Enterprise Information Mashups: 
*Integrating Information, Simply*

Anant Jhingran  
CTO, Information Management  
IBM
Outline

- Web 2.0 and Info 2.0
- Example and the research problems we see
- IBM efforts in this area
- Creativity v. Control
Information Technology Spend “had” been growing nicely

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1964</td>
<td>S/360 debuts</td>
</tr>
<tr>
<td>1971</td>
<td>First Intel Micro</td>
</tr>
<tr>
<td>1981</td>
<td>IBM PC</td>
</tr>
<tr>
<td>1994</td>
<td>Netscape Navigator</td>
</tr>
<tr>
<td>2000</td>
<td>Dot-com collapse</td>
</tr>
</tbody>
</table>

Information Technology Spend "had" been growing nicely.
Over time, complexity got built into the IT systems

Actual Application Architecture for Consumer Electronics Company
And using Information as a Strategic Asset to build better Architectures
Open Innovation is Here to Stay, Exemplified by Web 2.0

But...
Web 2.0 outside, and inside an enterprise will succeed *only with* a Info 2.0 Mashup Fabric

Enables the same separation of “data” and “logic” that revolutionized the use of databases in the ’80’s.

Enables the same separation of “information” and “process” that is now happening in Web 1.5
Within enterprises, it will...

- **Enable connections to information that does not make it into the enterprise IT Architectures:**
  - Email
  - Presentations and Documents
  - External Data (Web)
  - Spreadsheets
  - Decision Support Datasets…

- **And Enable it to be done “quickly”, as “assembly” as opposed to as “programming”**
How the Architecture could play out…

External Web  LOB Focus  IT Focus

- Situational Apps
- Info 2.0 Fabric
- Process Server/ESB
- Information Integration

SaaS Model  Software Model
Meet Pete, an insurance agent in Florida.

He sees a news report of a severe storm. What is the company’s risk?

He needs to forward a risk summary to executives.
Flood Risk Assessment Mashup

- Report
- Standardization
- Standardize
- www.floodlevels.com
- www.floodlevels.com
- policy XLS
- water.usgs.gov
- edc.usgs.gov
- dotd.florida.gov
- Screen Scraping
- Lineage
So how can Pete write his mashup simply?
So how can Pete write his mashup simply?

```php
<?php
// Get policy holders in a Policy object array
$url = "file://policies/myclients.xsl";
$content = file_get_contents($url);
$policyArr = getPolicy($content);

// Find high risk zones
$url = "http://www.floodlevels.com";
$content = file_get_contents($url);
// Do screen scraping to extract high risk zones
$zoneArr = findRiskyZones($content);

// Initialize the return array
$riskArr = array();
// Find corresponding policy holders for each city
foreach ($policyArr as $policy)
{
    if ($policy->amount < 250000)
    {
        continue;
    }
    // Standardize the address
    $policyZone = findZone($policy->address);
    // Check whether this policy affected
    foreach ($zoneArr as $zone)
    {
        if ($zone == $policyZone)
        {
            // This policy carries a high risk.
            // Insert into high risk array
            $riskArr [] = $policy;
        }
    }
}
// Send email to manager for high risk policies
sendEmail("suzan@trustinsurance.com", "High risk policies",
          $riskArr);
?>
```
So how can Pete write his mashup simply?

```xml
sendMail("suzan@trustinsurance.com",
<highRiskPolicies>
{
    for $i in url("file://policies/myclients.xsl")
    for $j in url("http://www.floodlevels.com")
    where $i//amount > 250000 and
    $i//address in $j//zone
    return <policy> {$i} </policy>
}
</highRiskPolicies>);
```
So how can Pete write his mashup simply?

Simplicity

Accuracy

GUIs, Spreadsheets, Wikis
So how can Pete write his mashup simply?

Flood risk for homes in myclients.xsl worth over 250000

How do we get there?
Research Agenda

- It is all about “simplicity” – do deep research and build deep technology, but make the job of application writer much easier!

- Much of our past research is applicable (including Information Manifold and its children), but new problems exist because of new target users.
Info 2.0 Mashup Fabric needs to address these issues, over time

- **How to create such a Mashup?**
  - Finding what exists, specifying what he wants, and creating what is needed (expressiveness vs. ease of use – DWIS vs. DWIM)

- **How to integrate the information?**
  - What is the minimal level of semantics that the Information 2.0 layer needs to have, and has the world evolved to make it easier now?

- **How to deal with unstructured data?**

- **How do Mashups evolve?**
How does Pete find the floodlevels.com Mashup?

- Pages on floodlevels.com are dynamically generated AJAX pages (produced by another mashup)
- Pete may have typed “Flood Levels Louisiana” into a search engine
- Similar to deep Web search problem, but now we have to deal with joins and other mashup operations, or even workflow
- Search has to understand the logic of the mashup
How does Pete specify his Mashup?

- **Pete is an insurance agent, not an expert Javascript or PHP/Java/Ruby/etc. programmer**

- **How does Pete specify a screen scraper if needed?**

- **How does Pete describe the Mashup flow?**
  - Current mashups are a hodge-podge of application and data access
  - Similarity to ETL Flow
  - Is the answer an XQuery-like language for mashups, or programming by example?

Web 2.0 needs simple methods to write mashups!
Can he create the Mashup by giving an example?
Could it have been even easier?

- Could Pete’s mashup have been dynamically constructed when he searched for “flood levels for zipcodes 33101, 34106, etc.”?

  - Test of Time Award: “Information Manifold”
    Querying Heterogeneous Information Sources Using Source Descriptions by A. Halevy, A. Rajamaran, and A. Ordille

  - automatically finding the right sources based on query

Extend Information Manifold to dynamically create Mashups!
How does one simplify “semantics”?  

**Helped by:**
- Microformats growing in popularity in the open community
- Standardization services increasingly available
- Master Data Management taking off in enterprises

**Issues:**
- Standardization is inherently uncertain. How is uncertainty handled?
- Quality of services differ. How to track the lineage of both data and integration services?
- Services vary in price. How to trade-off price, quality, and time?

**Search shows us some ways**
Issues in Unstructured Data

- *Everybody wants to run analytics on unstructured data, and create structured data, and then we are back in our favorite world.*
  
  *This poses two challenges:*
  
  - Analytics are hard and require some fundamentally new techniques.
  - The extracted structured (meta-) data is inherently imprecise.

But unstructured query systems have evolved to address this!
In another Web 2.0 sense, how does this co-exist and augment social tagging?

Manual tagging – By Professionals

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled vocabularies &amp; standard taxonomies</td>
<td>Costly</td>
</tr>
<tr>
<td>Higher quality</td>
<td>Human resource intensive</td>
</tr>
<tr>
<td></td>
<td>Cannot keep up</td>
</tr>
<tr>
<td>Example: ?</td>
<td></td>
</tr>
</tbody>
</table>

Social Tagging – By Users

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>User driven</td>
<td>Ambiguity</td>
</tr>
<tr>
<td>Emergent folksonomies</td>
<td>Uncontrolled vocabulary</td>
</tr>
<tr>
<td>Serendipitous browsing</td>
<td>Synonyms</td>
</tr>
<tr>
<td>Examples: Del.icio.us and Flickr</td>
<td></td>
</tr>
</tbody>
</table>

Automated Tagging – By Machine

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learns from professional &amp; user tagging</td>
<td>Requires training of models</td>
</tr>
<tr>
<td>Lower human cost</td>
<td>Lower quality than manual tagging</td>
</tr>
<tr>
<td>Example: Semantic tagging</td>
<td></td>
</tr>
</tbody>
</table>

Popularity

High-value content & enterprise data sources

Deep archives, large personal collections

“Long tail”

Consumer content

Digital item
Mashup Starter Kit – A Mashup Fabric for Intranet Applications being built @ IBM
As my Mom Used to say (perhaps still says!)

“How can you have any pudding if you don’t eat clean your feet?”

(Apologies to Pink Floyd, “The Wall”)
How do we unleash creativity, yet keep light control?

Transform
Create and Explore
Assemble & Use
Manage this
Unleash this

Web Content
Departmental Content
Personal Assets
Enterprise Information
Summary

- **Web style of architectures represent the next “sustainable” phase of IT spend**

- **The database research community can make a big difference!**
  - Re-enable the separation of data and logic: Web 2.0 built on Info 2.0!

- **New research problems exist**
  - Ease of use and ad-hoc integration.
  - Bringing Unstructured and (semi-) structured data

- **We at IBM are building such an Info 2.0 Fabric, targeting enterprise situational applications**

- **One of the biggest battles will be creativity vs. control**