Fundamentals/ICY: Databases 2013/14

WEEK 3: Friday
Theme intro contd; Intro to tables

John Barnden
Professor of Artificial Intelligence

School of Computer Science
University of Birmingham, UK

Reminder of Monday

SOME GENERAL THEMES

Data redundancy, data anomalies (inconsistencies).

Cross-references between places in a data repository.

Associative linking ***versus*** pointing.

Use of tables to represent things.

The mathematical underpinnings of the tables.

Data Redundancy and Anomalies

- Data Redundancy = replicating data in different places in a data repository.
  - ***E.g., in a recipe book, saying how to fry onions every time fried onions are needed in a recipe.***
- Redundancy encourages "data anomalies" and lack of "integrity" – basically, ***inconsistency*** between the different places.

  Such problems arise with insertions, deletions, and modifications in general.
- Redundancy also causes a type of ***inefficiency***: replicated updates.

Redundancy, etc., contd. 1

- Redundancy implies that if you want to modify/delete a piece of information, you need to
  - ***know whether*** there is replication, or ***check*** for possible replications
  - go to the ***effort*** of repeating changes when the item is replicated
  - ***avoid errors*** in such repeated changes.

New

Redundancy, etc., contd. 2

- Also, when something ***has*** to be repeated in different places, then:

  the more internally complex it is, then the more complex it is to use, and the more likely errors are to arise.

  - *E.g.,* repeating the tenant's full name throughout an apartment tenancy agreement.
  - Better to use a simple term such as "the tenant" and associate it in one place with the full name.
  - Another example: it's easier, and safer (in many respects), for a bank to identify customer accounts with special numbers than to use complex identifiers like "David and Samantha Cameron's lifestyle guru account".

Referential Integrity

- Referential integrity is relevant when one place in a data repository needs to refer to something in another place: ***cross-references.***
- Referential integrity is achieved when every such referring place contains a successful reference to another place or place-occupant (or no reference at all).
  - ***"Successful"*** there just means that the reference succeeds in specifying some other place(-occupant).
  - The reference may not be ***correct*** in the sense of being the one that you actually wanted!

Ways of Doing Cross-Reference

- Notice distinction above between referring to ***places*** or to ***place-occupants***: i.e., ***where*** or ***what***, respectively …

- ***Pointing*** or ***associative linking***, respectively.
  - **Your party-attendance plan for the month would use**

    pointing **if it referred to the party-givers by position, e.g. by page and line number in your address book,**

    associative linking **if it referred by means of party-givers' names.**

    Labels in a diagram are a means for *associative linking* between the diagram and the legend (= explanation of the labels, etc.) or other text.

- **The notion of "relational" database rests heavily on associative linking.**

Redundancy: Upside and More Downside

- Notice that associative linkages between different places constitute a specialized sort of ***needed*** redundancy.
- Redundancy in general does have some advantages, including:
  - ***efficiency*** **in some respects (reduce effort of following cross-references)**
  - **error or corruption in one place could in principle be *corrected* or at least *circumvented* by looking at another place (a type of *fault-tolerance*).**
- But redundancy can also prevent errors being ***detected***:
  - If the bank keeps detailed information about you (your address etc.) in just one place, associated with your account number, then an error in that detailed information ***is more likely to come to light*** precisely because its effects are more widespread.

An Analogy with Programming

- Analogous redundancy/anomaly issues arise in program text. *E.g.:*
  - If a constant numerical value such as π or g (gravitational acceleration) needs to be used in several places, best to give it a name and replicate the name, not the value. Aids consistency and maintainability.
  - If a sequence of operations needs to be invoked in many different places in the program, package it as a named procedure (function, method, …).

TABLES for REPRESENTATION

(theme introduction)

NAME	ADDRESS	PHONES	BIRTHDAY
Babloop Porkypasta	**107 Worm Drive, Hedgebarton, Birmngham, B15 9ZZ**	**0121-944-5677** **07979-888777**	11 January 1969
Coriolanus Zebedee O'Crackpotham	**The Wellyboots, Boring-under-Mosswood, Berks, HP11 1XX**	**016789-997710**	
Johnny	**Next to the Tesco's in Upper Street**	**H: 020-7111-2222** **W: 020-7111-2255** **M: 07887-842657**	???
Full Monty chip shop	**Harborne**		Oct 05
Hilary R. Clinton (grr!)	**The Old Black House, 15768 Aplanalp St., Las Cruces, NM 880011, USA**	ex-dir	16 Sep? (refused to tell me how old she was)

Problems with that Table

- Although that table illustrates the sort of table used in databases in ***some*** sense, it has many ***tricky features***:
  - **Empty entries – what's the interpretation?**
  - **Spelling error (*Birmngham)***
  - **Names/addresses of different forms (perhaps unavoidably)**
  - **Different numbers of alternatives in different cells**
  - **Different interpretations of "birthday" field (per year, or when born, or when shop opened)**
  - **Vague entries (*next to the Tesco's in Upper St.; Harborne)***
  - **Expressed uncertainty (the question marks, alone or attached)**
  - **Additional comments (*grr!, refused* …)**
  - **Exceptional entry types (*ex-dir,* and the contents of the chip-shop row)**

Question to You:

What other sorts of weird thing could happen in tables??

Restrictions on ***Database*** Tables: Overall Structure

- **Regular overall shape: rows all same length, similarly columns.**
- **No division into different regions (with a certain exception).**
- **No labels for rows, as opposed to columns.**

  Mostly no significance to the order or number of rows (but the number can change).
- **No additional comments, footnotes, etc.**

Restrictions on ***Database*** Tables: Nature of Entries

- **All cells in any one column are given the same intuitive interpretation.**
- **Each cell's item restricted to a pre-specified, fairly simple format, and all cells in any given column restricted to same format.**
- **No exceptional entries … with one exception!:**
  empty entries
- **One data item per cell (but it can be a variable-length character string, containing anything).**
- **Uncertainty and vagueness markers not supported.**

Extra, Crucial Restriction (on the main tables)

- ***No row can be repeated in a table***. (*I.e.,* no two rows can contain exactly the same values.)
- This is equivalent to saying:

  Rows are uniquely determined (picked out) by the values in some set of columns (possibly the whole set, but could be fewer).

  That is, if you imagine some values for those columns, there is at most one row that has exactly those values in those columns.

Table on next slide is closer to what might be in a database

LAST N.	FIRST N	MI	ADDRESS	Home Ph	Mobile	B year	B day
Porkypasta	Babloop		107 Worm Drive, Hedgebarton, Birmngham, B15 9ZZ	0121-944-5677	07979-888777	1969	Jan 11
O'Crackpotham	Coriolanus	Z	The Wellyboots, Boring-under-Mosswood, Berks, HP11 1XX	016789-997710			
Delfino	Johnny	-----	Next to the Tesco's in Upper Street	020-7111-2222	07887-842657		
Clinton	Hilary	R	The Old Black House, 15768 Aplanalp St., Las Cruces, NM 880011, USA		-----		Sep 16

Coordination between Tables

NAME	*PHONE*	EMPLOYER	AGE
Chopples	0121-414-3816	University of Birmingham	37
Blurp	01600-719975	Monmouth School for Girls	21
Rumpel	07970-852657	University of Birmingham	88

PHONE	TYPE	STATUS
0121-414-3816	office	OK
01600-719975	home	FAULT
0121-440-5677	home	OK
07970-852657	mobile	UNPAID

There should really be a FIRST NAME as well, in practice

EMPLOYER	ADDRESS	NUM. EMPLS	SECTOR
BT	BT House, London, …	1,234,5678	Private TCOM
Monmouth School for Girls	Hereford Rd, Monmouth, …	245	Private 2E
University of Birmingham	Edgbaston Park Rd, ….	4023	Public HE

Remember:

“Associative Linking”

This is how the tables are linked.

But:

What are the disadvantages of using character strings like “University of Birmingham” as linking values?

The Disadvantages

- In entering values, have to ensure exactly the same string of characters on each occasion
  - avoid typos on data entry
  - avoid variants: "***The*** University of Birmingham"
- Difficult to guarantee that two different entities won't have the same name.
- Inefficiency of comparing such complex values.

Reduce such problems by:

- Using artificial linking values that are simpler in form and easier to make distinct …..

Table Coordination: Revised

NAME	*PHONE*	EMPL. ID	AGE
Chopples	0121-414-3816	E22561	37
Blurp	01600-719975	E85704	21
Rumpel	07970-852657	E22561	88

PHONE	TYPE	STATUS
0121-414-3816	office	OK
01600-719975	home	FAULT
0121-440-5677	home	OK
07970-852657	mobile	UNPAID

EMPL. ID	EMPL. NAME	ADDRESS	NUM. EMPLS	SECTOR
E48693	BT	BT House, London, …	1,234,5678	Private TCOM
E85704	Monmouth School	Hereford Rd, Monmouth, …	245	Private 2E
E22561	University of Birmingham	Edgbaston Park Rd, ….	3023	Public HE

Redundancy between Tables

NAME	*PHONE*	STATUS	EMPLOYER	AGE
Chopples	0121-414-3816	OK	University of Birmingham	37
Blurp	01600-719975	FAULT	Monmouth School	21
Rumpel	07970-852657	UNPAID	University of Birmingham	88

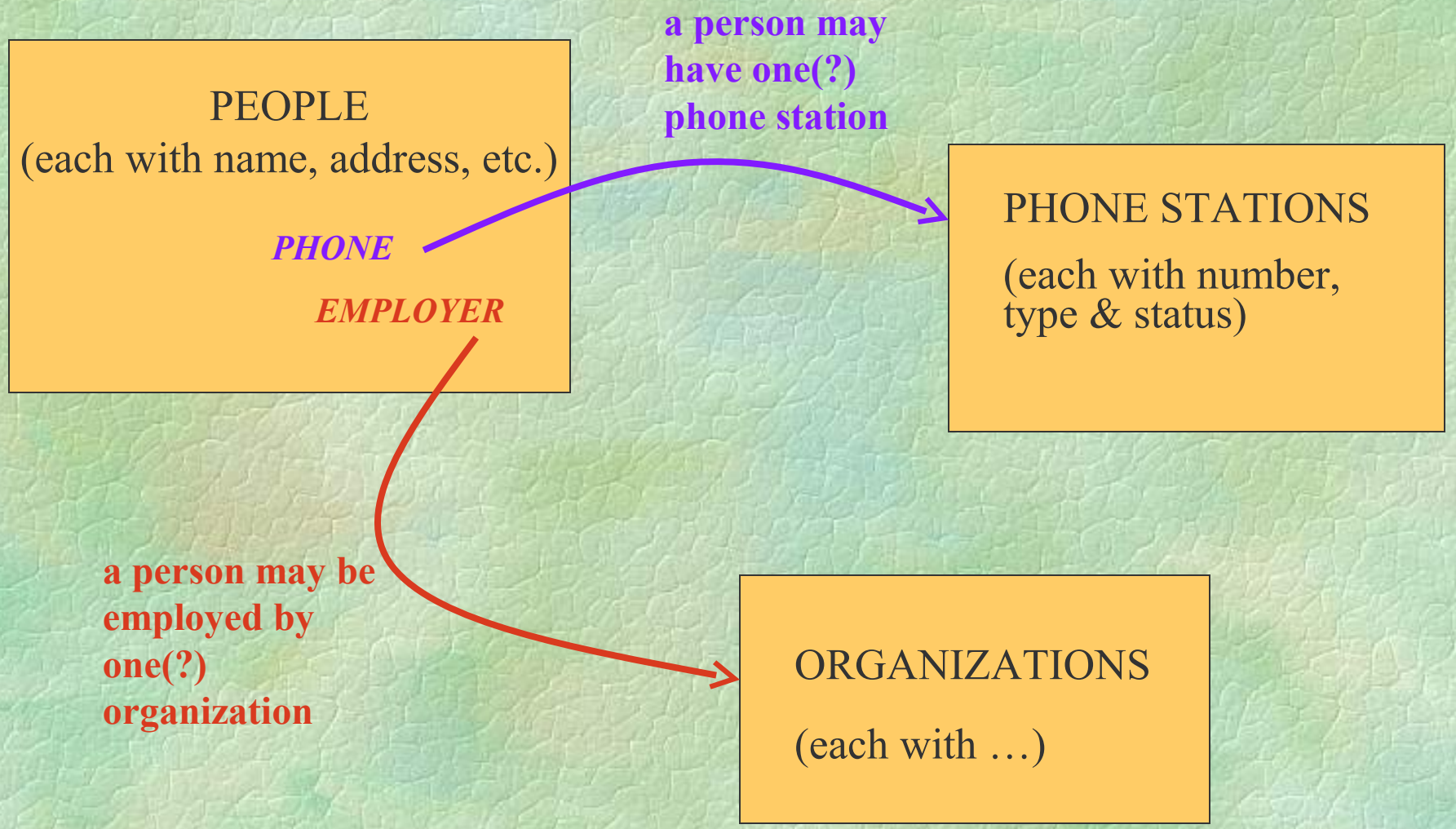
What are the advantages and disadvantages of the sharing of the STATUS attribute?

PHONE	TYPE	STATUS
0121-414-3816	office	OK
01600-719975	home	FAULT
0121-440-5677	home	OK
07970-852657	mobile	UNPAID

Tables and Things

- ***The example tables involve various types of thing:***
  - **People**
  - **People's names**
  - **Addresses**
  - **Phone numbers**
  - **Phone number types**
  - **Dates**
  - **Ages**
  - **Status indicators**
  - **etc.**

- ***and also various types of connection between things,*** *e.g.*:
  - A person having an address
  - A person being employed by an organization
  - An organization having some employees
  - A person having a birth date
  - A phone number being of a type
  - A phone number having a status
  - etc.


PEOPLE
(each with name, address, etc.)
PHONE
EMPLOYER
a person may have one(?) phone station
PHONE STATIONS
(each with number, type & status)
a person may be employed by one(?) organization
ORGANIZATIONS
(each with …)

You Judge ***Only Some Types*** of Thing to Merit Tables

- **In the example above we have decreed that** only the following TYPES of thing --- ***people, employing organizations,*** and ***phone stations ---*** correspond to WHOLE TABLES.
  - **In one table, each row represents a person.**
  - **In another table, each row represents an employing organization.**
  - **In yet another table, each row represents a phone station.**
- **We have decreed that** the other types of things, such as ***people's names, addresses, phone numbers, phone-number types,*** etc. correspond only to COLUMNS of tables, not whole tables, and each individual thing is just represented as a value in a cell.

Meriting Tables, contd.

- The question of what types of thing should correspond to tables depends on the application and your design judgment.
- It all depends on things like:
  - what range of information is needed about something
  - how separate the pieces of info about a given thing are
  - what operations are needed
  - how often they're needed.
- *For example:*

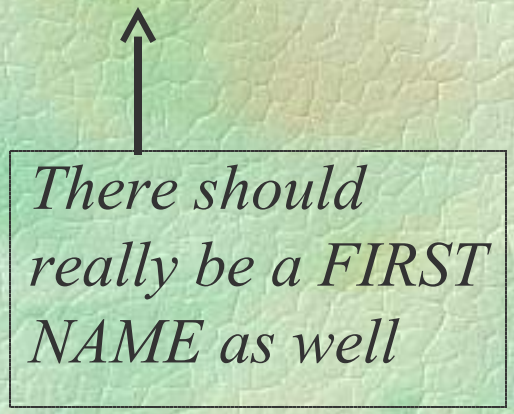
Typical Approach to Phone Numbers

NAME	*PHONE*	EMPLOYER	AGE
Chopples	**0121-414-3816**	E12345	37
Blurp	**01600-719975**	E54321	21
Rumpel	**07970-852657**	E12345	88

(There should really be a FIRST NAME as well)

But the following is ***possible*** …

NAME	*PHONE ID*	EMPLOYER	AGE
Chopples	**ABC123**	E12345	37
Blurp	**ABC137**	E54321	21
Rumpel	**DEF678**	E12345	88



PHONE ID	AREA CODE	BODY
ABC123	0121	414-3816
ABC137	01600	719975
DEF101	0121	440-5677
DEF678	07970	852657

Some Operations on Individual Tables

- Creating a new empty table of a particular "shape" (mainly, particular column names and value-types for the columns)
- Changing the "shape" of an existing table (e.g., adding/deleting a column, or changing the type of a column)
- Adding a row or rows to a table
- Deleting a row or rows (question: how identified?)
- Updating values in an individual cell (column specified by name; but how identify the row?)

More Operations on Individual Tables

- Retrieving values from an individual cell; doing calculations on them
- Retrieving the values in the cells in some or all columns for some or all rows
- Calculating statistics concerning values in particular columns across all rows, a subset of rows, or several subsets of rows (count, max, min, average, standard deviation, …)
- Ordering rows in different ways in displays of a table.

Operations on Coordinated Tables

- Need to be able to combine data from related tables in a variety of ways. *E.g.:*
  - Join tables together in various ways
  - Select things from one table on the basis of information in others
- Need to ensure consistency between related tables. *E.g.:*
  - Deletion of something in one table may require deletions from or other modifications to other tables.

ENTITIES, RELATIONSHIPS & ATTRIBUTES

(Introduction)

Entities

- Basically, entities are just things of the "important types" that we judged above to merit tables. So we had ***entity types*** such as:
  - People
  - Employing Organizations
  - Phone Stations (as opposed to just phone numbers as such)
- So what the entity types are in a given working environment are partly a matter of judgment, as explained earlier.

  But we'll see that in designing a DB we may need to introduce new, not immediately obvious, entity types.
- ***"Entities"*** are, or should be, the things of a type: *e.g.*, ***individual*** people. An entity is represented by a ***row*** in the appropriate table.

Entity Terminology

- ***Unfortunately:***

  "**entity**" is often used to mean entity type.

  "**entity set**" is often used for entity type.

  "**entity occurrence**" is often used to mean individual entity.