

Data Integration and Exchange: Course info

- Mondays, 11:10-13:00, FH 1B01 (at least for now)
- Prerequisites: Database Systems
- Text:
 - For data integration: none (because there isn't one...)
 - For data exchange: [Relational and XML Data Exchange](#), by Arenas, Barceló, Libkin, and Murlak, published by Morgan&Claypool, 2010. Not yet free from UoE.
- Slides will be posted on the course webpage:
<http://homepages.inf.ed.ac.uk/libkin/teach/dataintegr10>
- Surveys by Lenzerini and Halevy (see links on the webpage)
- 3 assignments (each worth 10%) and final exam (70%)
- Office hours: by appointment (usually works better for UG4)

Why do you need this course

- Databases are everywhere these days ($> \$2 \cdot 10^{10}$ /year business — whatever that means today)
- Every enterprise has a database; they merge, combine data – hence data integration
- In addition, a lot of data is available on the web, but often one needs many sources to answer a query
- Hence (almost) everyone needs to integrate data
- Huge investment from leading companies, IBM, Oracle, Microsoft
- Very ad hoc solutions; but finally we understand what the real problems in data integration are, and have some solutions (but not all!)

Data Integration and Exchange

- Traditional approach to databases:
 - A single large repository of data.
 - perhaps distributed across several sites
 - Database administrator in charge of access to data.
 - Users interact with the database through application programs.
 - Programmers write those (embedded SQL, etc)
 - Queries dominate; updates less common.
 - DMBS takes care of lots of things for you
- But the world is changing.

What happens these days

- Many huge repositories are publicly available
- Many queries **cannot** be answered using a single source.
- Often data from various sources needs to be combined, e.g.
 - company mergers
 - restructuring databases within a single organisation
 - combining data from several private and public sources
- Different sources have different structures/models
- Only portions of the data from some database could be available.
- Our view of the world may be very different from the view of the databases we need to use

Integration and Exchange

- Integration: answer queries using multiple sources:
 - virtual approach, or
 - materialization
- Exchange: transfer data between two legacy database schemas
- What changes:
 - no clear notion of an **answer to a query**
 - data is not clean: incomplete, inconsistent
 - data may not even exist (virtual integration)
- Our goal: study the main concepts and techniques for creating and querying integrated/exchanged data

Main topics

- data integration basics
 - scenarios, overview of products, techniques
- integration and views
- incomplete information
- relational data exchange
- overview of commercial tools
- XML data exchange
- schema mappings
- inconsistent databases
- top-k queries