Building Mashups

Craig Knoblock
University of Southern California

Thanks to Rattapoom Tuchinda and
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

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
Outline

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
Existing Approaches

Goal: Create Mashups without Programming

- Addresses syntax issued, 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.

Can we come up with a framework that addresses all of the issues while still making the Mashup building process easy?
Our Integrated 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.
Our system: Karma

Embedded Browser

Table

Interaction Modes
### Data Retrieval: Extraction

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.

---

#### HTML Syntax:
- `<tbody>`
- `<tr>`
- `<td>`
- `<a>`
- `<br>`

**Example:**
- `<tbody><tr><td><a>Japon Bistro</a></td><td>Upscale yet affordable...</td></tr><tr><td><a>Sushi Doro</a></td><td>Intimate and charming...</td></tr></tbody>`

---

```
<table>
<thead>
<tr>
<th>select one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
</tr>
</tbody>
</table>
```

---
Data Retrieval: Navigation

1. **Japon Bistro**
   927 E Colorado Blvd, Pasadena, CA, 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

2. **Sushi Dokoro Ki Ra La**
   9777 S Santa Monica Blvd, Beverly Hills, CA, 90211
   Intimate and charming Japanese restaurant offers wide range of hand-selected sushi and sashimi.

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

4. **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.

5. **Sushi Roku**
   8445 W 3rd St, Los Angeles, CA, 90048
   High fashion, rock and roll and Hollywood buzz converge over innovative sushi.

---

**Japon Bistro**
927 E Colorado Blvd
Pasadena, CA 91106
Phone: (626) 744-1751

31 Reviews | Write a Review

Map | Print | Website | Email to a Friend | SMS/Text

Business Cards

Promote this business
Link to this page
Have Updated Info? Suggest a correction

---

**Select One**
**Address**
**Select One**
**Select One**

---

| Japon Bistro | 927 E Coloro... | Upscale yet ... | 28 Reviews |
| Sushi Dokoro... | 9777 S Sant... | Intimate an... |
| Hokusai | 8400 Wilshir... | Chic eleganc... |
| Sushi Sasab... | 12400 Wilshi... | Authentic Ja... |
| Sushi Roku | 8445 W 3rd ... | High Fashion... |
| Hide Sushi | 2040 Sawtel... | No fuss, jus... |
| Fat Fish | 616 N Rober... | Inventive ro... |
| Sushi Katsu-ya | 11680 Vent... | The MOCA o... |
| Gindi Thai / ... | 4017 W Riv... | Burbank res... |
| Katana | 8439 W Sun... | Rustic Japa... |
| Echigo | 12217 Sant... | Stellar sushi... |
Source Modeling (Attribute selection)

Possible Attribute
\{a | a, s: a ∈ att(s) ∧ (val(a, s) ⊂ V)\} 

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

<table>
<thead>
<tr>
<th>Newly extracted data</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>

<table>
<thead>
<tr>
<th>Data repository</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>restaurant name</strong></td>
</tr>
<tr>
<td><strong>Address</strong></td>
</tr>
<tr>
<td><strong>Health Rating</strong></td>
</tr>
<tr>
<td>Hokusai</td>
</tr>
<tr>
<td>Katana</td>
</tr>
<tr>
<td>Japon Bistro</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zagat</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>restaurant name</strong></td>
</tr>
<tr>
<td><strong>zagat Rating</strong></td>
</tr>
<tr>
<td>Sushi Sasabune</td>
</tr>
<tr>
<td>Sushi Roku</td>
</tr>
<tr>
<td>Katana</td>
</tr>
</tbody>
</table>
Data Cleaning: using predefined rules

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
<th>number of reviews</th>
<th>suggest</th>
<th>user defined</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iora...</td>
<td>Upscale yet...</td>
<td>28 Reviews</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ant...</td>
<td>Intimate and...</td>
<td>3 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shir...</td>
<td>Chic elegance...</td>
<td>30 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ilshi...</td>
<td>Authentic Japan...</td>
<td>66 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd...</td>
<td>High Fashion...</td>
<td>62 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>itell...</td>
<td>No fuss, just...</td>
<td>25 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iber...</td>
<td>Inventive roll...</td>
<td>38 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intu...</td>
<td>The MOCA o...</td>
<td>49 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rive...</td>
<td>Burbank rest...</td>
<td>29 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sun...</td>
<td>Rustic Japan...</td>
<td>96 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>santa...</td>
<td>Stellar sushi ...</td>
<td>49 Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

28 Reviews → 28

Subset Rule:
\((s_1s_2...s_k) \rightarrow (d_1d_2...d_t) \land (k \leq t) \land s_i \in \{d_1,d_2,...,d_t\} \land d_i \not= d_j\)
Data Integration
Based on [tuchinda 2007]
Data Integration (cont.)

<table>
<thead>
<tr>
<th>restaurant</th>
<th>address</th>
<th>description</th>
<th>number of...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
<td>927 E Color...</td>
<td>Upscale yet...</td>
<td>28 Reviews</td>
</tr>
<tr>
<td>Sushi Dokor</td>
<td>9777 S Sant...</td>
<td>Intimate an...</td>
<td>3 Reviews</td>
</tr>
<tr>
<td>Hokusai</td>
<td>8400 Wilsh...</td>
<td>Chic eleganc...</td>
<td>30 Reviews</td>
</tr>
<tr>
<td>Sushi Sasab...</td>
<td>12400 Wilshi...</td>
<td>Authentic Ja...</td>
<td>66 Reviews</td>
</tr>
<tr>
<td>Sushi Roku</td>
<td>8445 W 3rd...</td>
<td>High fashion...</td>
<td>62 Reviews</td>
</tr>
<tr>
<td>Hide Sushi</td>
<td>2040 Sawtel...</td>
<td>No fuss, jus...</td>
<td>25 Reviews</td>
</tr>
<tr>
<td>Fat Fish</td>
<td>616 N Rober...</td>
<td>Inventive ro...</td>
<td>38 Reviews</td>
</tr>
<tr>
<td>Sushi Katsu-ya</td>
<td>11680 Vent...</td>
<td>The MOCA o...</td>
<td>49 Reviews</td>
</tr>
<tr>
<td>Gindi Thai / ...</td>
<td>4017 W Riv...</td>
<td>Burbank res...</td>
<td>29 Reviews</td>
</tr>
<tr>
<td>Katana</td>
<td>8439 W Sun...</td>
<td>Rustic Japa...</td>
<td>96 Reviews</td>
</tr>
<tr>
<td>Echigo</td>
<td>12217 Sant...</td>
<td>Stellar sushi ...</td>
<td>49 Reviews</td>
</tr>
</tbody>
</table>

LA Health Rating

<table>
<thead>
<tr>
<th>restaurant</th>
<th>Address</th>
<th>Health Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hokusai</td>
<td>8400</td>
<td>90</td>
</tr>
<tr>
<td>Katana</td>
<td>8439.</td>
<td>99</td>
</tr>
<tr>
<td>Japon Bistro</td>
<td>927 E..</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</td>
<td>27</td>
</tr>
<tr>
<td>Sasabune</td>
<td></td>
</tr>
<tr>
<td>Sushi Roku</td>
<td>25</td>
</tr>
<tr>
<td>Katana</td>
<td>23</td>
</tr>
</tbody>
</table>

\{a\}_R = possible new attribute selection for row \(i\).
\{x\} = Set intersection(\{a\}) over all the value rows.
\{v\} = val(\(a, s\)) where \(a \{x\}\)
s is any source where att(s) \{x\} ≠ {}
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
Related Work

• Building Blocks
  – Programming by Demonstration [Cypher 1993, Lau 2001]
  – Simile [Huynh 2005]
  – Potter’s Wheel [Raman 2001]
  – Building Queries by Demonstration [Tuchinda 2007]

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

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

• Misc.
  – Intel’s Mashmaker [Ennals 2007]
  – Google MyMap
# Preliminary Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Data Retrieval</th>
<th>Source Modeling</th>
<th>Data Cleaning</th>
<th>Data Integration</th>
<th>Total Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task1 K</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Task1 DP</td>
<td>8</td>
<td>10</td>
<td>21</td>
<td>9</td>
<td>48</td>
</tr>
<tr>
<td>Task2 K</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Task2 DP</td>
<td>18</td>
<td>30</td>
<td>0</td>
<td>28</td>
<td>76</td>
</tr>
<tr>
<td>Task3 K</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>Task3 DP</td>
<td>8</td>
<td>11</td>
<td>16</td>
<td>12</td>
<td>47</td>
</tr>
</tbody>
</table>

- 3 Mashup building tasks
- Compare Karma (K) and Dapper + Yahoo (DP).
- Record the number of steps:
  - typing values in textbox, clicking a button, select an option from a list, dragging and dropping, and connecting one widget to one another
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
Outline

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
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

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 (1 spreadsheet, 2 programmers) complete the 4 tasks in one hour.
  – Novice 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

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
Intel Mashmaker

• Multi-tier user
  – Naïve users
  – Expert users

• Experts do all the dirty works 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.
Conclusion

• Tradeoffs in each approach
  – Karma
    • Pro: end-to-end approach based on a programming by demonstration paradigm
    • Con: May not work on all web sites
  – Marmite
    • Pro: Easy integration of capabilities
    • Con: Limited types of integration
  – Mashmaker
    • Pro: Browser-based integration
    • Con: Requires an expert to add new functionality