



Making the Web searchable

Towards Semantic Search

Peter Mika
Researcher, Data Architect
Yahoo! Research

Previously in search

- Horizontal search
 - Yahoo...
 - Keyword-based indexing
 - Minimal natural language processing
 - Limited experiments with ontologies (query expansion)
- Vertical search
 - e.g. shopping.com, Kelkoo
 - Faceted search, browsing
 - Fixed ontology
- Combinations
 - Google Base, Google Co-op
 - Web-scale, but fixed ontologies
 - Proprietary technology
- Can we do better with the Semantic Web?
 - Address the long tail of queries (88% of queries)
 - Use standard technology
- Not a new question. But the answer may be new.



Which Semantic Web?

- Two visions
 - Data Web
 - Bringing the content of databases to the Web (linkeddata.org)
 - Rich data, heavyweight semantics
 - Deep Web
 - Annotated Web
 - Annotating the content of Web resources (documents, mm)
 - Simple data, lightweight semantics
 - Shallow Web
- This presentation is about the Annotated Web.

- 3 -

Updated: 19-May-09



Brief history of the Annotated Web

- 1995: HTML meta tags
- 1996: Simple HTML Ontology Extensions (SHOE)
- 1998: RDF/XML
 - RDF/XML in HTML
 - RDF linked from HTML
- 2003: Web 2.0
 - Tagging
 - Microformats
 - Metadata in Wikipedia
 - Machine tags in Flickr
- 2005: eRDF
- 2008: RDFa



- 4 -

Updated: 19-May-09



HTML meta tags

```
<HTML>
<HEAD profile="http://dublincore.org/documents/dc9-html/">
<META name="DC.author" content="Peter Mika">
<LINK rel="DC.rights copyright"
      href="http://www.example.org/rights.html" />
<LINK rel="meta" type="application/rdf+xml" title="FOAF"
      href= "http://www.cs.vu.nl/~pmika/foaf.rdf">
</HEAD>
...
</HTML>
```

- 5 -

Updated: 19-May-09



SHOE example (Hefflin & Hendler, 1996)

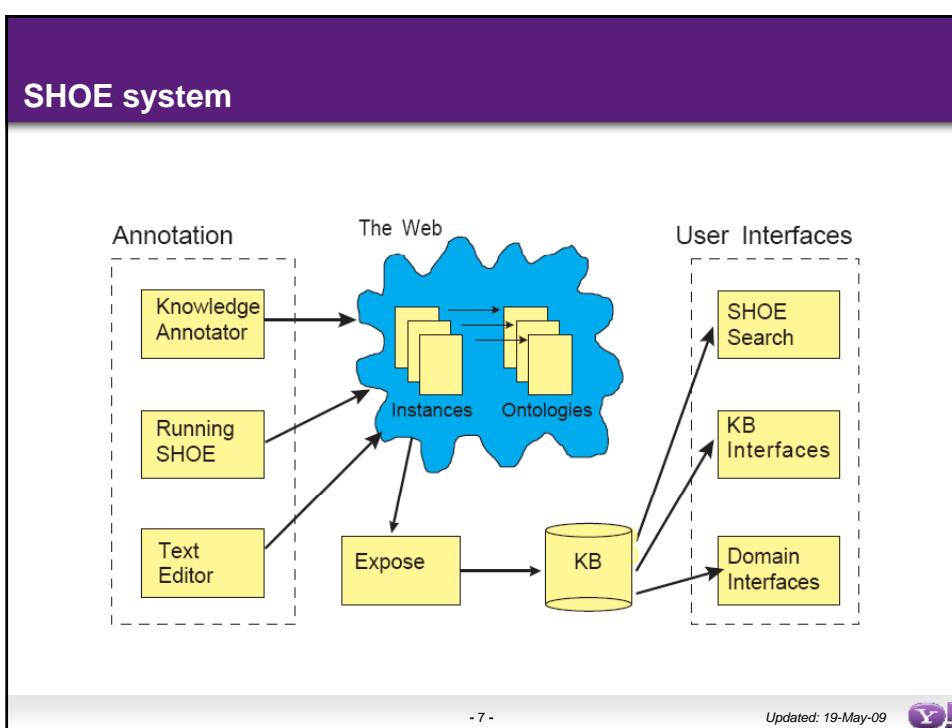
```
<ONTOLOGY "our-ontology" VERSION="1.0">
<ONTOLOGY-EXTENDS "organization-ontology" VERSION="2.1" PREFIX="org"
  URL="http://www.ont.org/orgont.html">
<ONTDEF CATEGORY="Person" ISA="org.Thing">
<ONTDEF RELATION="lastName" ARGS="Person STRING">
<ONTDEF RELATION="firstName" ARGS="Person STRING">
<ONTDEF RELATION="marriedTo" ARGS="Person Person">
<ONTDEF RELATION="employee" ARGS="org.Organization Person">
</ONTOLOGY>

<HEAD>
<META HTTP-EQUIV="Instance-Key" CONTENT="http://www.cs.umd.edu/~george">
<USE-ONTOLOGY "our-ontology" VERSION="1.0" PREFIX="our" URL="http://ont.org/our-ont.html">
</HEAD>
<BODY>
<CATEGORY "our.Person">
<RELATION "our.marriedTo" TO="http://www.cs.umd.edu/~helena">
<RELATION "our.employee" FROM="http://www.cs.umd.edu">
My name is
<ATTRIBUTE "our.firstName"> George </ATTRIBUTE>
<ATTRIBUTE "our.lastName"> Cook </ATTRIBUTE> and I live at...
```

- 6 -

Updated: 19-May-09





SHOE Text-based query interface

The screenshot shows the SHOE Text-based query interface. The top menu bar includes File, Help, and a dropdown for Ontology (set to "cs-dept-ontology, v.1.0"). The main window has two panes: a left pane for selecting a category and a right pane for entering a query.

Left Pane (Select a category):

- AssociateProfessor
- Chair
- Dean
- FullProfessor
- VisitingProfessor
- Publication
- Article** (selected)
- ConferencePaper
- JournalArticle
- TechnicalReport

Right Pane (Query):

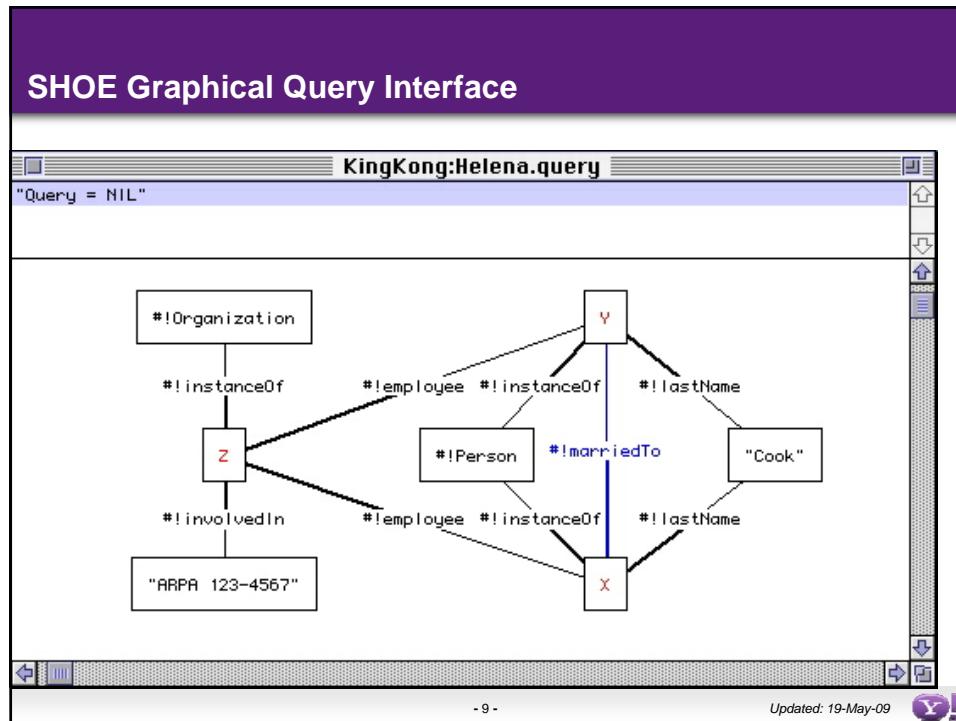
Subject	Category	Value
publicationResearch	Simple HTML Ontology Extension	
publicationAuthor	Jeff Heffler	
publicationDate		
name		
softwareDocumentationOf		
orgPubl icationOf		

Bottom Table (Results):

Article	publicationResearch(name)	publicationAuthor(name)	name
http://www.1ht	Simple HTML Ontology Extension	Jeff Heffler	Coping with Changing Ontologies in a
http://www.2ht	Simple HTML Ontology Extension	Jeff Heffler	Applying Ontology to the Web: A Case
http://www.3ht	Simple HTML Ontology Extension	Jeff Heffler	Reading Between the Lines: Using SH

Buttons at the bottom include Web Search, Get Page, Add To Query, and 3 answers.

- 8 - Updated: 19-May-09



Example: Creative Commons

Embedding CC license in HTML (now deprecated):

```
<HTML>
<HEAD>... </HEAD>
<BODY>
...
<!--
<rdf:RDF xmlns="http://creativecommons.org/ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<Work rdf:about="http://www.yergler.net/averages/">
<dc:title>The Law of Averages</dc:title>
<dc:description>...because eventually i'll be right...</dc:description>
<license rdf:resource="http://creativecommons.org/licenses/by-nc/1.0/" />
</Work>
<License rdf:about="http://creativecommons.org/licenses/by-nc/1.0/">
<requires rdf:resource="http://web.resource.org/cc/Notice" />
<permits rdf:resource="http://web.resource.org/cc/Reproduction" />
<permits rdf:resource="http://web.resource.org/cc/Distribution" />
<prohibits rdf:resource="http://web.resource.org/cc/CommercialUse" />
</License>
</rdf:RDF>
-->
```

- 10 - Updated: 19-May-09

Example: Creative Commons

- Current: rel attribute (HTML4)

This work is licensed under a [a Creative Commons Attribution 3.0 United States License.](http://creativecommons.org/licenses/by/3.0/us/)

- Use of the “rel” attribute for semantic annotation is the birth of the microformat...

- 11 -

Updated: 19-May-09



Example: microformats

```
<div class="vcard">
  <a class="email fn" href="mailto:jfriday@host.com">Joe Friday</a>
  <div class="tel">+1-919-555-7878</div>
  <div class="title">Area Administrator, Assistant</div>
</div>
```

```
<cite class="vcard">
<a class="fn url" rel="friend colleague met"
  href="http://meyerweb.com/">Eric Meyer</a>
</cite> wrote a post (<cite>
<a href="http://meyerweb.com/eric/thoughts/2005/12/16/tax-relief/">
  Tax Relief</a></cite>) about an unintentionally humorous letter
  he received from the <span class="vcard">
<a class="fn org url" href="http://irs.gov/">
  Internal Revenue Service</a> </span>.
```

- 12 -

Updated: 19-May-09

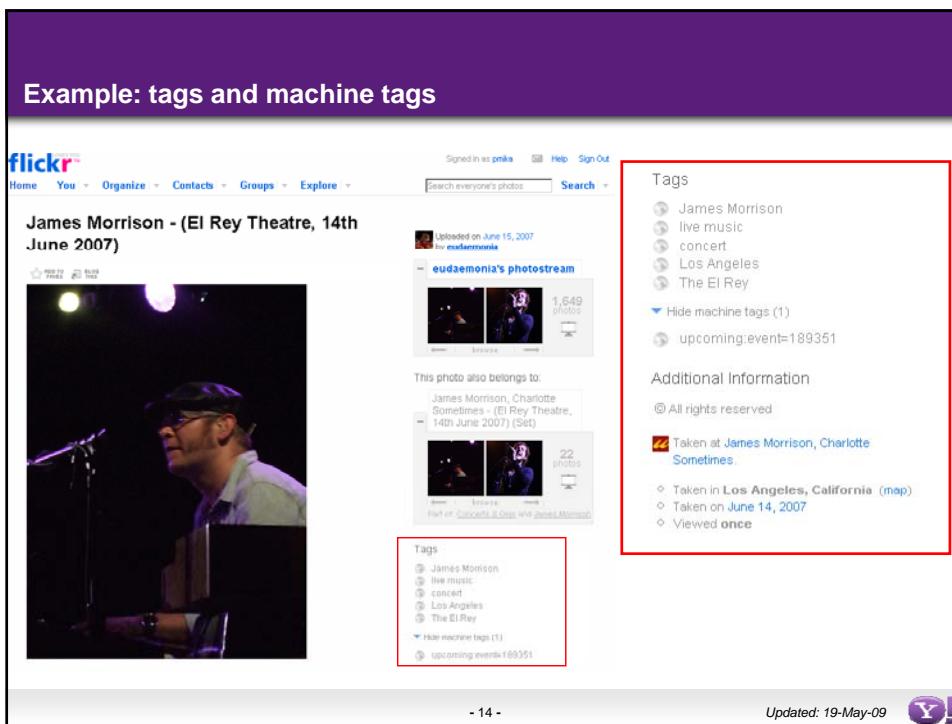


microformats

- microformats.org
- Originated by Tantek Celik and others
- Agreements on the way to encode certain kinds metadata in HTML
 - Reuse of semantic-bearing HTML elements
 - Based on existing standards
 - Community process
 - Persons, events, listings etc. but also syntactic metadata: licenses, tags
- Microformats have no shared syntax
 - Each microformat has a separate syntax tailored to the vocabulary
- Microformats are not ontologies
 - No formal descriptions of schema, only text
 - Limited reuse, extensibility of schemas
 - No datatypes
- No namespaces, unique identifiers (URIs)
 - no interlinking
 - mapping between instances is required
- Relationship to page context is unclear
- Widely used in millions of documents
 - User-generated as well as automatically generated

- 13 - Updated: 19-May-09 

Example: tags and machine tags



Tags

- James Morrison
- live music
- concert
- Los Angeles
- The El Rey

Additional Information

- All rights reserved
- Taken at James Morrison, Charlotte Sometimes.
- Taken in Los Angeles, California (map)
- Taken on June 14, 2007
- Viewed once

- 14 - Updated: 19-May-09 

Example: Tags and machine tags

- Tags
 - User defined keywords
 - Minimal agreement
 - Is ‘rock’ on Flickr same as ‘rock’ on myspace?
 - Is ‘rock’ by me on Flickr is the same as ‘rock’ by you on Flickr?
 - Is ‘rock’ by me on Flickr today the same as ‘rock’ by me on myspace tomorrow?
- Machine tags
 - User defined values for user defined properties
 - Possibility to define the namespace (but not enforced)
 - Limited use

- 15 -

Updated: 19-May-09



RDF-based annotation #1: eRDF

- eRDF
 - Ian Davis (Talis)
 - Embedding RDF in HTML
 - Straightforward mapping to RDF triples (XSLT available)
 - HTML4 compatible
 - More complex than microformats
 - Use any RDF/OWL vocabulary
 - Reuse of semantic-bearing HTML elements is limited
 - More limited than RDF
 - No blank nodes
 - No data types
 - No statements about subjects other than the current document
 - Limited usage

- 16 -

Updated: 19-May-09



RDF-based annotation #2: RDFa

- RDFa
 - World Wide Web Consortium (W3C) last call document
 - Similar intent as eRDF, but full RDF support
 - Requires XHTML
 - Big question: user complexity (→ data quality)

```
<p typeof="contact:info" about="http://example.org/staff/jo">
<span property="contact:fn">Jo Smith</span>.
<span property="contact:title">Web hacker</span> at
<a rel="contact:org" href="http://example.org"> Example.org </a>.
  You can contact me <a rel="contact:email"
  href="mailto:jo@example.org">
  via email </a>.
</p> ...
```

- 17 -

Updated: 19-May-09



Metadata is out there

- Question:
 - Just how much data is out there?
 - What is the quality?
- Idea: bring metadata to the surface of search
- How does it work?
 - User enters query
 - Metadata is extracted dynamically
 - Entity reconciliation
 - Metadata is used to display
 - rich abstracts,
 - related pages
 - spatial, temporal visualization
- Microsearch prototype
 - Play at <http://www.yr-bcn.es/demos/microsearch/>

- 18 -

Updated: 19-May-09



Example: ivan herman

The screenshot shows a Yahoo search results page for the query "ivan herman". The results include:

- Rich abstract:** A summary of the first result, "Ivan's blog", which links to <http://www.ivan-herman.net/>.
- Related pages based on metadata:** A list of related pages including "Ivan Herman's Photos" and "Ivan Herman".
- Geolocation:** A map showing the location of "Herman O" in a specific area.
- Events from personal calendar, Conferences, and bio from LinkedIn:** A calendar view showing events like "WW2008" and "Robo's BD".

Annotations highlight the "Geolocation" and "Events from personal calendar, Conferences, and bio from LinkedIn" sections.

Example: peter site:flickr.com

The screenshot shows a Yahoo search results page for the query "peter site:flickr.com". The results include:

- Flickr users named "Peter" by geography:** A world map showing numerous orange markers representing users named "Peter" located across different continents.

Annotations highlight the "Flickr users named "Peter" by geography" section.

Example: san francisco conference

Search Results

1. [San Francisco Conference](#) - Encyclopaedia Britannica
2. [Oracle OpenWorld - San Francisco](#)
3. [ad:tech Interactive Media Conference in San Francisco Events](#)
4. [Only in San Francisco - The Official Visitors Site for San Francisco](#)
5. [Gibane San Francisco 2008](#)
6. [South San Francisco Conference Center](#)
7. [5th Annual Conference on Arteriosclerosis, Thrombosis and Vascular Biology](#)

Persons: 20 Events: 135 Unfinished: 1031 | 1 - 10 of about 2725643 for san francisco conference - 10 sec. (About this query)

Map Hyb Sat

Imagery 2006-early 2007 @media 2007 America
Workshop: Transcending CSS @media 2007 Europe
Open Source Conferenc
phony Future of Web Design Europe

Conferences in San Francisco by date

- 21 - Updated: 19-May-09

Example: greater st. peter

Search Results

1. [Greater St. Peter @ River's Edge](#)
2. [Greater St. Peter Missionary Baptist Church](#)
3. [Greater Saint Peter](#)
4. [St Peter's RC Sch](#)
5. [Jobs at The YMCA](#)
6. [St. Peter's Episcopal](#)
7. [St. Louis Honda de](#)
8. [Greater St. Peter S](#)

Name: Senior Pastor: Isaiah J. Waddy, M. Div.
Address: 9540 Fayetteville Road, Jonesboro,
GA 30229-4100
tel:+1-770-210-4100
Add to your AddressBook

Call phone number

Save to address book

(other actions)

- 22 - Updated: 19-May-09

Lessons

- More metadata than we expected
 - 53% of unique queries have at least one metadata-enabled page in top 10 (n=7848)
- Performance is poor
 - Metadata needs to come from the index for performance
- Metacrap does exist
 - Users have to see metadata to spot mistakes in their markup, warn others
- RDF templating is hard
 - Adds extra complexity
- Scalability

- 23 -

Updated: 19-May-09



SearchMonkey



- Creating an ecosystem of publishers, developers and end-users
 - Motivating and helping publishers to implement semantic annotation
 - Providing tools for developers to create compelling applications
 - Focusing on end-user experience
- Rich abstracts as a first application
- Addressing the long tail of query and content production
- Standard Semantic Web technology
 - dataRSS = Atom + RDFa
 - Industry standard vocabularies

- 24 -

Updated: 19-May-09



What is SearchMonkey?



an open platform for using structured data to build more useful and relevant search results

Before

Topics for Getting Pregnant - BabyCenter
Find out how to boost your chances of getting pregnant, what you can do if you're having a problem conceiving, and more ... do before you try to get pregnant ...
www.babycenter.com/getting-pregnant • 91k • Cached

WebMD Allergies Health Center - Find allergy information and latest...
Information and articles on the diagnosis, symptoms, treatment, and the prevention of allergies
www.webmd.com/diseases_and_conditions/allergies.htm • 122k • Cached

Italy Travel Guide and Travel Information - Lonely Planet
Lonely Planet Italy includes information on events, attractions, activities, and transportation for the independent traveler.
www.lonelyplanet.com/worldguide/destinations/europe/italy • 57k • Cached

After

Topics for Getting Pregnant - BabyCenter
BabyCenter's getting pregnant tools and information can boost your chances of conception by helping you understand your cervical mucus, and pinpoint ovulation.
www.babycenter.com/getting-pregnant • 76k • Cached

WebMD Allergies Health Center - Find allergy information and latest...
Allergies are an abnormal reaction of the immune system. People with allergies have an overactive system that reacts to a usually harmless substance in the environment. This substance (pollen, mold, dander, etc.) is called an allergen.
www.webmd.com/allergies • 111k • Cached

Italy Travel Guide: Overview - Lonely Planet WorldGuide
From 10 to 12, Italy is at its best in spring (April-May) and autumn (October-November). During these seasons, the scenery is beautiful, the temperatures are pleasant and are relatively few crowds.
www.lonelyplanet.com/worldguide/destinations/europe/italy • 40k • Cached

- 25 - Updated: 19-May-09 

Enhanced Result



deep links [points to a link labeled "The Art of Pizza - Chicago, IL - Citysearch.com"]

image [points to a thumbnail image of a pizza from the Citysearch result]

name/value pairs or abstract [points to the star rating and review count from the Citysearch result]

1 - 10 of about 8,140,000 for art of pizza chicago (About this page) - 0.23 sec.

Sponsor Results

Chicago Pizza Chicago
Pizza New - Cast Today Chicago Style, World Class Taste.
dineonline.com/ctech/pizza

Nick & Bruno's Pizzeria
Chicago style pizza restaurant
Home of the Monday Pizza
Nick-Brunos.com

Pizza Place in Chicago
Find & review local restaurants and cafes. Free search.
Chicago.Citysearch.com

Pizza Art
Find Pizza Art and Compare prices at Smarter.com.
www.smarter.com

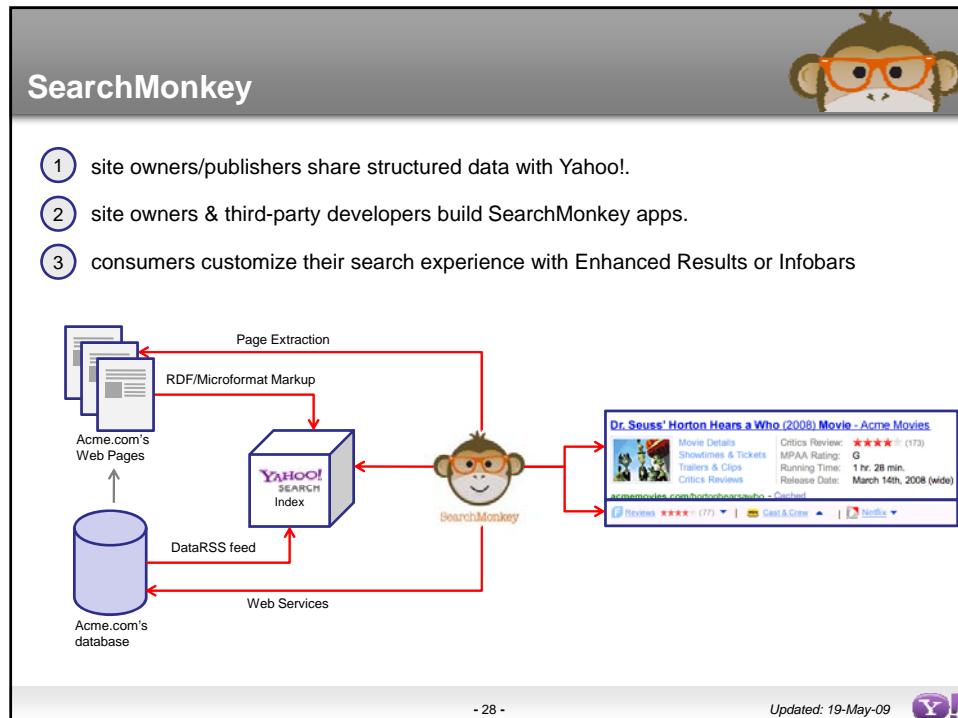
See your message here...

YAHOO!

- 26 - Updated: 19-May-09 

Infobar

YAHOO! CONFIDENTIAL | 27 - 27 - Updated: 19-May-09



Developer tool

Yahoo! | My Yahoo! | Mail | More ▾ Make Y! Your Home Page Welcome, Peter ▾ | Sign Out | Help

YAHOO! APPLICATION PLATFORM SearchMonkey WEB SEARCH

Application Dashboard



Presentation Applications

Presentation Applications are small PHP apps that display enhanced search results using data services. You can use an existing data service or create a custom service below.

Create a New Application | Start with a sample application | Import application

LinkedIn: Presentation Application (SAMPLE)

A sample presentation application for LinkedIn. Displays hResume data directly from profile pages, using a previously-created sample LinkedIn custom data service.

Created: May 20th, 2008, 3:49pm

[Edit](#) [Delete](#) [Export](#)

Custom Data Services

Data Services provide structured information to display in Yahoo! Search results. You can create a custom data service by calling an API or by extracting structured data from web pages.

Create a new data service | Start with a sample data service | Import data service

- 29 - Updated: 19-May-09 

Developer tool

Yahoo! | My Yahoo! | Mail | More ▾ Make Y! Your Home Page Welcome, Peter ▾ | Sign Out | Help

YAHOO! APPLICATION PLATFORM SearchMonkey WEB SEARCH

Application Dashboard / LinkedIn: Presentation Application (SAMPLE)

Basic Info URLs Data Services Appearance Confirmation

Step 1: Specify your application's name, type, and description.

Name: The external name of your app. (* Required)

Type: Enhanced Result (reconfigures a result) Infobar (provides more info under a result)

Acme Movies Dr. Seuss' Horton Hears a Who (2008) Movie Info
 **Dr. Seuss' Horton Hears a Who - Acme Movies**
 Movie Details | Cast & Crew | Storyline | Tickets | Trailers & Clips | Critics' Reviews
 MPAA Rating: PG | Running Time: 1 hr. 28 min. | Release Date: March 14th, 2008 (wide)
movies.acme.com/movie18009736093/info • 248k + Cached

Description:
 A sample presentation application for LinkedIn. Displays hResume data directly from profile pages, using a previously-created sample LinkedIn custom data service.

Category:

- 30 - Updated: 19-May-09 

Developer tool



Yahoo! | My Yahoo! | Mail | More ▾ Make Y! Your Home Page Welcome, Peter ▾ Sign Out | Help

YAHOO! APPLICATION PLATFORM SearchMonkey

Search WEB SEARCH

Application Dashboard / LinkedIn: Presentation Application (SAMPLE)

Basic Info URLs Data Services Appearance Confirmation

Step 3: Choose your data services.

Data Service

Select one or more Data Services to supply your presentation application with interesting information to display. For example, the default Yahoo! Index contains data gathered by the Yahoo! Search Crawler, such as the page's title and file size. You will specify how to display this data on the next screen.

Default: Available from the Yahoo! Index
+ Contains 16 data fields

Feed: com.yahoo.uf.hcard
- Contains 6 data fields

- com.yahoo.uf.hcard/rel:Card.2/vcard:fn
- com.yahoo.uf.hcard/rel:Card.2/vcard:org
- com.yahoo.uf.hcard/rel:Card.3/vcard:fn
- com.yahoo.uf.hcard/rel:Card.3/vcard:org
- com.yahoo.uf.hcard/rel:Card/vcard:fn
- com.yahoo.uf.hcard/rel:Card/vcard:photo

Feed: sm031-LinkedIn
+ Contains 12 data fields

- 31 - Updated: 19-May-09 

Developer tool



Yahoo! | My Yahoo! | Mail | More ▾ Make Y! Your Home Page Welcome, Peter ▾ Sign Out | Help

YAHOO! APPLICATION PLATFORM SearchMonkey

Search WEB SEARCH

Application Dashboard / LinkedIn: Presentation Application (SAMPLE)

Basic Info URLs Data Services Appearance Confirmation

Step 4: Map your data services to your presentation layer.

Modify Appearance with PHP:

Use PHP to map data from your data services to different components of your presentation application. Clicking one of the **Link** to the right inserts a short PHP snippet that retrieves the specified data. Once you have mapped your data, click **Save and Refresh** to check your work in the **Preview Pane** at the bottom of the screen.

```
public static function getOutput() {
    $ret = array();

    // Image
    $ret['image'][0]['src'] =
        Data::get('com.yahoo.uf.hcard/rel:Card/vcard:photo');
    $ret['image'][0]['alt'] = "linkedin image";
    $ret['image'][0]['allowResize'] = TRUE;

    // Key Value pairs - up to 4
    $ret['dict'][0]['key'] = "Summary";
    $ret['dict'][0]['value'] =
        Data::get('sm031-LinkedIn/rel:Profile/reference:summary');
    $ret['dict'][1]['key'] = "Industry";
    $ret['dict'][1]['value'] =
        Data::get('sm031-LinkedIn/rel:Profile/job:industry');
    $ret['dict'][2]['key'] = "Location";
```

Data Inputs: Click the **Link** to paste the data into your PHP at the cursor.

-yahooindex

- dc:identifier
- media:fileSize
- dc:format
- dc:language
- finance:stockTickers
- rel:Posting
 - dc:type
 - context:position
 - dc:title
 - dc:description
 - dc:date
 - rel:link

ay-09 

The screenshot shows the Yahoo! Application Platform interface. At the top, there's a banner with the text "Developer tool" and a cartoon monkey wearing glasses. The main header reads "YAHOO! APPLICATION PLATFORM SearchMonkey". Below the header, the title "Application Dashboard / LinkedIn: Presentation Application (SAMPLE)" is displayed. A navigation bar below the title includes tabs for "Basic Info", "URLs", "Data Services", "Appearance", and "Confirmation", with "Confirmation" being the active tab. The main content area shows a LinkedIn profile for "LinkedIn: Amit Kumar" with a thumbnail, name, industry (internet), location (San Francisco Bay Area, United States), and current work experience (President, IIT Delhi Alumni Association). Below this, there's a link to the LinkedIn URL with a "Cached" option. A note says "Now you have a SearchMonkey application. Check it out above. We are using the first URL in your list of Test URLs to generate the preview." It also mentions "Here's what to do next: Try this Presentation Application in Yahoo! Search. (this link only works for you) Make the latest version of your application sharable so that others may use it." A small note at the bottom says "Don't forget to kick back and have a banana." At the bottom of the page are "Home" and "Back" buttons, and a "y-09 Y!" logo.

Future Work

- Encouraging reception from publishers and developers
- Next steps
 - Increasing relevance through better understanding of query intent
 - Focusing on task completion instead of single results
 - Opening up more of search, other parts of Yahoo!
 - Increasing the complexity of applications
- Get engaged
 - Help in bringing metadata to the Semantic Web, preserving quality
 - Talk to us about applications you foresee
 - Facilitate the creation of industry standard vocabularies
- Contact
 - developer.yahoo.com/searchmonkey/
 - mailing lists
 - searchmonkey-developers@yahoogroups.com
 - searchmonkey-siteowners@yahoogroups.com
 - forums
 - http://suggestions.yahoo.com/searchmonkey
 - http://suggestions.yahoo.com/datarss

- 34 - Updated: 19-May-09 Y!



SearchMonkey

the monkey is out!

- 35 -

Updated: 19-May-09

