Relational data bases

- 3rd Module EPISOUTH -

Madrid, 18.06.2009

Martin Mengel
ISCIII / EPIET

What is a relational database

What is a database?

a collection of data (information) on a specific topic stored in an organized manner

e.g.:A book, a paper sheet, an Excel sheet

What is a relational database

a fractionized collection of data (information) on a specific topic stored in various parts but linked by a common key

E.g.: various books put together

e.g.:an Encyclopedia, a library,

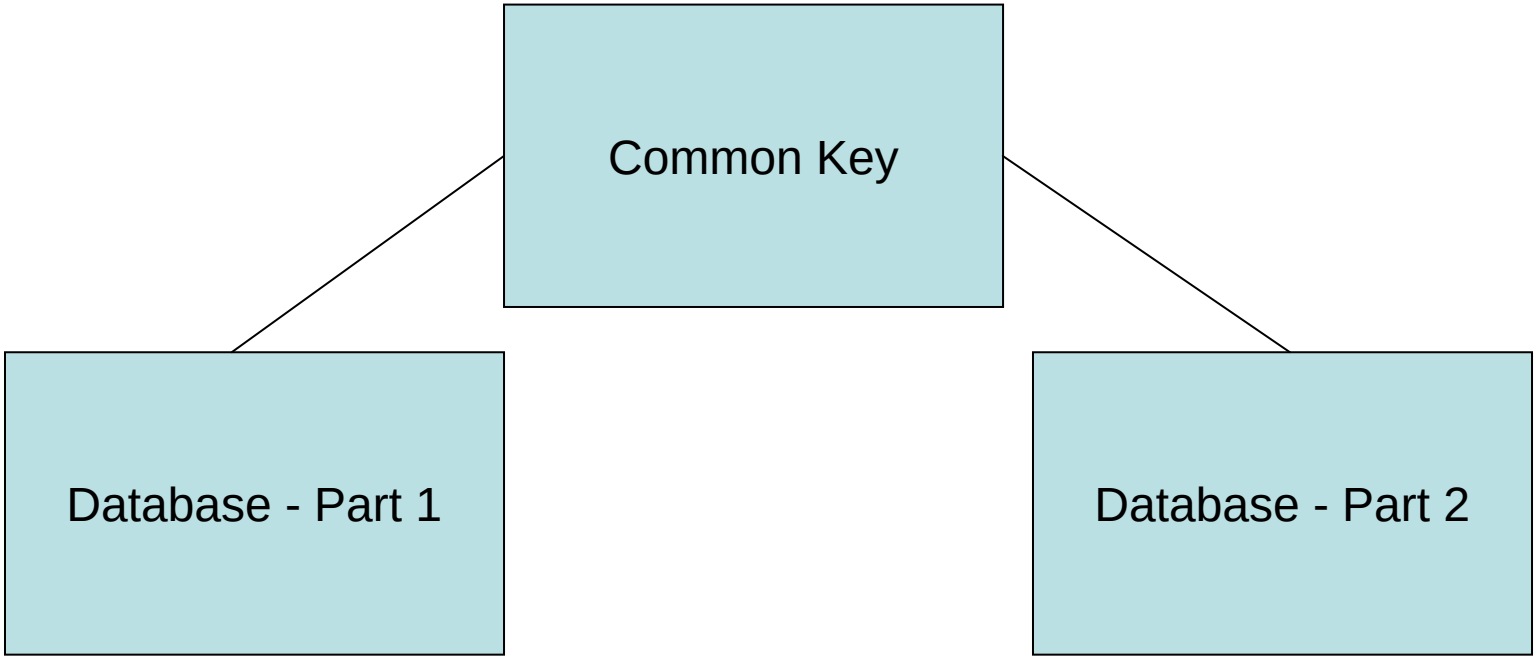
an Access database

Databases nowadays

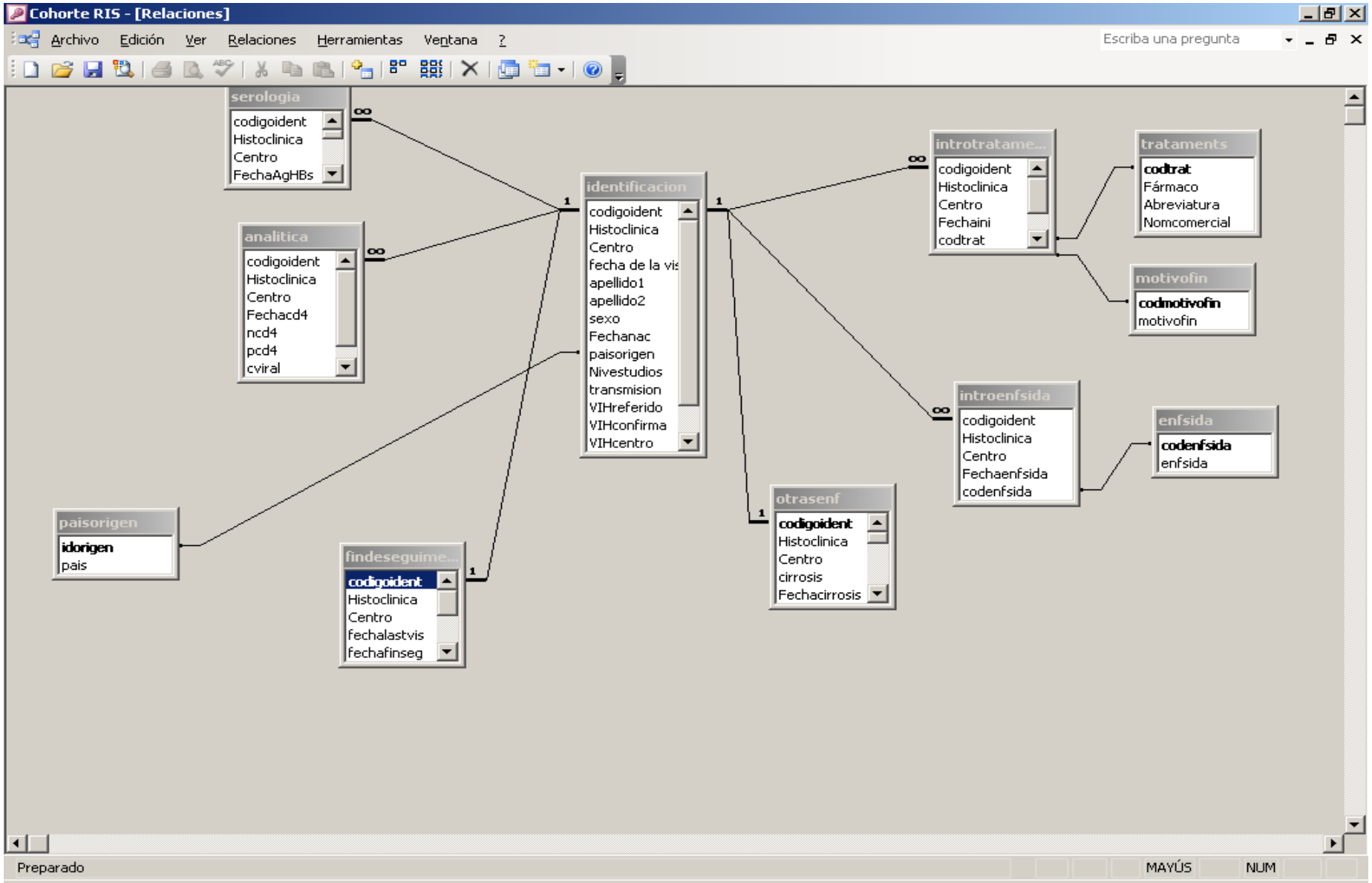
ID	Name	Date of birth	Sex	Lab result	...
1	Smith Mike	10.10.1970	male	neg	…
2	Roth Karl	10.05.1947	male	neg	…
3	Mia Carole	08.09.1956	female	pos	…
4	Schmid Jan	07.11.1935	male	pos	…
…	…	…	…	…	…

Relational Database – Basic scheme

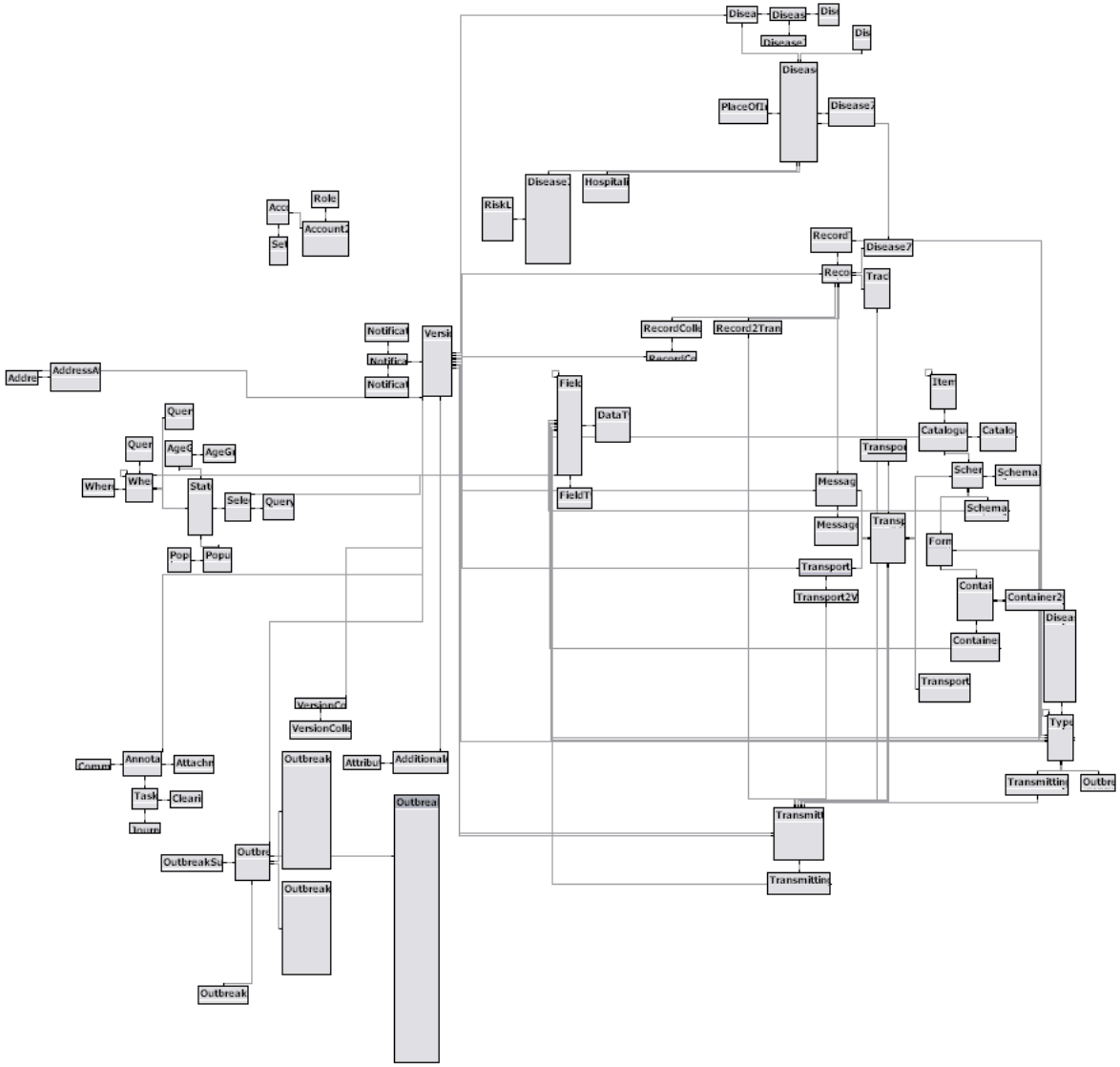
- is a collection of relations (tables)
- related records are linked together with a "key"



HIV Database



Tables of SurvNet@RKI



Examples for databases

- SurvNet@RKI
- Consists of more than one table
- 115,000 outbreaks, 2.9 million cases
- Including versions:
  202,000 outbreaks, 4.3 million cases
- Size: approx. 15GB

Examples for databases

- Outbreak of gastroenteritis in a nursing home
- Single table (flat file database)
- 1 outbreak, 103 entries
- Size: < 10 kB

How to design a relational database?

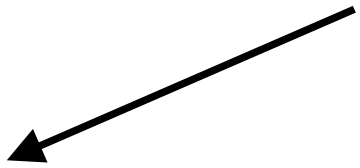
- **Approach: From flat file to relational database**

ID	Name	Date of birth	Sex	Lab result	...
1	**Smith Mike**	**10.10.1970**	**male**	**neg**	**…**
2	**Roth Karl**	**10.05.1947**	**male**	**neg**	**…**
3	**Mia Carole**	**08.09.1956**	**female**	**pos**	**…**
4	**Schmid Jan**	**07.11.1935**	**male**	**pos**	**…**
…	**…**	**…**	**…**	**…**	**…**

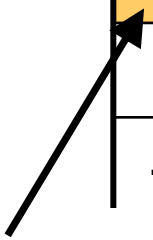
Relational database terminology

- **Relation (=Table)**

Attribute (=Column)



ID	Name	Date of birth	Sex	Lab result	...
1	**Smith Mike**	**10.10.1970**	**male**	**neg**	**…**
2	**Roth Karl**	**10.05.1947**	**male**	**neg**	**…**
3	**Mia Carole**	**08.09.1956**	**female**	**pos**	**…**
4	**Schmid Jan**	**07.11.1935**	**male**	**pos**	**…**
…	**…**	**…**	**…**	**…**	**…**



Tuple (=Row)

Data collection form for Tb surveillance

Name: Smith Mike Sex x□M □F

Date of birth: 10/10/1970 Age: 33

Country:UK City: London

ZIP: 011222

Date of diagnosis: 1/05/2003

Case x□Yes □No

Visit 1	Date: 28/04/2003	Lab result □P x□N
Visit 2	Date: 05/05/2003	Lab result x□P □N
Visit 3	Date:___/___/___	Lab result □P □N
Visit 4	Date:___/___/___	Lab result □P □N

Data entry

Name	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith Mike	male	10.10.70	UK	London	1222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth Karl	male	10.05.47	GER	Berlin	10405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia Carole	female	08.09.56	FRA	Paris	75000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid Jan	male	07.11.35	GER	Munich	8900	11.06.03	no	01.03.03	neg	01.03.03	neg

Normalisation: atomic values

Name	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith Mike	male	10.10.70	UK	London	1222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth Karl	male	10.05.47	GER	Berlin	10405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia Carole	female	08.09.56	FRA	Paris	75000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid Jan	male	07.11.35	GER	Munich	8900	11.06.03	no	01.03.03	neg	01.03.03	neg

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith	Mike	male	10.10.70	UK	London	1222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth	Karl	male	10.05.47	GER	Berlin	10405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia	Carole	female	08.09.56	FRA	Paris	75000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid	Jan	male	07.11.35	GER	Munich	8900	11.06.03	no	01.03.03	neg	01.03.03	neg

Normalisation

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith	Mike	male	10.10.70	UK	London	1222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth	Karl	male	10.05.47	GER	Berlin	10405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia	Carole	female	08.09.56	FRA	Paris	75000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid	Jan	male	07.11.35	GER	Munich	8900	11.06.03	no	01.03.03	neg	01.03.03	neg

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith	Mike	male	10.10.70	UK	London	001222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth	Karl	male	10.05.47	GER	Berlin	010405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia	Carole	female	08.09.56	FRA	Paris	075000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid	Jan	male	07.11.35	GER	Munich	008900	11.06.03	no	01.03.03	neg	01.03.03	neg

Coding

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith	Mike	male	10.10.70	UK	London	11222	14.04.03	yes	28.04.03	neg	05.05.03	pos
Roth	Karl	male	10.05.47	GER	Berlin	10405	15.04.03	no	28.04.04	neg	03.05.03	neg
Mia	Carole	female	08.09.56	FRA	Paris	75000	20.05.03	yes	22.05.03	pos	22.05.03	pos
Schmid	Jan	male	07.11.35	GER	Munich	89000	11.06.03	no	01.03.03	neg	01.03.03	neg

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	DVisit2	Lab2
Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2	05.05.03	1
Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.04	2	03.05.03	2
Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1	22.05.03	1
Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2	01.03.03	2

Normalisation: eliminate redundant information

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab	Visit2	Lab2
Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2	05.05.03	1
Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.03	2	03.05.03	2
Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1		
Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2		

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	DVisit	Lab
Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2
Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	05.05.03	1
Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.03	2
Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	03.05.03	2
Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1
Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2

but not so efficient

Normalisation: eliminate redundant information

ID	FirstN	LastN	Date of birth	...
1	**Mike**	**Smith**	**10.10.1970**	**...**
2	**Karl**	**Roth**	**10.05.1947**	**...**
3	**Carole**	**Mia**	**08.09.1956**	**...**
…	**…**	**…**	**…**	**…**

ID	DVisit	Lab
1	**28.04.03**	**neg**
1	**05.05.03**	**pos**
2	**28.04.03**	**neg**
2	**03.05.03**	**neg**
3	**22.05.03**	**pos**
…	**…**	**…**

Creating an identifier

LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	Dvisit	Lab	Visit2	Lab2
Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2	05.05.03	1
Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.03	2	03.05.03	2
Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1		
Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2		

Better adding flexibility not using the contents of the records.

ID	LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	Visit1	Lab1	Visit2	Lab2
1	Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2	05.05.03	1
2	Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.03	2	03.05.03	2
3	Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1		
4	Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2		

Referenced and referencing table

Referenced table

ID	FirstN	LastN	Date of birth	...
1	**Mike**	**Smith**	**10.10.1970**	**...**
2	**Karl**	**Roth**	**10.05.1947**	**...**
3	**Carole**	**Mia**	**08.09.1956**	**...**
...	**...**	**...**	**...**	**...**

Primary key

Foreign key

Referencing table

ID	DVisit	Lab
1	**28.04.03**	**neg**
1	**05.05.03**	**pos**
2	**28.04.03**	**neg**
2	**03.05.03**	**neg**
3	**22.05.03**	**pos**
...	**...**	**...**

Linking tables

Referenced table

ID	FirstN	LastN	Date of birth	...
1	**Mike**	**Smith**	**10.10.1970**	**...**
2	**Karl**	**Roth**	**10.05.1947**	**...**
3	**Carole**	**Mia**	**08.09.1956**	**...**
...	**...**	**...**	**...**	**...**

Primary key

Foreign key

Referencing table

ID	DVisit	Lab
1	**28.04.03**	**neg**
1	**05.05.03**	**pos**
2	**28.04.03**	**neg**
2	**03.05.03**	**neg**
3	**22.05.03**	**pos**
...	**...**	**...**

Good practice: always primary key!

Referenced table

ID	FirstN	LastN	Date of birth	...
1	**Mike**	**Smith**	**10.10.1970**	**...**
2	**Karl**	**Roth**	**10.05.1947**	**...**
3	**Carole**	**Mia**	**08.09.1956**	**...**
...	**...**	**...**	**...**	**...**

Primary key

Referencing table

ID_Lab	ID	DVisit	Lab
1	**1**	**28.04.03**	**neg**
2	**1**	**05.05.03**	**pos**
3	**2**	**28.04.03**	**neg**
4	**2**	**03.05.03**	**neg**
5	**3**	**22.05.03**	**pos**
...	**...**	**...**	**...**

Foreign key

Primary key

1: n

2 tables !!!

ID	LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case	Visit1	Lab1	Visit2	Lab2
1	Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1	28.04.03	2	05.05.03	1
2	Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2	28.04.03	2	03.05.03	2
3	Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1	22.05.03	1		
4	Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2	01.03.03	2		

Tab Main

ID	LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns	Case
1	Smith	Mike	1	10.10.70	UK	London	011222	14.04.03	1
2	Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03	2
3	Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03	1
4	Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03	2

Tab Laboratory

ID	Visit	Lab
1	28.04.03	1
1	05.05.03	2
2	28.04.03	2
2	03.05.03	2
3	22.05.03	1
4	01.03.03	2

Relationships

More than 2 tables? Yes...

Tab Main

ID	LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns
1	Smith	Mike	1	10.10.70	UK	London	011222	14.04.03
2	Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03
3	Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03
4	Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03

1:n

1:1

Tab Laboratory

ID	Visit	Lab
1	28.04.03	1
1	05.05.03	2
2	28.04.03	2
2	03.05.03	2
3	22.05.03	1
4	01.03.03	2

Casestatus

ID	Case
1	1
2	2
3	1
4	2

One-to-many relationship

Referenced table

ID	FirstN	LastN	Date of birth	...
1	**Mike**	**Smith**	**10.10.1970**	**...**
2	**Karl**	**Roth**	**10.05.1947**	**...**
3	**Carole**	**Mia**	**08.09.1956**	**...**
...	**...**	**...**	**...**	**...**

Primary key

Foreign key

Referencing table

ID	DVisit	Lab
1	**28.04.03**	**neg**
1	**05.05.03**	**pos**
2	**28.04.03**	**neg**
2	**03.05.03**	**neg**
3	**22.05.03**	**pos**
...	**...**	**...**

1: n

One-to-many relationship

- **A one-to-many relationship occurs when one entity is related to many occurrences in another entity**

Anonymise the data !

Tab Main

ID	Sex	DBirth	Coun	City	ZIP	DOns
1	1	10.10.70	UK	London	011222	14.04.03
2	1	10.05.47	GER	Berlin	010405	15.04.03
3	2	08.09.56	FRA	Paris	075000	20.05.03
4	1	07.11.35	GER	Munich	089000	11.06.03

Tab Laboratory

ID	Visit	Lab
1	28.04.03	1
1	05.05.03	2
2	28.04.03	2
2	03.05.03	2
3	22.05.03	1
4	01.03.03	2

Casestatus

ID	Case
1	1
2	2
3	1
4	2

ID	LastN	FirstN
1	Smith	Mike
2	Roth	Karl
3	Mia	Carole
4	Schmid	Jan

Anonymise data

Referenced table

ID	Date of birth	Street	City	ZIP	...
1	10.10.1970	Leeke St	London	WC1X 9JF	...
2	10.05.1947	Phoenix Rd	London	NW1 1HB	...
3	08.09.1956	York Way	London	N1 9AA	...
...	...	...	...	...	...

Primary key

1:1

confidential

Explain the codes

Tab Main

ID	LastN	FirstN	Sex	DBirth	Country	City	ZIP	DOns
1	Smith	Mike	1	10.10.70	UK	London	011222	14.04.03
2	Roth	Karl	1	10.05.47	GER	Berlin	010405	15.04.03
3	Mia	Carole	2	08.09.56	FRA	Paris	075000	20.05.03
4	Schmid	Jan	1	07.11.35	GER	Munich	089000	11.06.03

Tab Laboratory

ID	Visit	Lab
1	28.04.03	1
1	05.05.03	2
2	28.04.03	2
2	03.05.03	2
3	22.05.03	1
4	01.03.03	2

Casestatus

ID	Case
1	1
2	2
3	1
4	2

Sex_code

code	Case
1	Male
2	Female

Case_code

code	Case
1	Neg
2	Pos

Relational database

- We can use the relational database in the forms for data entry (the data will be stored on different tables)
- During the analysis
  - Analyse patients records
  - Analyse visit records
  - Analyse change of status between visits
  - Relate table to analyse visits using patients characteristics

Essential questions

- What information is needed when by whom?
- Who updates the data & how often?
- What kind of analysis and analysis software?
- Do you need more than a flat file database?

  ➔ Data manager? Work Plan!!!

What makes a „good“ database?

- **Avoid redundant information**
- **Allow manipulation of data in efficient way**
- **Ensure data integrity**
- **Fit with the logical flow of information**
- **The database model in most common use today is the** relational **database model**

Advices

- Use always an unique identifier (preferably automatic) in every table
- Consider to use more than a table
- Enter data as numeric codes in the main tables
- Use label to store the meaning of the codes
- When you analyze the data, use relate functions
- Be careful and don’t hesitate to ask for help

Summary

- Normalization is to optimize the data for the analysis and management
- Relational databases permit to use and manage different tables
- Choose between a big table and relation database in the planning phase! Prefer the relational one
- Relational database could avoid duplication and errors
- Relational databases may require a data manager