Building Mashups

Craig Knoblock
University of Southern California

Thanks to Rattapoom 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
Outline

- Manual Mashup Construction
- Manual Integration Specification
- Widget-based Approach to Integration
- Programming by demonstration
Outline

• Manual Mashup Construction
• Manual Integration Specification
• Widget-based Approach to Integration
• Programming by demonstration
Manual Mashup Construction

• User simply specifies the data and the integration with a map
• Easy to use tools in Google Maps to build and share your own application
• But, requires the user to specify and maintain all of the data
Google MyMap Video
Outline

• Manual Mashup Construction
• Manual Integration Specification
• Widget-based Approach to Integration
• Programming by demonstration
Intel Mashmaker

• Multi-tier user
  – Naïve users
  – Expert users

• Experts do all the hard work to customize the integration between sources

• Naïve users browse web pages normally
  – If the page that the user is viewing contain an existing wrapper or predefined integration, the user can get those information by pressing a button
Intel Mashmaker: Design Principles

• Program as you browse
  – view Mashup creation as an extension of the normal web browsing habits

• Direct manipulation
  – work on data without having to think about abstract concepts such as programs

• Pay as you go
  – Unskilled users should be able to gain some benefit with very little effort
  – Experts should be able to do more advanced stuff
Intel Mashmaker: Features

• Look at Dapper to see if the wrapper for a particular site exist

• Direct manipulation of data through operations such as map, fold, and filter

• User can interact with Mashmaker at a number of different levels depending on the skill
Intel Mashmaker: Users

• Basic: know nothing
• Normal: Occasionally expand the widget panel to edit form parameters
• Skilled: Connecting sources
• Semi-Expert: Extract data from new sites
• Expert: Write complex expression directly in Mash-Maker’s core language
• Gurus: Teach Mashmaker to understand the content of the new website.
Mashmaker Video
Outline

- Manual Mashup Construction
- Manual Integration Specification
- Widget-based Approach to Integration
- Programming by demonstration
Goal: Create Mashups without Programming

- Addresses syntax issues, but users still required to understand programming concepts

Widget Paradigm

- Widgets (i.e., 43 for Pipes, 300+ for MS) represents an operation on the data.
- Locating and learning to customize widget can be time-consuming.
- Most tools focus on particular issues and ignore others.
Marmite

• Widget/Workflow approach similar to Yahoo’s Pipes and Microsoft’s Popfly
• Firefox extensions
• The interface is divided into three sections
  – Widget selection
  – Workflow
  – Intermediate results

### Step 2: Extract Address

**Workflow**

1. **Select Links From Page**

   ![Select Links From Page](select_links_from_page.png)

2. **Extract Address**

   ![Extract Address](extract_address.png)

- **Input**
  - Get: from column:
  - Web Page: B: URL

- **Output**
  - Write: to column:
  - Street Address: new column
  - City: new column
  - State: new column

**Next operator suggestions.**

Replace this placeholder by selecting the next operator from the operator list or the suggestions list.

Show suggestions
Marmite Approach

• Based on Apple Automator
• One of a few that design the system by doing user studies prior implementation
  – Showing intermediate result
  – Suggestion for the next operators
Marmite Evaluation

• 6 People
  – 2 novices
  – 2 people who know how to use spreadsheet
  – 2 programmers

• 4 Tasks
  – Retrieve a set of addresses and geocode an address
  – Search and filter out events further than a week away
  – Compile a list of events from two event services and plot them on a map.
  – Recreate the map from housingmaps website
Marmite Result

• 3 people (1 spreadsheet, 2 programmers) complete the 4 tasks in one hour.
  – Novices did not finish all the tasks.

• The biggest problem for them is understanding data flow
  – Confusion about the input/output concept
  – Did not understand that the data flow and the spreadsheet result are linked.
Marmite Video
Outline

- Manual Mashup Construction
- Manual Integration Specification
- Widget-based Approach to Integration
- Programming by demonstration
Programming by Demonstration Approach

- Focus on data, not on the process
  - Users are already familiar with data.
  - Capture and model the Mashup building process from examples (PBD)
- Consolidate rather than Divide-And-Conquer
  - Solving one issue can help solve other issues.
  - Use one interaction platform -- a table
- Leverage existing database
  - Helps with source modeling, cleaning, and data integration.
1. **Japon Bistro**
   970 E Colorado Blvd, Pasadena, CA, 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

2. **Hokusai**
   8400 Wilshire Blvd, Beverly Hills, CA, 90211
   Chic elegance and modern Zen style surround Japanese French this paean to haute cuisine and stylized sushi.

3. **Sushi Sasabune**
   12400 Wilshire Blvd, Ste 150, Los Angeles, CA, 90025
   Sushi is the singular star at this Zen Westside palace that bows only to the royalty of chef and fish.

4. **Sushi Roku**
   8445 W 3rd St, Los Angeles, CA, 90048
   High fashion, rock and roll and Hollywood buzz converge over innovative sushi.
1. **Japon Bistro**  
627 E Colorado Blvd, Pasadena, CA, 91106  
Upscale yet affordable Japanese eatery offers the city’s largest sake selection.

2. **Hokusai**  
8400 Wilshire Blvd, Beverly Hills, CA, 90211  
Chic elegance and modern Zen style surround Japanese-French paeans to haute cuisine and stylized sushi.

3. **Sushi Sasabune**  
12400 Wilshire Blvd Ste 160, Los Angeles, CA, 90025  
Sushi is the singular star at this Zen Westside palace that bows only to the royalty of chef and fish.

4. **Sushi Roku**  
8445 W 3rd St, Los Angeles, CA, 90048  
High fashion, rock and roll and Hollywood buzz converge over innovative sushi.
Source Modeling (Attribute selection)

Possible Attribute
\{a | a, s: a \in att(s) \land (val(a, s) \subset V) \}

restaurant name (3)
artist name (1)
Data Cleaning: using existing values

Newly extracted data

<table>
<thead>
<tr>
<th>Restaurant Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
</tr>
<tr>
<td>Hokusai</td>
</tr>
<tr>
<td>Sushi Sasabune</td>
</tr>
<tr>
<td>Sushi Roka</td>
</tr>
</tbody>
</table>

Data repository

<table>
<thead>
<tr>
<th>Restaurant Name</th>
<th>Address</th>
<th>...</th>
<th>Health Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hokusai</td>
<td>8400..</td>
<td>...</td>
<td>90</td>
</tr>
<tr>
<td>Katana</td>
<td>8439..</td>
<td>...</td>
<td>99</td>
</tr>
<tr>
<td>Japon Bistro</td>
<td>927 E..</td>
<td>...</td>
<td>95</td>
</tr>
</tbody>
</table>

Zagat

<table>
<thead>
<tr>
<th>Restaurant Name</th>
<th>Zagat Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sushi Sasabune</td>
<td>27</td>
</tr>
<tr>
<td>Sushi Roka</td>
<td>25</td>
</tr>
<tr>
<td>Katana</td>
<td>23</td>
</tr>
</tbody>
</table>
Data Cleaning: using predefined rules

Subset Rule:
\[(s_1s_2...s_k) \rightarrow (d_1d_2...d_t) \land (k \leq t) \land s_i \in \{d_1,d_2,...,d_t\} \land d_i \neq d_j\]

28 Reviews $\rightarrow$ 28
Data Integration

Based on [tuchinda 2007]

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Address</th>
<th>Description</th>
<th>Number of Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
<td>927 E Color Pl</td>
<td>Upscale yet</td>
<td>28 Reviews</td>
</tr>
<tr>
<td>Sushi Dokoro</td>
<td>9777 S Sant</td>
<td>Intimate an</td>
<td>3 Reviews</td>
</tr>
<tr>
<td>Hokusai</td>
<td>8400 Wilshire</td>
<td>Chic elegance</td>
<td>30 Reviews</td>
</tr>
<tr>
<td>Sushi Sasaba</td>
<td>12400 Wilshire A</td>
<td>Authentic Japanese</td>
<td>66 Reviews</td>
</tr>
<tr>
<td>Sushi Roku</td>
<td>8445 W 3rd Ave</td>
<td>High fashion</td>
<td>62 Reviews</td>
</tr>
<tr>
<td>Hide Sushi</td>
<td>2040 Sawtelle</td>
<td>No fuss, just sushi</td>
<td>25 Reviews</td>
</tr>
<tr>
<td>Fat Fish</td>
<td>616 N Robert</td>
<td>Inventive restaurant</td>
<td>38 Reviews</td>
</tr>
<tr>
<td>Sushi Katsu-ya</td>
<td>11680 Venice Ave</td>
<td>The MOCA on</td>
<td>49 Reviews</td>
</tr>
<tr>
<td>Gindi Thai</td>
<td>4017 W Riverside</td>
<td>Burbank restaurant</td>
<td>29 Reviews</td>
</tr>
<tr>
<td>Katana</td>
<td>8439 W Sunset</td>
<td>Rustic Japanese</td>
<td>96 Reviews</td>
</tr>
<tr>
<td>Echigo</td>
<td>12217 Sunset Dr</td>
<td>Stellar sushi</td>
<td>49 Reviews</td>
</tr>
</tbody>
</table>

![Data Integration Tool](image)
Data Integration (cont.)

\[
\{a\}_R = \text{possible new attribute selection for row } i.
\]

\[
\{x\} = \text{Set intersection(}\{a\}\text{) over all the value rows.}
\]

\[
\{v\} = \text{val(a, s) where } a \{x\}
\]

\[
s \text{ is any source where } att(s) \{x\} \neq \{\}
\]
Map Generation

restaurant_name: Hokusai
address: 8400 Wilshire Blvd, Beverly Hills, CA, 90211
description: Chic elegance and modern Zen style surround Japanese French this paean to haute cuisine and stylized sushi.
number_of_reviews: 30
health_rating: 90
Evaluation: Average

Time comparison average over three tasks

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Times (minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>2.22x</td>
</tr>
<tr>
<td>Source Modeling</td>
<td>0.67x</td>
</tr>
<tr>
<td>Cleaning</td>
<td>4.16x</td>
</tr>
<tr>
<td>Integration</td>
<td>6.49x</td>
</tr>
<tr>
<td>Overall</td>
<td>3.32x</td>
</tr>
</tbody>
</table>

Dapper/Pipes  Karma
Discussion

• Contribution: An approach to build Mashups by combining four common information integration techniques into a unified framework.
  – Data extraction
  – Source modeling
  – Data Cleaning
  – Data Integration
Karma Video
Related Work

• Data Extraction
  – Simile [Huynh 2005], Dapper, D.Mix [Hartman 2007], OpenKapow

• Data Cleaning
  – Potter’s Wheel [Raman 2001]

• Manual Mashup Construction
  – Google MyMap

• Manual Integration
  – Intel’s Mashmaker [Ennals 2007]

• Widget Approach to Integration
  – Yahoo’s Pipes, Microsoft’s Popfly, IBM’s QED Wiki, Bungee Labs, Proto Software, Marmite [Wong 2007]

• Programming by Demonstration
  – Programming by Demonstration [Cypher 1993, Lau 2001]
  – Building Queries by Demonstration [Tuchinda 2007]
Conclusion

• Tradeoffs in each approach
  – Manual: Google MyMaps
    • Pro: Easy to define final result
    • Con: Labor intensive
  – Manual Specification: Mashmaker
    • Pro: Flexible, browser-based integration
    • Con: Requires an expert to add new functionality
  – Widget-based Approach: Marmite
    • Pro: Easy integration of capabilities
    • Con: Dataflow model is difficult for users to understand
  – Programming by Demonstration: Karma
    • Pro: Easy for users to specify integration
    • Con: May not work on all web sites