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.

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

HTML meta tags

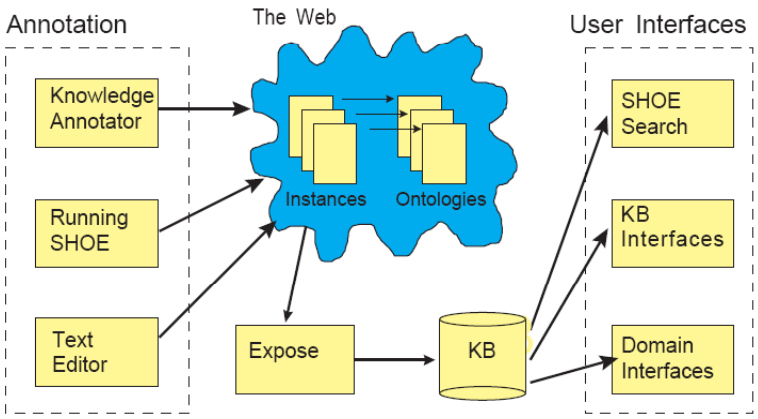
```
<HTML>
<HEAD profile="http://dublincore.org/documents/dcq-html/">
<META name="DC.author" content="Peter Mika">
<LINK rel="DC.rights copyright"
  href="http://www.example.org/rights.html" />
<LINK rel="meta" type="application/rdf+xml" title="FOAF"
     href= "http://www.cs.vu.nl/~pmika/foaf.rdf">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

SHOE example (Hefflin & Hendler, 1996)

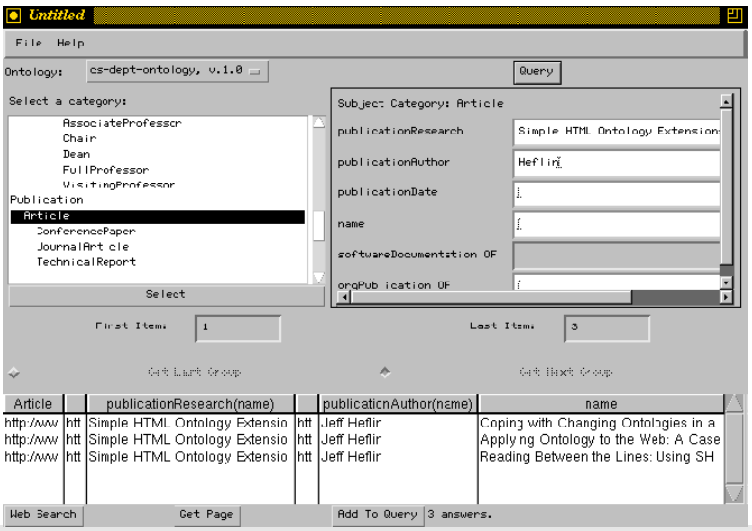
```
<ONTOLOGY "our-ontology" VERSION="1.0">
<ONTOLOGY-EXTENDS "organization-ontology" VERSION="2.1" PREFIX="org"
   URL="http://www.ont.org/orgont.html">
<ONTDEF CATEGORY="Person" ISA="org.Thing">
<ONTDEF RELATION="lastName" ARGS="Person STRING">
<ONTDEF RELATION="firstName" ARGS="Person STRING">
<ONTDEF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

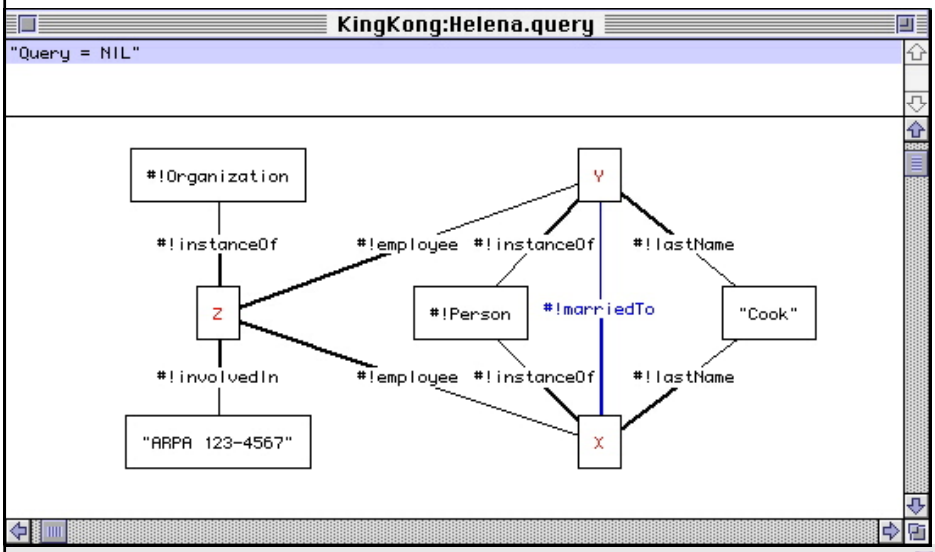
SHOE system



SHOE Text-based query interface



SHOE Graphical Query Interface



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&apos;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>
-->
```

Example: Creative Commons

- Current: rel attribute (HTML4)

```
This work is licensed under a <a rel="license"
href="http://creativecommons.org/licenses/by/3
.0/us/">Creative Commons Attribution 3.0
United States License</a>.
```

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

Example: microformats

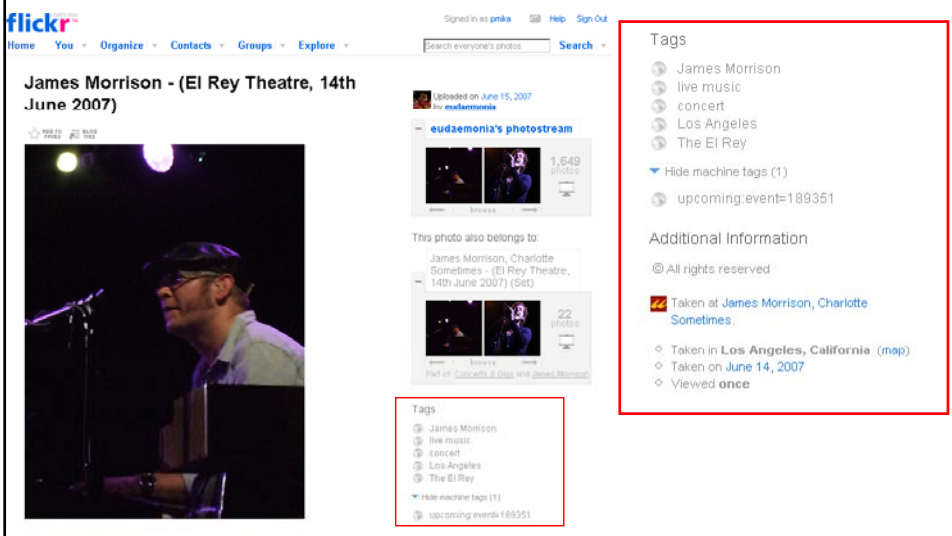
```
<div class="vcard">
  <a class="email fn" href="mailto:jfriday@host.com">Joe Friday</a>
  <div class="tel">+1-919-555-7878</div>
  <div class="title">Area Administrator, Assistant</div>
</div>
```

```
<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>.
```

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

Example: tags and machine tags



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

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

RDF-based annotation #2: RDFa

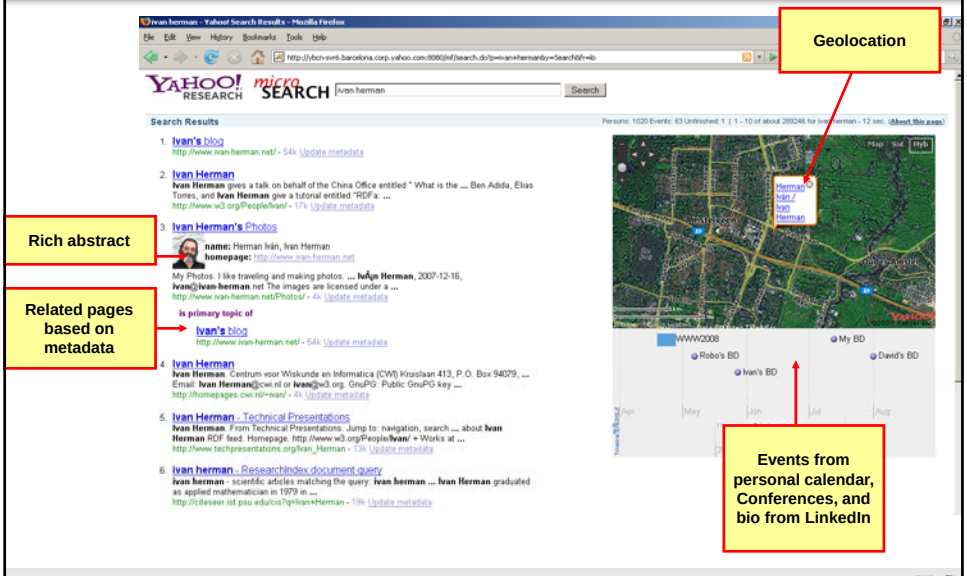
- RDFa
  - World Wide Web Consortium (W3C) last call document
  - Similar intent as eRDF, but full RDF support
    - Requires XHTML
  - Big question: user complexity (→ data quality)

```
<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> ...
```

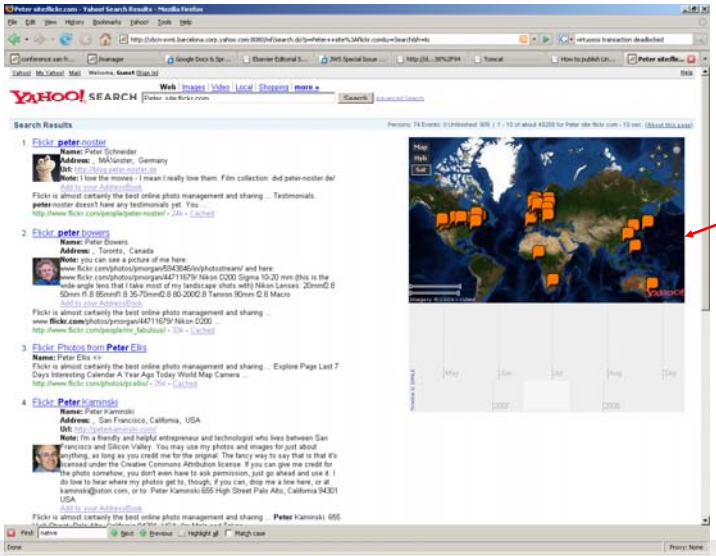

Metadata is out there

- 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/

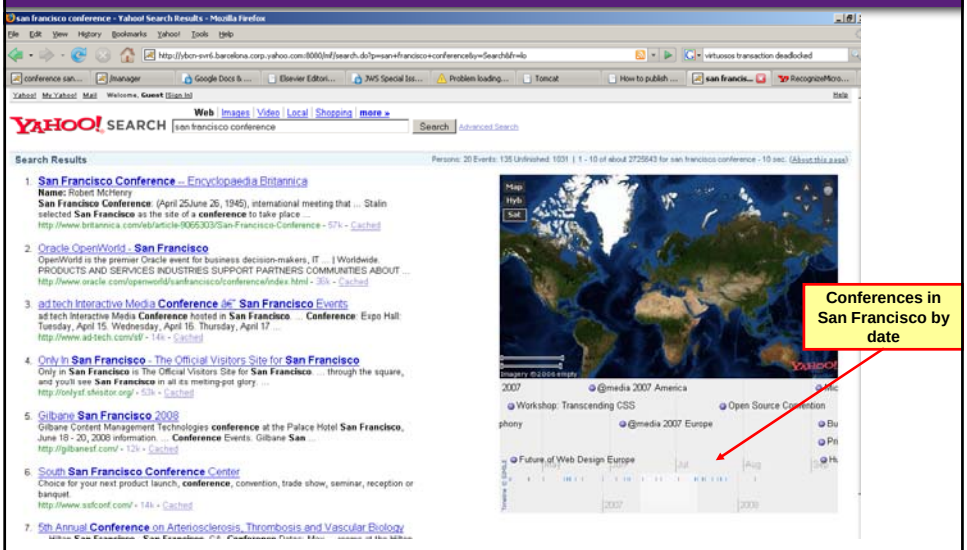
Example: ivan herman



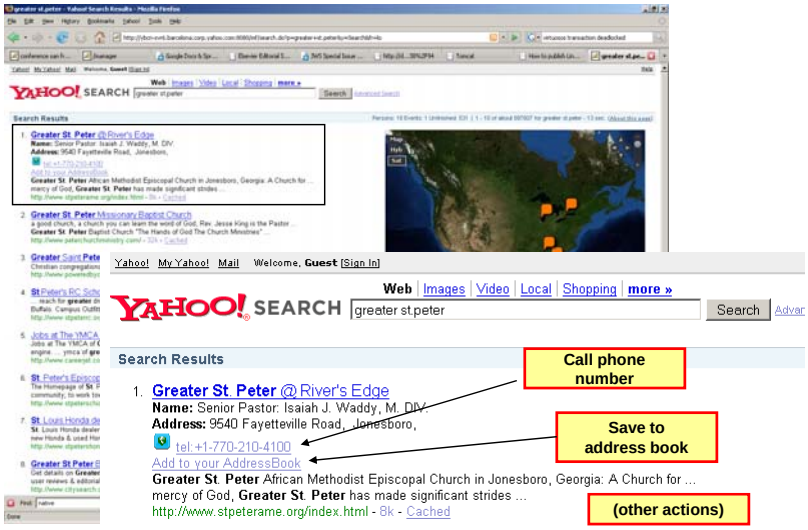
Example: peter site:flickr.com



Example: san francisco conference



Example: greater st. peter



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

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

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 - 132k - 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** - 97k - Cached

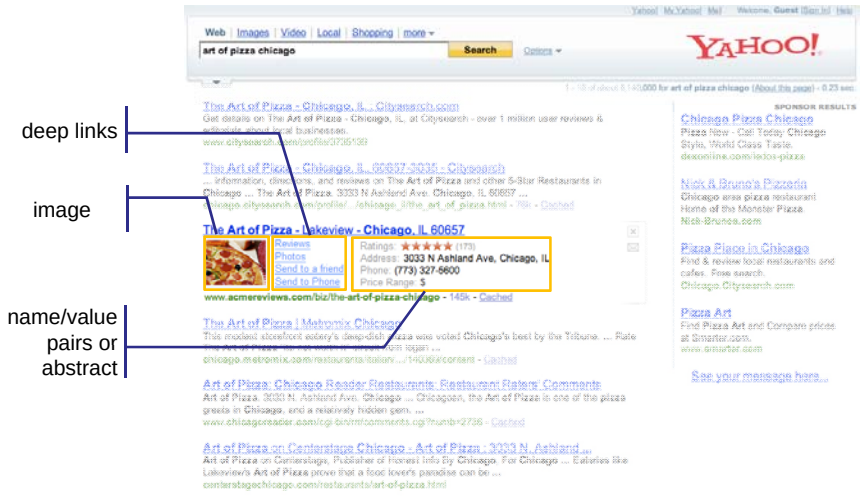
After

Topics for **Getting Pregnant** - BabyCenter
Tools
Before you Try
Active Trying
Having Trouble
BabyCenter's getting **pregnant** tools and information can boost your chances of conception by helping you chart your cycle, read your cervical mucus, and pinpoint ovulation.
www.**babycenter.com/getting-pregnant** - 76k - Cached

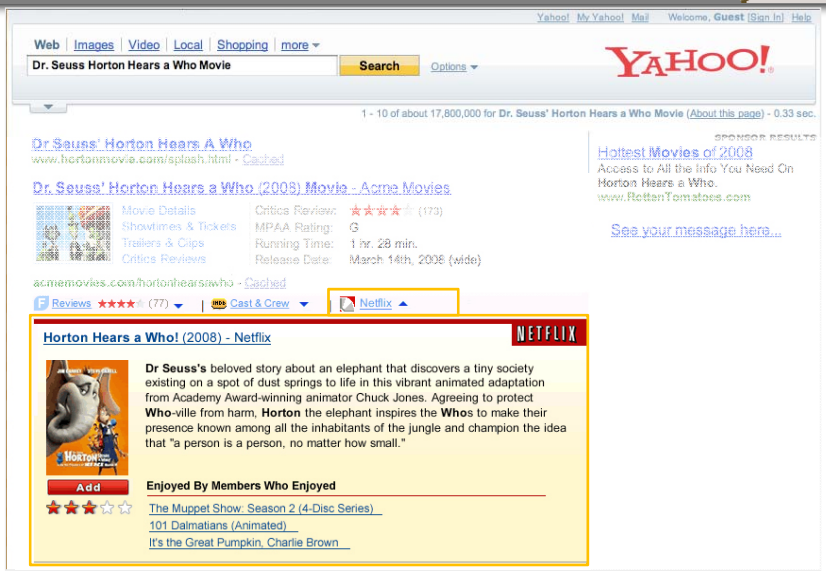
WebMD Allergies Health Center - Find allergy information and latest ...
Overview
Symptoms
Diagnosis
Treatment
Allergies are an abnormal response of the immune system. People who have allergies have an immune system that reacts to a usually harmless substance in the environment. This substance (pollen,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

Italy Travel Guide: Overview - Lonely Planet WorldGuide
Overview
Money & Costs
Itineraries
Sights
When to go: **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** - 46k - Cached

Enhanced Result

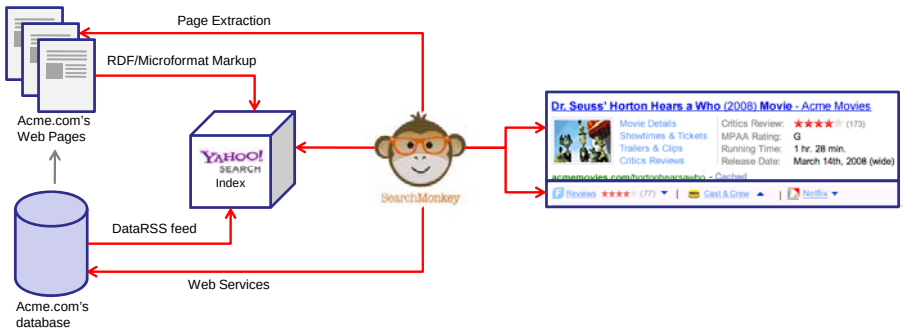


Infobar

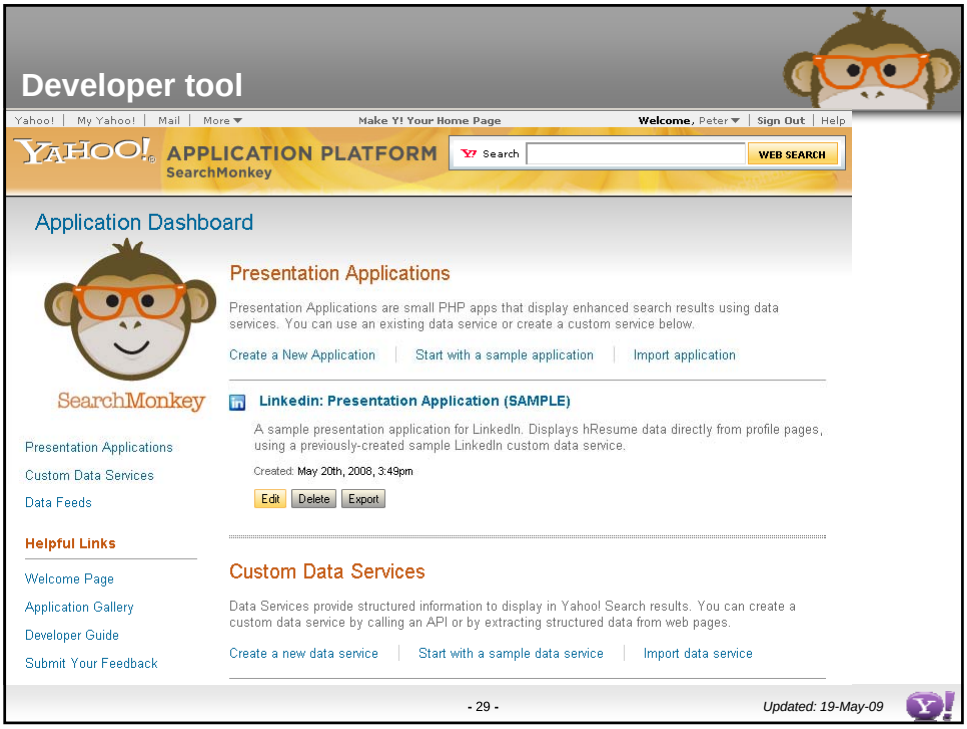


SearchMonkey

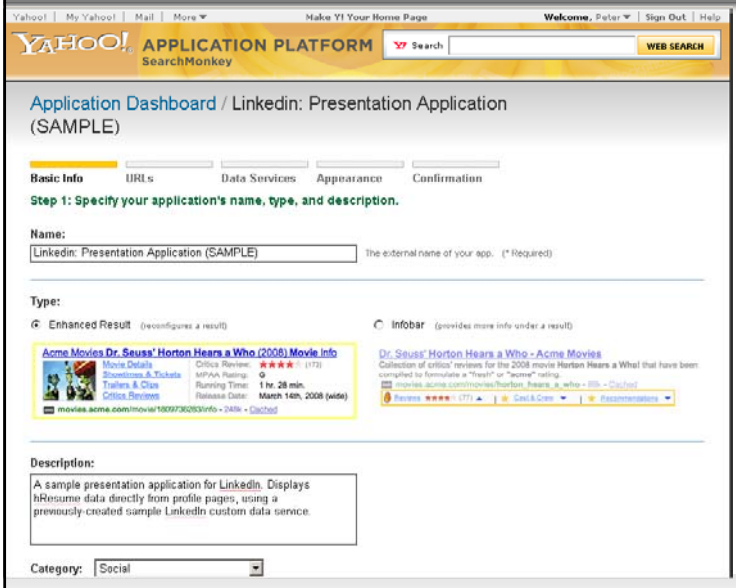
1. site owners/publishers share structured data with Yahoo!.
2. site owners & third-party developers build SearchMonkey apps.
3. consumers customize their search experience with Enhanced Results or Infobars

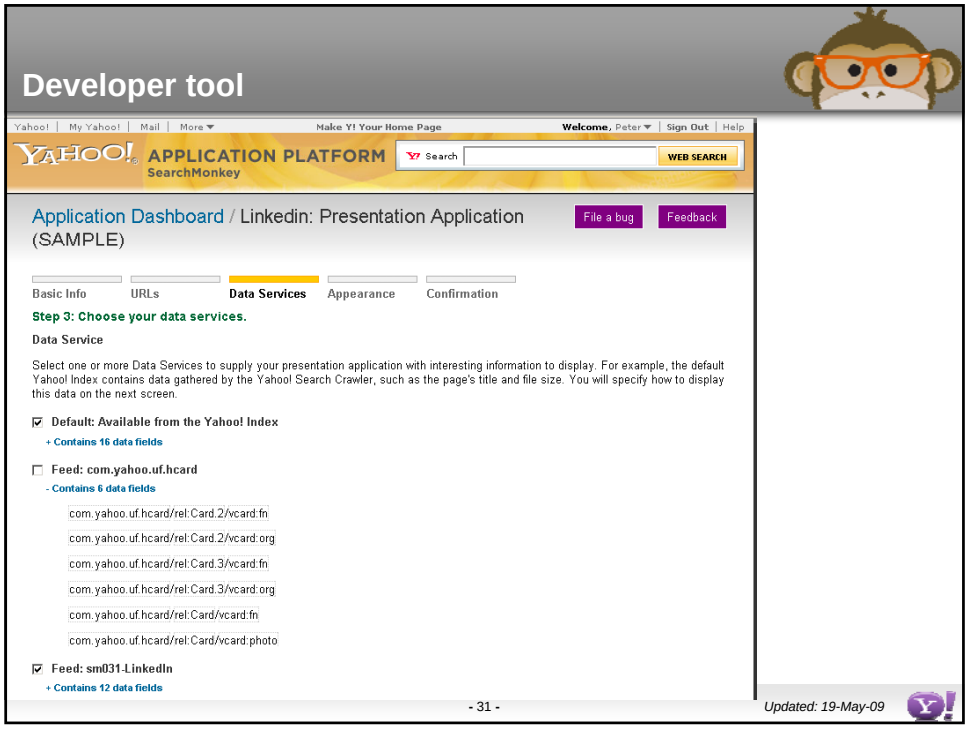


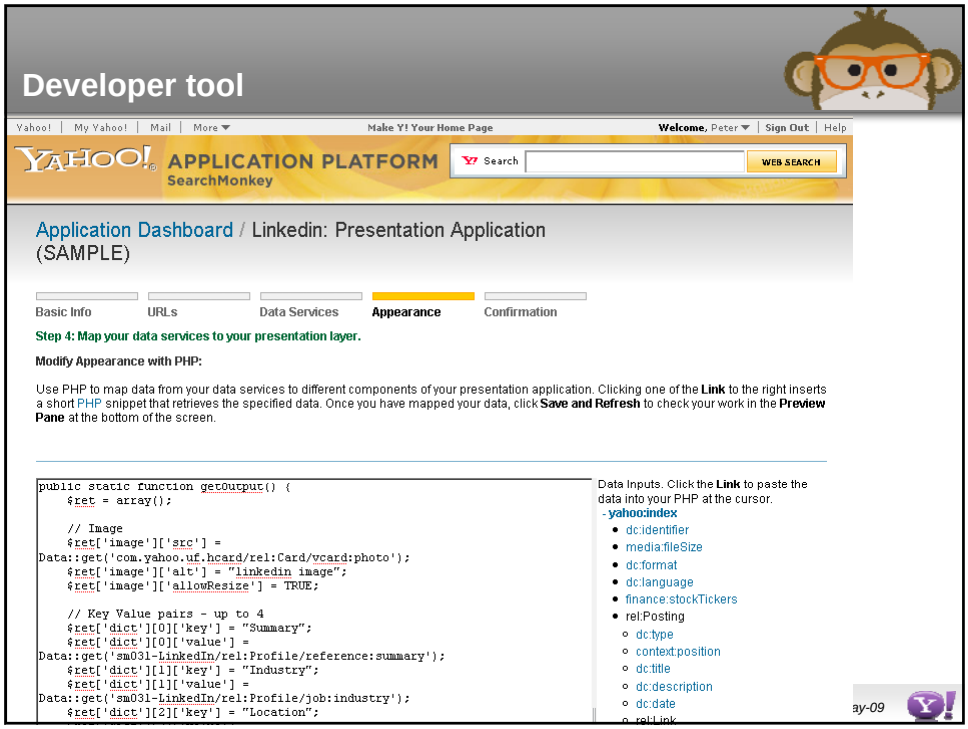
Developer tool



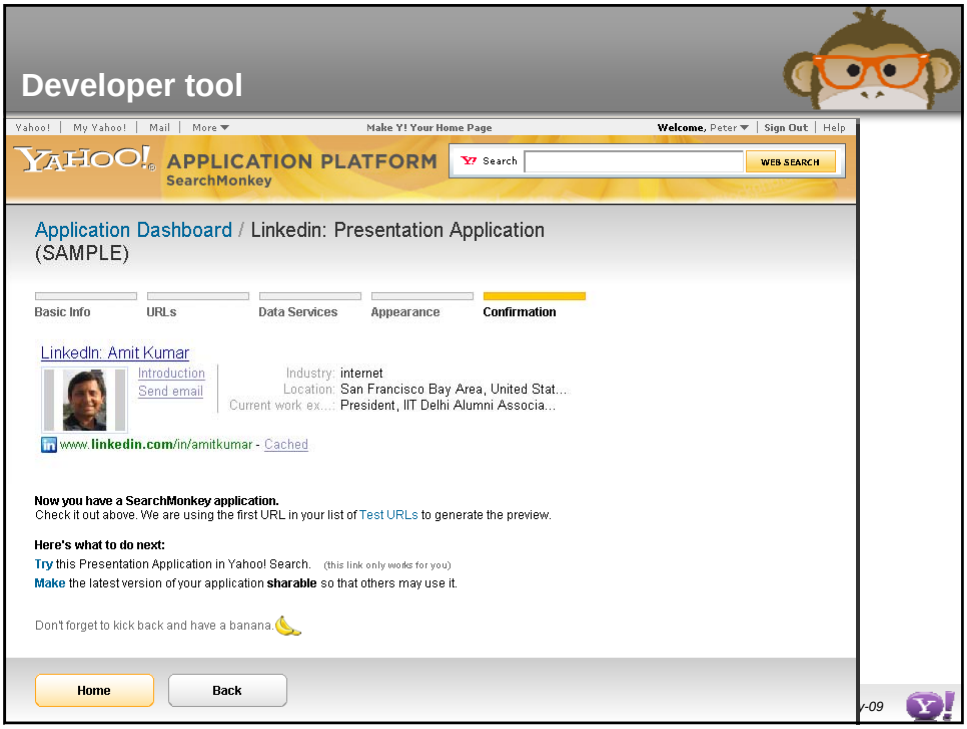
Developer tool




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)
File a bug
Feedback
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']['src'] =
Data::get('com.yahoo.uf.hcard/rel:Card/vcard:photo');
$ret['image']['alt'] = "linkedin image";
$ret['image']['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.
- yahoo:index
dc:identifier
media:fileSize
dc:format
dc:language
finance:stockTickers
rel:Posting
dc:type
context:position
dc:title
dc:description
dc:date
rel:Link

Developer tool



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

the monkey is out!