CM3010 Databases and Advanced Data Techniques

Course Notes

WEEK 1 COMPLETED

UPDATED APRIL 18, 2021 BY TONI HE

Week 1

1.001 Welcome to the module

In this module, we'll be focusing on ***Data***, in particular:

- Understanding structure of data
- Getting and cleaning data
- Storing and retrieving data
- Using and sharing data

1.003 Where can you find usable Data?

Whilst many organisations and individuals make large amounts of data openly available, it can be hard to find. The Open Data Institute founded by Sir Tim Berners-Lee and Sir Nigel Shadbolt is dedicated to getting large-scale open publication of useful data started.

The ODI guides cover a broad range of topics, covering policy, data resources and projects. Choose a domain that interests you and explore what they have.

1.101 Data Sources, where does data come from?

- New data, building/generating new data
- Pre-existing ie: internal 'legacy data', external data ie: provided by suppliers.

NEW DATA

- Easiest when you build it with the database application, adding data as you go.
Ex: organization with appointments and bookings with clients, any new bookings gets added to the database, can be tested with the smaller dataset.
- Bulk data entry

PRE-EXISTING DATA

May need:

- Extraction
- Conversion
- Cleaning

EXTERNAL DATA SOURCES:

Pros:

- No costs for data entry
- No costs for quality checks
- Delegated expertise

Cons:

- No control over data quality
- No control over data structure
- May be incomplete
- May be ambiguous
- Questions of trustworthiness

1.103 Ordering some data: What's on the menu?

In two blog posts, Trevor Munoz (Interim Head of the Maryland Institute for Technology in the Humanities), introduces and explores a dataset of menus. The menu transcriptions have been released by the New York Public Library as Open Data.

Post 1: Trevor Munoz, 'What IS on the menu'
Post 2: Trevor Munoz, 'Refining the problem'

The menu project has a full website which offers the ability to explore the menus, download transcribed data and contribute to the dataset:

What's on the menu

1.105 What does your data look like?

Sometimes the external source of information may seem ambiguous or incomplete:

The way data is used will be different amongst different groups.

When you model real life data and prepare to put it into a database you have to be thinking about:

- The use of the data
- The sort of information you'd be interested in

For example a book.

You might want to include it's:

- title
- author
- publisher
- date
- weight
- number of pages

1.203 Licenses, sharing and ethics

Decisions about whether your data will be proprietary or open are often taken at an institutional level. Research funding bodies, normally require your data to be open, and many governments and their agencies have similar restrictions.

Whichever way these decisions go, the exact nature and terms of the license can have a significant impact on the legal re-use of your data. The Ball article summarises some of the issues, the advantages and the disadvantages of different licenses for open data. As you will see, there are some differences between the issues that arise for data and those for software.

Alex Ball, How to License research data

Week 2
Week 3

Week 4

Week 5

Week 6

Week 7

Week 8

Week 9

Week 10

Week 11

Week 12

Week 13

Week 14

Week 15

Week 16

Week 17

Week 18

Week 19

Week 20

Week 21

Week 22

Learning Objectives

- describe the concepts taught in this course
- apply the concepts taught in this course to solve problems
- analyse algorithms and data structure in terms of the concepts taught in this course

EXAM: September 6, 2021