# Database Design

**Author: Joshua Thompson**
**Created: January 14, 2009**
**The LINGUIST List**

# Overview

- There is no one right design for all cases
- Design is largely dependent on usage patterns
- Trade-offs: speed vs size
- Think ahead: extensible designs

# Naming Conventions

- The exact naming convention isn't as important as being consistent
- Case isn't important
- Use underscores for spaces if needed
- Be descriptive, but not overly descriptive. People have to type your column names!

# Choosing a Data Type

- Choosing the appropriate data type lets the DBMS help you enforce data integrity
- Watch out for NULL
- Set default values

# Normalization

- Remove redundant data
- Avoid potential data inconsistency
- Don't prefer one class of query
- Normalization levels: 1st normal form, 2nd normal form, 3rd normal form
- Anything in a given normal form is implicitly in the lower forms as well
- Shoot for third normal form (3NF) whenever possible

# First Normal Form

- No specific ordering of rows or columns required
- No duplicate rows
- Only one value per column in a given row
- No repeated columns (that is, no two columns holding data in the same domain)

# First Normal Form

- Not in 1NF, because the telephone number field contains multiple values:

| Customer ID | First Name | Surname | Telephone Number |
|---|---|---|---|
| 123 | Robert | Ingram | 555-861-2025 |
| 456 | Jane | Wright | 555-403-1659, 555-776-4100 |
| 789 | Maria | Fernandez | 555-808-9633 |

- Not strictly in 1NF, because the multiple telephone number columns are in the same data domain

| Customer ID | First Name | Surname | Tel. No. 1 | Tel. No. 2 |
|---|---|---|---|---|
| 123 | Robert | Ingram | 555-861-2025 | |
| 456 | Jane | Wright | 555-403-1659 | 555-776-4100 |
| 789 | Maria | Fernandez | 555-808-9633 | |

# First Normal Form

- The same data in 1NF:

| Customer ID | First Name | Surname |
|---|---|---|
| 123 | Robert | Ingram |
| 456 | Jane | Wright |
| 789 | Maria | Fernandez |

| Customer ID | Telephone Number |
|---|---|
| 123 | 555-861-2025 |
| 456 | 555-403-1659 |
| 456 | 555-776-4100 |
| 789 | 555-808-9633 |

# Second Normal Form

- Everything from 1NF
- For every candidate key, the other columns depend only on the whole key and not just a part of it

# Second Normal Form

- The following table is not in 2NF; given the candidate key <Employee, Skill>, the value of Current Work Location is dependent only on the value of Employee.

| Employee | Skill | Current Work Location |
|---|---|---|
| Jones | Typing | 114 Main Street |
| Jones | Shorthand | 114 Main Street |
| Jones | Whittling | 114 Main Street |
| Roberts | Light Cleaning | 73 Industrial Way |
| Ellis | Alchemy | 73 Industrial Way |
| Ellis | Juggling | 73 Industrial Way |
| Harrison | Light Cleaning | 73 Industrial Way |

# Second Normal Form

- The same data in 2NF:

| Employee | Skill |
|---|---|
| Jones | Typing |
| Jones | Shorthand |
| Jones | Whittling |
| Roberts | Light Cleaning |
| Ellis | Alchemy |
| Ellis | Juggling |
| Harrison | Light Cleaning |

| Employee | Current Work Location |
|---|---|
| Jones | 114 Main Street |
| Roberts | 73 Industrial Way |
| Ellis | 73 Industrial Way |
| Harrison | 73 Industrial Way |

# Third Normal Form

- Everything from 2NF
- All columns that are not part of any candidate key are directly dependent on the keys, and not dependent via an intermediate column (*transitive dependence*)
- The simplest way to remember this is that all non-key columns express a fact about “The key, the whole key, and nothing but the key.”

# Third Normal Form

- The following is not in 3NF, because the Winner Date of Birth is functionally dependent on <Tournament, Year> indirectly via Winner, which is not part of the key:

| Tournament | Year | Winner | Winner Date of Birth |
|---|---|---|---|
| Indiana Invitational | 1998 | Al Fredrickson | 21 July 1975 |
| Cleveland Open | 1999 | Bob Albertson | 28 September 1968 |
| Des Moines Masters | 1999 | Al Fredrickson | 21 July 1975 |
| Indiana Invitational | 1999 | Chip Masterson | 14 March 1977 |

# Third Normal Form

- The same data in 3NF form:

| Tournament | Year | Winner |
|---|---|---|
| Indiana Invitational | 1998 | Al Fredrickson |
| Cleveland Open | 1999 | Bob Albertson |
| Des Moines Masters | 1999 | Al Fredrickson |
| Indiana Invitational | 1999 | Chip Masterson |

| Player | Date of Birth |
|---|---|
| Al Fredrickson | 21 July 1975 |
| Bob Albertson | 28 September 1968 |
| Chip Masterson | 14 March 1977 |

# Beyond 3NF

- Boyce-Codd Normal Form (BCNF)
- $4^{th}$ normal form, $5^{th}$ normal form, $6^{th}$ normal form
- These are rather complex; see Wikipedia if you are curious