Extracting and Integrating Data
Rattapom Tuchinda

What’s a Mashup?
A website or application that combines content from more than one source into an integrated experience [wikipedia]
a) LA crime map  b) zillow.com  c) Ski bonk
Combined Data gives new insight / provides new services

Existing Approaches
Goal: Create Mashups without Programming
• Addresses syntax issues, but users still required to understand programming concepts

Mashup Building Issues
Data Retrieval
Wrapper
Wrapper
Calibration
- source modeling
- cleaning
Clean
Clean
Integration
Combine
Display

Dapper (www.dapper.net)
• Data Extraction
• Light Filtering
• Linking Sources (dapps)
• Output
  – Feed
  – XML
  – Google Maps

Outline
• Introduction
• Dapper
• Yahoo’s Pipes
• Assignment

Can we come up with a framework that addresses all of the issues while still making the Mashup building process easy?
How does it work?

Creating a new Dapp

Creating a new Dapp

Layout

Adding samples

Tutorials

Extracting an element
Parallel element extraction

Simple text filtering

Defining a field

Defining a field (2)

Errors

Getting rid of extraction errors
Defining another attribute

Grouping

Grouping (2)

Grouping (3)

Saving

Exporting and executing
Yahoo's Pipes (pipes.yahoo.com)

- Data Extraction
- Data Processing
- Linking Sources
- Output
  - Feed
  - Google Maps

How does it work?

Widget
Connector
Intermediate output

Fetching Feed

Checking result: Don’t panic if sometimes it doesn’t show up
Advance Features

- Dapper
  - Linking Dapp
  - Form Input
  - Parameterize URL

Linking: Searching for dapps

Dapp Linker

Linking: matching attributes

Form Input

Extracting
Grouping

Customizing Input

Customizing Input: define a variable

Customizing Input: define a variable (2)

Using input variables

Using input variables: execution
Using input variables: as a URL

Advance Functionalities

- Regex
- Looping
- And more!

Advance widgets

Homework

- Task
  - Create a Mashup using both Dapper and Yahoo’s Pipes.
  - You can use existing feeds as parts of the Mashup, but at least one source must be extracted using Dapper
- Hints
  - You need to sign-up on both services
  - Important: Make sure that you “publish” it, so it won’t get deleted in 24 hrs
  - Look at existing dapps and pipes to understand how it works
  - Look at programmable web (http://www.programmableweb.com/) for Mashups ideas.
  - Dapper is known to go down once in a while. Make sure you work on Dapper early.
  - Dapper --> Pipes --> Dapper --> Pipes???

Homework cont.

- What to turn in
  - 1 page writeup
  - Snapshot of your Mashups/output
  - What’s it? Why is it interesting? What sources are you using? What operators?
  - Links to the Mashup that you’ve created
- Score
  - Creativity 1 point
  - Mashup runs 1 point
  - Writeup 0.5 point

Questions?