Semi-structured Data

1 - Introduction

Andreas Pieris and Wolfgang Fischl, Summer Term 2016

Outline

- Structured Data

- Semi-structured Data

- Why Semi-structured Data?

- The Data Model

- Store Semi-structured Data

Structured Data

- Data is structured in semantic chunks - entities

> VIE, Vienna International, Vienna
>
> LHR, London Heathrow, London
>
> VIE, LHR, BA
>
> VIE, LHR, OS
>
> BA, British Airways
>
> OS, Austrian Airlines

- Similar entities are grouped together - classes
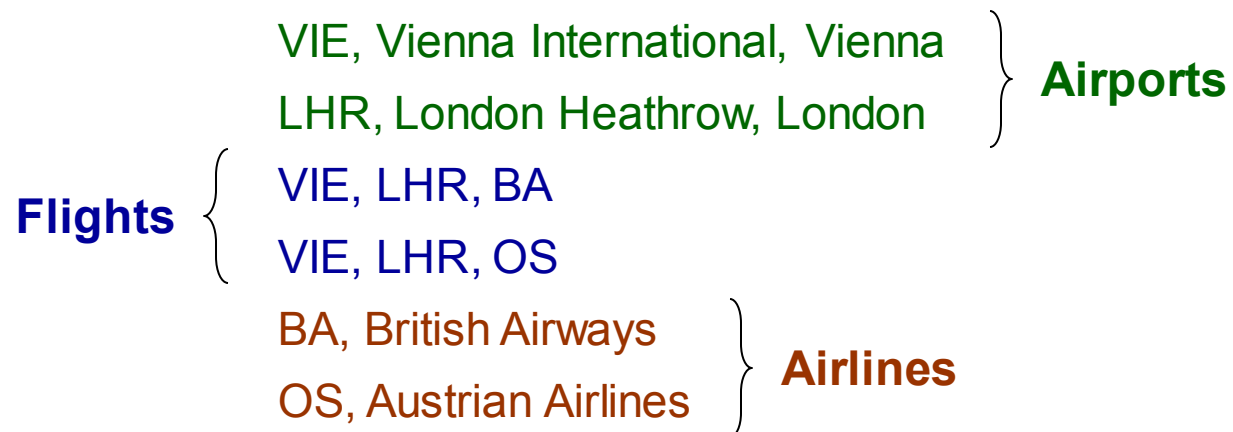
> VIE, Vienna International, Vienna
>
> LHR, London Heathrow, London   **Airports**
>
> VIE, LHR, BA
>
> VIE, LHR, OS   **Flights**
>
> BA, British Airways
>
> OS, Austrian Airlines   **Airlines**

Structured Data

- Entities in the same class have the same descriptions - attributes

Airports

(VIE, Vienna International, Vienna)
(LHR, London Heathrow, London)

(Airport_Code, Name, City)

Flights

(VIE, LHR, BA)
(VIE, LHR, OS)

(Origin, Destination, Airline)

Airlines

(BA, British Airways)
(OS, Austrian Airlines)

(Airline_Code, Name)

Structured Data

- Entities in the same class have the same descriptions - attributes

| Airports	Flights	Airlines
(VIE, Vienna International, Vienna)	(VIE, LHR, BA)	(BA, British Airways)
(LHR, London Heathrow, London)	(VIE, LHR, OS)	(OS, Austrian Airlines)
(Airport_Code, Name, City)	**(Origin, Destination, Airline)**	**(Airline_Code, Name)**

- Attributes in similar entities
  - same format (string, integer, date, etc.)
  - predefined length
  - all present
  - same order

… strict structure forced by a schema!!!

Structured Data - Relational Model

- Database model for structured data: entities   →   records (or tuples)

  classes   →   tables (or relations)

- Records grouped in tables

attribute

record {

table
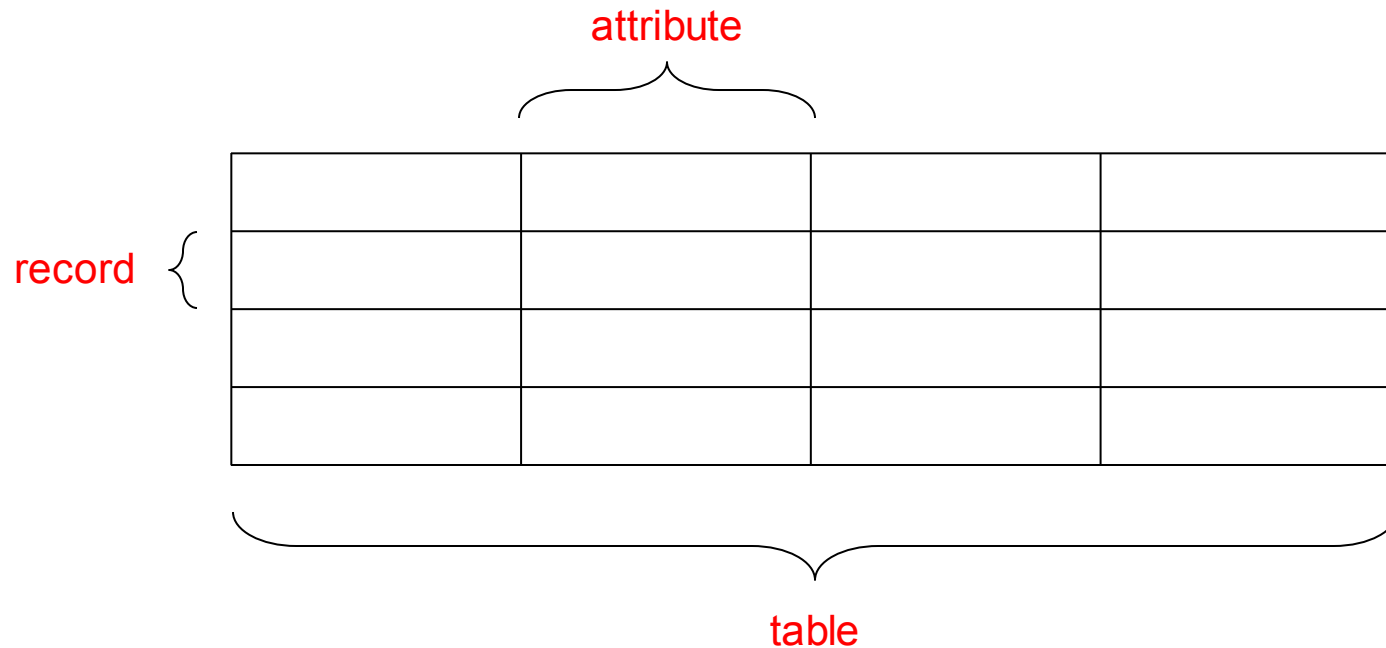
Structured Data: "On the Fly" Example

Airports	Code	Name	City
	VIE	Vienna International	Vienna
	LHR	London Heathrow	London
	LGW	London Gatwick	London
	LCA	Larnaca International	Larnaca
	GLA	Glasgow	Glasgow
	EDI	Edinburgh	Edinburgh

Flights	Origin	Destination	Airline
	VIE	LHR	British Airways
	VIE	LHR	Austrian Airlines
	LHR	EDI	British Airways
	LGW	GLA	EasyJet

Airlines	Code	Name
	BA	British Airways
	OS	Austrian Airlines
	U2	EasyJet

"Persons" Example

Gerti Kappel, 18870, 18896, gerti@big.tuwien.ac.at

Andreas, Pieris, pieris@dbai.tuwien.ac.at, 740072, 18493

Wolfgang Fischl, wfischl@dbai.tuwien.ac.at, 740050

Bill, Robert, 188316, bill@big.tuwien.ac.at

Semi-structured Data (SSD)

- Data is structured in semantic entities

  there is structure

- Similar entities are grouped in classes

- Entities in the same class may not have the same attributes

- Attributes of similar entities
  - may have different format
  - may have different length
  - not all required
  - may have different order

  but not too much structure

Semi-structured Data: "Persons" Example

Gerti Kappel, 18870, 18896, gerti@big.tuwien.ac.at

Andreas, Pieris, pieris@dbai.tuwien.ac.at, 740072, 18493

Wolfgang Fischl, wfischl@dbai.tuwien.ac.at, 740050

Martin, Fleck, 58801, fleck@big.tuwien.ac.at

- There is structure
  - Each row is a semantic entity - person
  - All entities are grouped in a class - persons

- But not too much structure
  - Entities have no regular structure
  - Structure of future entities is unpredictable

Why Semi-structured Data?

- There are data sources that we would like to treat as databases, but which cannot be constraint by a schema

- Flexible format for exchanging data between different places

… the WEB

GOAL: Reconcile document view (web) with strict structures (databases)
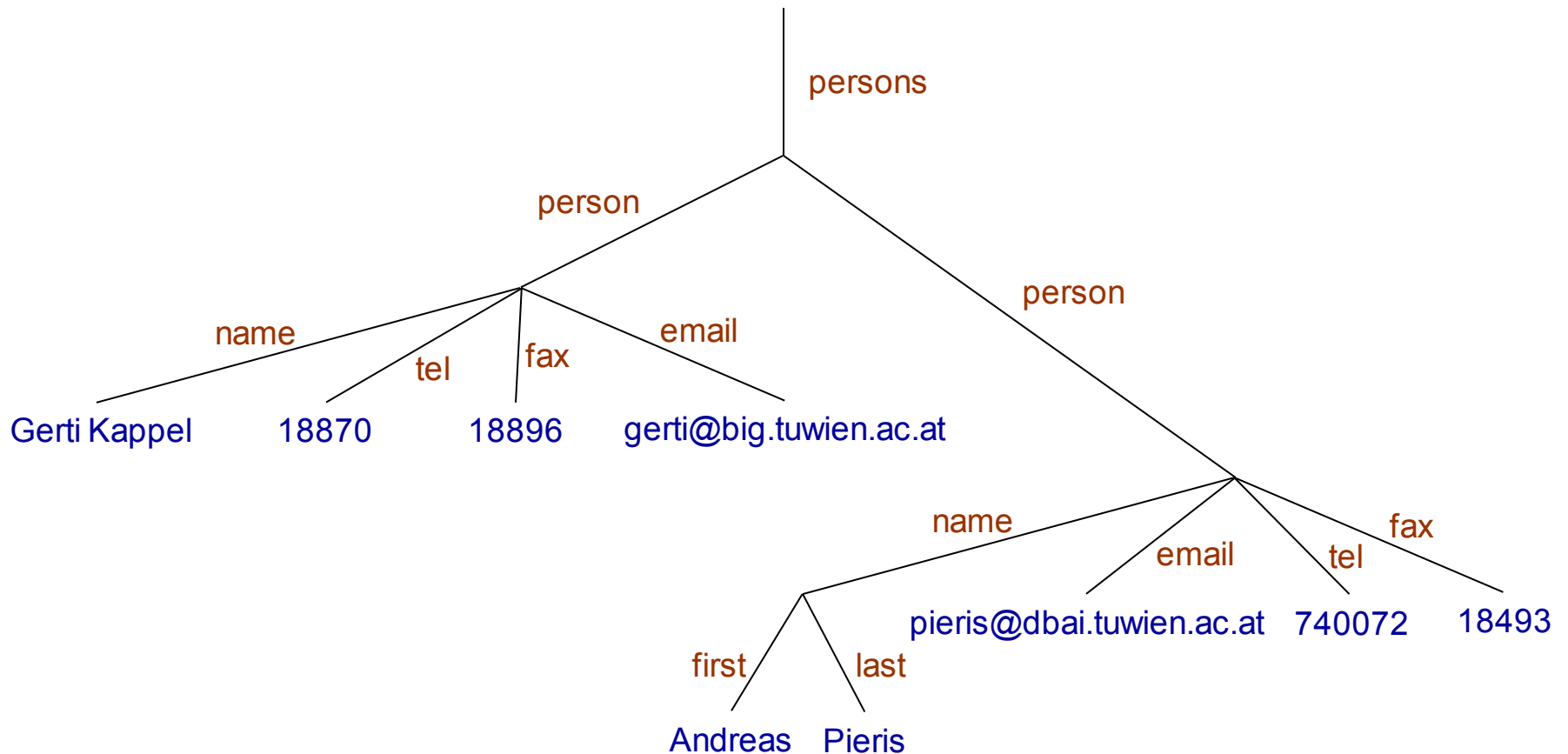
Data Model

- We need an effective way to represent semi-structured data

- Like the relational model for structured data

… any ideas?

Trees as Data Model

Gerti Kappel, 18870, 18896, gerti@big.tuwien.ac.at

Andreas, Pieris, pieris@dbai.tuwien.ac.at, 740072, 18493

Trees as Data Model

- SSD can be represented as a (labelled) tree:
  - ○ leaf nodes standing for single data items
  - ○ inner nodes have no label
  - ○ edges labelled with elements

```
                                                    │ persons
                                                    │
                                              person    person
                                             /              \
                   name    tel   fax   email                 ...
                  /       /    │      \
          Gerti Kappel  18870  18896  gerti@big.tuwien.ac.at
```
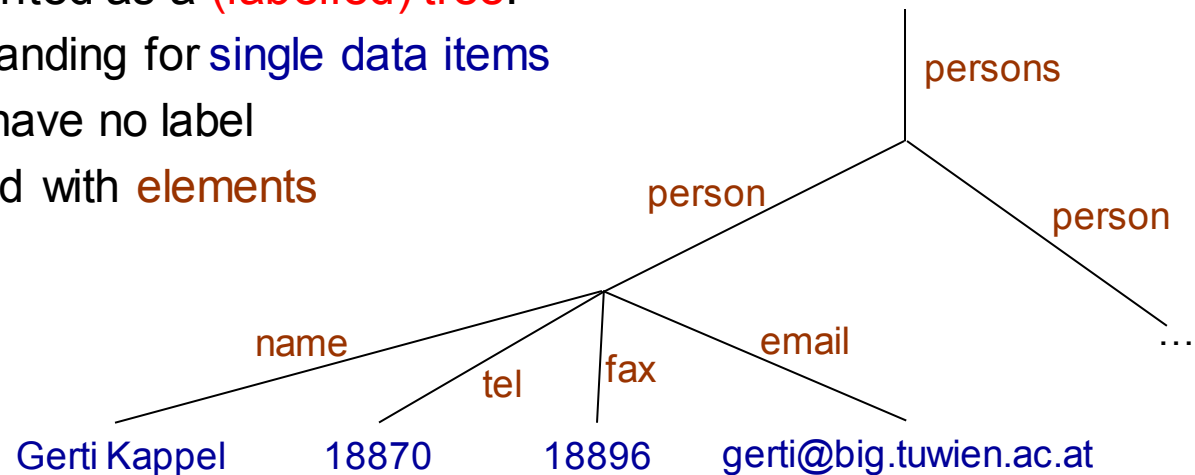
- Such a model is called self-describing - information that is usually associated with a schema is contained within the data
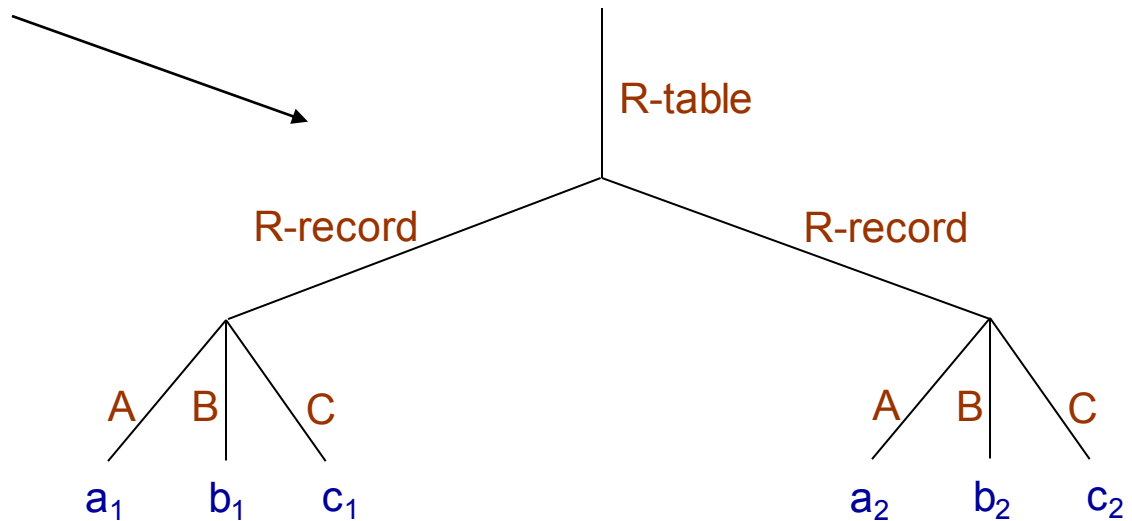
- Data carries its own description

SSD: Representing Relational Data

Structured data is a special case of semi-structured data

$\Downarrow$

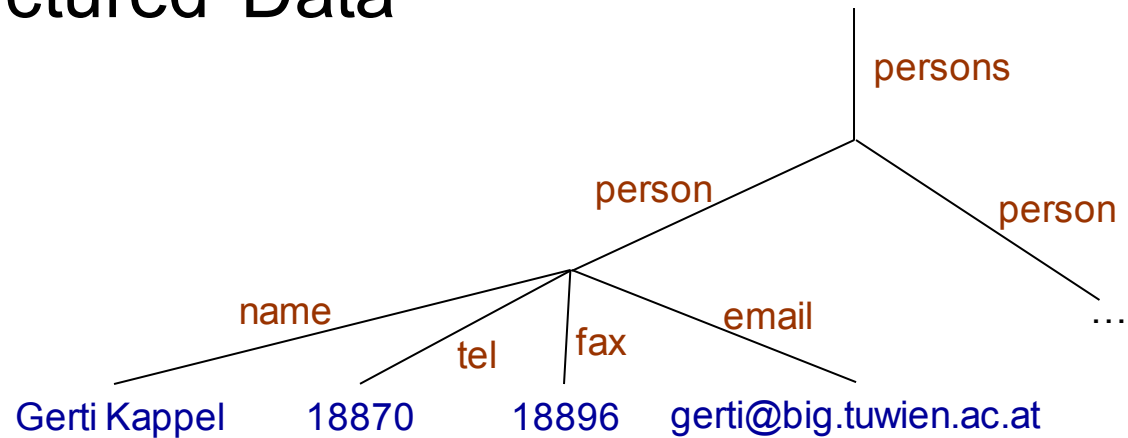relational data can be represented as a tree (with an overhead)

R	A	B	C
	a_1	b_1	c_1
	a_2	b_2	c_2

R-table

R-record   R-record

A   B   C   A   B   C

a_1 b_1 c_1 a_2 b_2 c_2
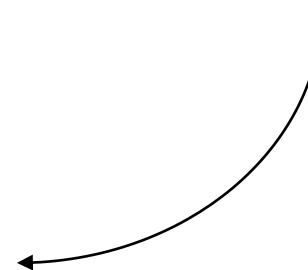
Store Semi-structured Data

- There are various formalisms to store semi-structured data
    - Object Exchange Model (OEM)
    - JavaScript Object Notation (JSON)
    - eXtensible Markup Language (XML)

Store Semi-structured Data

persons

person   person

name   …
            tel   fax   email

Gerti Kappel   18870   18896   gerti@big.tuwien.ac.at

```
{persons:
    {person:
        {name: "Gerti Kappel"
         tel: 18870
         fax: 18896
         email: "gerti@big.tuwien.ac.at"}
    }
    {person:
        {name:
            {first: "Andreas",
             last: "Pieris"}
         email: "pieris@dbai.tuwien.ac.at"
         tel: 740072
         fax: 18493}
    }
}
```
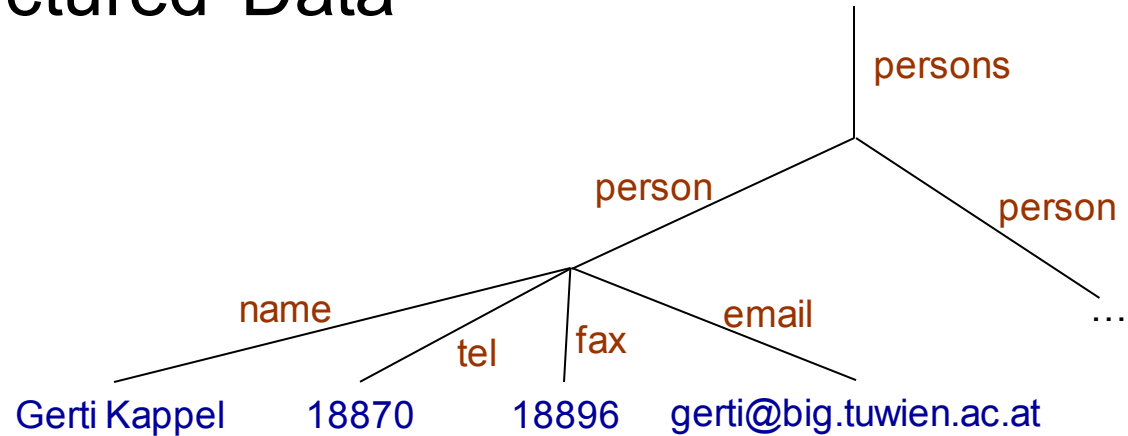
OEM Representation

Store Semi-structured Data

persons

person   person

name   tel   fax   email   ...

Gerti Kappel   18870   18896   gerti@big.tuwien.ac.at

XML Representation

```
<persons>
    <person>
        <name> Gerti Kappel </name>
        <tel> 18870 </tel>
        <fax> 18896 </fax>
        <email> gerti@big.tuwien.ac.at </email>
    </person>
    <person>
        <name>
            <first> Andreas </first>
            <last> Pieris </last>
        </name>
        <email> pieris@dbai.tuwien.ac.at </email>
        <tel> 740072 </tel>
        <fax> 18493 </fax>
    </person>
</persons>
```
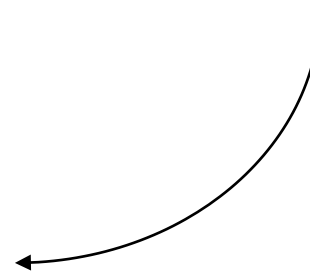
Store Semi-structured Data

- There are various formalisms to store semi-structured data
    - Object Exchange Model (OEM)
    - JavaScript Object Notation (JSON)
    - eXtensible Markup Language (XML)

- Different syntax

- Different mechanisms for self-describing

but the goal is the same:

store SSD

- Different description mechanisms
    - Which attributes are allowed/required
    - Which values are allowed/required

- Different query languages and manipulation mechanisms

Sum Up

- **Structured Data**
  - Similar entities grouped in classes
  - Similar entities have a regular structure
  - Relational Model

- **Semi-structured Data**
  - Similar entities grouped in classes
  - Similar entities have irregular structure
  - Trees as a Model

- **Store Semi-structured Data**
  - Various formalisms
  - eXtensible Markup Language (XML)