Roerich  Valentina
Danforth Frank
```

The semi-colon at the end of the query tells the database manager that the query is complete and ready to run. We have written our commands and column names in lower case, and the table name in Title Case, but we don't have to: as the example below shows, SQL is case insensitive (../../gloss.html#case-insensitive).

```
%%sqlite survey.db
SeLeCt FaMiLy, PeRsOnAl FrOm PeRsOn;
```

```
Dyer    William
Pabodie Frank
Lake    Anderson
Roerich Valentina
DanforthFrank
```

Whatever casing convention you choose, please be consistent: complex queries are hard enough to read without the extra cognitive load of random capitalization.

Going back to our query, it's important to understand that the rows and columns in a database table aren't actually stored in any particular order. They will always be *displayed* in some order, but we can control that in various ways. For example, we could swap the columns in the output by writing our query as:

```
%%sqlite survey.db
select personal, family from Person;
```

```
William  Dyer
Frank    Pabodie
Anderson Lake
ValentinaRoerich
Frank    Danforth
```

or even repeat columns:

```
%%sqlite survey.db
select ident, ident, ident from Person;
```

```
dyer    dyer    dyer
pb      pb      pb
lake    lake    lake
roe     roe     roe
danforthdanforthdanforth
```

As a shortcut, we can select all of the columns in a table using `*`:

```
%%sqlite survey.db
select * from Person;
```

```
dyer    William  Dyer
pb      Frank    Pabodie
lake    Anderson Lake
roe     ValentinaRoerich
danforthFrank    Danforth
```

Challenges

1. Write a query that selects only site names from the `Site` table.

2. Many people format queries as:

   ```
   SELECT personal, family FROM person;
   ```

   or as:

   ```
   select Personal, Family from PERSON;
   ```

   What style do you find easiest to read, and why?

Key Points

- A relational database stores information in tables, each of which has a fixed set of columns and a variable number of records.
- A database manager is a program that manipulates information stored in a database.
- We write queries in a specialized language called SQL to extract information from databases.
- SQL is case-insensitive.

Email (admin@software-carpentry.org) **Twitter (https://twitter.com/swcarpentry)**
RSS (http://software-carpentry.org/feed.xml) **GitHub (https://github.com/swcarpentry)** **IRC (irc://moznet/sciencelab)**
License (../../LICENSE.html)
Bug Report (mailto:admin@software-carpentry.org?subject=bug%20in%20novice/sql/01-select.md)