Mashups

Jeff Wong
Human-Computer Interaction Institute
Carnegie Mellon University

jeffwong@cmu.edu

Who am I?

• 2nd Year PhD Student, HCII @ CMU
  – Research: mashups and end-user programming (EUP)
  – Creator of Marmite

• Masters HCI, Carnegie Mellon
• Computer Science, UC Berkeley (2000)

• Former millionaire
Last Time

• Tomek introduces you to Web technologies (Web 2.0)

• DOM
• CSS (don’t)
• Javascript
• AJAX
  – XMLHttpRequest
  – XML (avoid if possible)
  – JSON

This Time

• Jeff shows you mashups and how to make your own.

• Mashup Patterns
• Getting other People’s stuff
  – Scraping
  – RSS
  – Using APIs

• Mashup Architecture & Tools
• Business Models
• The Future
Mashup Patterns

What Are Mashups?

• “A mashup is a web application that combines data from more than one source into a single integrated tool”  
  - Wikipedia

• A good start, but not a complete definition, better to show examples
Craigslist Rentals + Google Maps =

- $475 / 2br - 2 Bedroom Apartment Available Immediately (Belleview)
- $450 / 1br - Beautiful 1BR apartment available from 16th NOV, 2007 (Booms Ave)
- $750 / 3br - 3 bedroom townhouse (Mountains)
- $850 / 2br - LIVE 2 BEDROOM APT DOWNTOWN GREENSBURG [[ VIEW ]] LOADS OF CHARACTER! (Greenbush)
- $1150 / 3br - Shawnee Townhouse unique rooms left (Kentucky near Noger) (pic)
- $295 / 2br - Great FOX CHAPEL 2br Available NOW! (Gypsy Run Rd)
- $450 / 1br - Bellvue 1 BR Available for Jan 1st (Pearl Ave)
- $650 / 1br - Squirrel Hill 1BR ALL UTILITIES INCLUDED! (Many)
- $475 Large Efficiency! Available Jan 1st! (Adams)

5000 Forbes Ave
Pittsburgh, PA 15213
Make this my default location
Pattern from HousingMaps

- HousingMaps is
  - Craigslist: geographic data
  - Map as a widget
  - Metadata extraction
  - Filtering
  - New index for existing “leaves”
- Visualization is cheap
  - Google Maps is overdone
    - Google MyMaps today
  - Let’s-All-Do-Same-Thing
    - Flickr
    - Twitter
Patterns in Mashups

• by me and my Advisor
• accepted to Workshop for End-User Software Engineering
• Understand patterns...
  – tool design
  – make things easy for certain kinds
  – suggest new avenues for mashup innovation

Patterns in Mashups

• Method
  – Look at lists of good mashups
    • Programmable Web Mashup of the Day
    • MashupAwards.com (every day)
  – Describe how data sources are used
  – Flag attributes
  – “Guess” at user needs

  – Make a tables of attributes
    • toss bad mashups (50%)
Aggregation

50 MATCHES

Enter search term: Go

We only crawl web sites that were bookmarked or voted for by people, in sites like del.icio.us, digg and reddit. They link to sites with the best content, and it gives you the best quality search engine on the web. You can also call it Social Powered Search. More about us »

Alternate UI

InSuggest

Link to this selection

SUGGESTIONS  SEARCH  RANDOM

Golden Gate Bridge

Link to this selection

Link to this selection

Link to this selection

Link to this selection
Personalization

![Year End Summary](image)

**Focused-View of the Data**

*lazylibrary*

Aesthetics

**read less. get more.**

Ever read a book that was a few hundred pages longer than it needed to be? Yeah, so have we. Fortunately, there are authors out there that would rather have a concise and efficient book than a lengthy and dilated tome, and that’s where we come in.

Welcome to the LazyLibrary, where you can find books on any topic without having to worry about high page counts. If it's over 200 pages, you won't even see it. Read all about anything, in less time, for (usually) less money.
Real-Time Monitoring

Patterns found

- Aggregation
- Alternate UI & In-situ use
- Personalization
- Focused View of Data
- Real-time Monitoring

- others??? (maybe some obvious ones here)
Other sources for inspiration

- Greasemonkey scripts
- Firefox plugins
- Explore APIs!
  - Helpful to do this in a group
  - Report back and brainstorm together
  - AVOID popular APIs
    - so many "lonely" APIs
- Your own personal experience!
  - Long tail of mashups
    - Hacking, Mashing, & Gluing (Hartmann, Doorley & Klemmer)

Find More Mashups

- ProgrammableWeb
  - Lots of crap here
  - Use the search form to get all mashups
Other mashup possibilities

• Government
  – Census data
  – Democracy
  – Political discourse
• Use Wikipedia
  – Some APIs out there
• Mixed initiative mashups
  – Humans can sort, rate, and comment
  – Re-categorize
• Apply novel algorithms!
  – Data-mining or semantic stuff

How to re-use other people’s stuff (APIs and more)
Re-using other people’s stuff

- Screen-scraping
- RSS Feeds
- APIs
  - Web Services (WSDL + SOAP)
  - REST
  - Wrappers for favorite programming language
    - Java, Python, Perl, PHP, Flash
- Prototype in an open, interactive environment

Screen-scraping

- Basically: visit web pages, extract information
- Do this when there is no API
- Variety of techniques
  - Query languages
  - Demonstrational interfaces
    - Dapper
- BUT
  - Slow
  - Unstable
Query Languages

• Need to inspect page’s DOM structure
• Need to test interactively

• XPath (yuck)
  – Powerful, but complicated
• BeautifulSoup (Python)
• JQuery (Javascript)

• XML parsers not so helpful

Demonstrational Interfaces

• Dapper is the best
  – Works on web pages where DOM structure is regular
  – Creates a scraper that takes a URL and turns the web page into XML, CSV, RSS, etc
• Some semantic scrapers exist
  – Mostly research prototypes
  – MIT Sifter
RSS Feeds

- XML for news articles
- Also supports fields and tags for names
- RSS 2.0 is simpler than RSS 1.1

- A feed is a list of items
- News oriented tags: title, description, date, etc

- Services often use extensions to insert custom fields

What are API’s (on the Web)

- Libraries for programming
  - Java
    - draw a box
    - sort this list
    - the list

- API’s on the web
  - Upcoming.org: search for events (in an area), lookup venue
  - Google: search for something
  - Rent-o-meter: add my listing to your database
How to use an API

- Find an API worth using
  - involves prototyping
- Get a developer key (sometimes)
- Read the documentation
- Explore how the API will work
  - Pick a prototyping language or method
    - Ex: Javascript, Python, Java, your browser, PHP, etc
  - Establish a debugging process
  - Avoid busy work or waterfall process
  - Keep in mind architectural decisions while you are testing
    - performance may be important

Web Services vs REST

- Web Services
  - Service-Oriented Architecture (SOA) or Web Services
    - WSDL
      - Web-Services Description Language
    - SOAP
      - Simple-Object Access Protocol
      - Simple In Theory
    - You download a WSDL file, generate function call stubs (using tools), call your web services
Web Service-based Weather Service

- Define some methods in WSDL file
  - `getWeather(country, state, city)`
    - Define to be accessible via http on certain url
- Return XML-formatted content

- To get latest weather, create a program that
  POSTs the right SOAP call to the URL
  - Works for programs only

Web Services vs. REST

- Web Services are complicated
  - WSDL is complicated to make
  - SOAP is hard to implement on providers end
  - High-barrier to entry for simple services
  - User needs to locate a parser and stub generator
REST to the rescue!

- REST (Representational State Transfer) related idea
  - Philosophy of how web works
- Main ideas:
  - Few verbs, many nouns, applied uniformly to all data
    - GET, POST, PUT, DELETE
    - Standard HTTP methods
  - URLs are implicit API
  - Minimal state maintained between transactions
- You can test it by entering URL in your browser!

REST-based Weather Service

- Rely exclusively on pre-defined methods
  - GET, POST, PUT, DELETE
- Define URLs
  - http://weather.com/usa/pennsylvania/pittsburgh
  - http://weather.com/usa/georgia/atlanta
- Return content in some form
  - HTML web page or XML-formatted
- To get the latest weather, just access the URL
  - Works for people, programs
REST Drawbacks

• Inspectable, simple for providers, low overhead BUT

• Hard to deal with a lot of different calls
• Have to read documentation to figure out options
• Calls don’t look like function calls
  1. URL formation
  2. Make asynchronous call, setup callback
  3. Setup post-processing of results

• Possible solution: WADL (WSDL for REST)
  — generate using REST Describe and Compile

Wrappers

• Service providers make a library to access their service in your programming language

• Usually big providers
  – Amazon, eBay, Google, Flickr
Interesting APIs

• Mindmeister
  – Maybe with del.icio.us
• Remember the Milk
• BaseCamp
  – Project management
• 43Things
  – Social goal setting
• Ning
  – Community Web Services

API exploration advice

• Use an interactive programming language
  – Javascript (in shell)
  – Python
  – Ruby?

• Look for developer API exploration tools
  – Tend to be interactive
  – Similar to above but usually tailored

• Use Firefox or Internet Explorer to pick at REST results
Jeff’s Tool Demo

• Hope this works!

Go beyond APIs

• Interesting Data is not always behind APIs
  – Government and economic data
  – Environmental and scientific data
    • Fish Data + Recipes: EcoEats
    • HeyWhatsThat
  – Swivel

• Ask and ye *may* receive
  – Sometimes you get proprietary data

• Boundaries are always expanding
  – Semantic
Mashup Architecture & Tools

Dissecting a Mashup

- HousingMaps.com
EcoEats architecture

- OceansAlive.org
  - Python
  - BeautifulSoup scraper
- MealMaster Recipes
  - Perl parsing
  - Python injection
- SQL Database
  - Zend+PHP
- user

No-cross site scripting

- Security model of Javascript
  - “Same origin” policy
    - Can only talk to the server where your page comes from
  - Like Java applets
  - This means you need to setup a proxy
No-cross site scripting

• BUT
  – If you sign your scripts this might work
  – JQuery Javascript library supports this I think
  – IANASE

Mashup Tools

• Programming mashups quite hard
  – HTML, text parsing, regular expressions, databases, integration with other web services, etc

• Many tools recently created:
  – Marmite, Dapper, Intel MashMaker
  – Microsoft Popfly, Yahoo Pipes
Marmite

- Goal: Make it easy to extract data from sites and then visualize as desired
  - HousingMaps.com
Extracting Generic Content

- Can’t have pre-defined extractor operators for every possible web site
  - Need a more general way of extracting data from pages

- Developed a generic wizard UI for selecting links
  - Content from that set could be extracted via other operators
  - Uses Solvent (MIT), an XPath-based algorithm for finding patterns in web pages
    - Finds “groups” of related web content based on how HTML is structured
Using Marmite (Envisioned)

- Extract content from one or more web pages
  - names, addresses, dates, phone #, URLs
- Process it in a data flow manner
  - filtering out values or adding metadata
  - integrating with other data sources (similar to a database join operation)
- Direct the output to a variety of sinks
  - databases, map services, text files, visualizations, web pages, or source code that can be further edited

Marmite

- No way to integrate into an automatic process
- Not supported (by Me)
- Hard to explore Web Services
- No supporting tools or community
Yahoo Pipes

Microsoft Popfly

- Blocks of functionality shared in a programmer community
- Web Services you want might already be coded!
- Integrated environment at different levels of code
- Javascript (YAY)
- Probably your best bet for not having to build something from scratch
  - Or setup a web server, environment (PITA)
Spinning 3D translucency!
Business Models
How to succeed with mashups

- Making money is nice
  - Get bought
  - Make affiliate money
  - Drive advertising (sketchy)

- Get hired
  - Make something cool by reusing something
  - “Innovation is invention in context”

- Make something cool
  - Once people use it, you can then decide what success means
  - Requires some PR and polish

The Future
The Future

• Semantic Web
  – Not my expertise.... can some speak to this?

• Facebook?
  – Advantages: ???
  – Disadvantages: ???

Thanks!

Jeff Wong
Human-Computer Interaction Institute
Carnegie Mellon University
jeffwong@cmu.edu