Big Data from the Web

Georg Gottlob, TU Wien & University of Oxford

The WWW is older than we may think...
This talk is about extracting data from the Web... ...and about our (ad)ventures in:

- Theory
- Technology
- Business
- Commercialization

(Kaplan Turbine)
The Web is the Largest Database
The Web is the Largest Database
The Web is not a Database

- No common schema for data.
- No structured queries possible
- Only keyword search
The Web is not a Database!

- No common schema for data.
- No structured queries possible
- Only keyword search

For example, the following queries cannot be answered:

- List all flats for sale in Vienna having a balcony that cost less than €500,000 and that are in an area with above-average many Italian restaurants.

- List all used cars for sale in or around Oxford that are not older than 5 years and that cost no more than £6,000.

- Find translation offices in Europe that translate German to Basque, charge no more than €20 per page, and accept VISA card payments.
Web Data Extraction

<table>
<thead>
<tr>
<th>ref-code</th>
<th>postcode</th>
<th>bedrooms</th>
<th>bathrooms</th>
<th>available</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>33453</td>
<td>OX2 6AR</td>
<td>3</td>
<td>2</td>
<td>15/10/2013</td>
<td>£1280 pcm</td>
</tr>
<tr>
<td>33433</td>
<td>OX4 7DG</td>
<td>2</td>
<td>1</td>
<td>18/04/2013</td>
<td>£995 pcm</td>
</tr>
</tbody>
</table>
Function \( f: \) HTML Parse tree \( \Rightarrow \) Subtrees

Leaves of subtrees are among leaves of orig. tree
A HTML page

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html> <body>
<h1>People @ DBAI</h1>
<table border="1" cellpadding="3" cellspacing="1">
  <tr> <td>Georg Gottlob</td>
   <td>gottlob@dbai.tuwien.ac.at</td>
   <td>18420</td>
  </tr>
  <tr> <td>Christoph Koch</td>
   <td>koch@dbai.tuwien.ac.at</td>
   <td>18449</td>
  </tr>
</table>
</body> </html>

People @ DBAI

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georg Gottlob</td>
<td>gottlob@...</td>
<td>18420</td>
</tr>
<tr>
<td>Christoph Koch</td>
<td>koch@...</td>
<td>18449</td>
</tr>
</tbody>
</table>
People @ DBAI

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georg Gottlob</td>
<td><a href="mailto:gottlob@dbai.tuwien.ac.at">gottlob@dbai.tuwien.ac.at</a></td>
<td>18420</td>
</tr>
<tr>
<td>Christoph Koch</td>
<td><a href="mailto:koch@dbai.tuwien.ac.at">koch@dbai.tuwien.ac.at</a></td>
<td>18449</td>
</tr>
</tbody>
</table>
Predicate \textit{employee}

```html
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html> 
<body> 
<h1>People @ DBAI</h1> 
<table border="1" cellspacing="1" style="margin:5px 5px 5px 5px; border:1px solid #000000; width:100%; text-align: left; ">
  <tr> 
    <td>Georg Gottlob</td> 
    <td>gottlob@dbai.tuwien.ac.at</td> 
    <td>18420</td> 
  </tr>
  <tr> 
    <td>Christoph Koch</td> 
    <td>koch@dbai.tuwien.ac.at</td> 
    <td>18449</td> 
  </tr> 
</table>
</body> </html>
```
**Predicate phone**

```html
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html> <body>
<h1>People @ DBAI</h1>
<table border="1" cellpadding="3" cellspacing="1">
  <tr> <td>Georg Gottlob</td> <td>gottlob@dbai.tuwien.ac.at</td> <td>18420</td> </tr>
  <tr> <td>Christoph Koch</td> <td>koch@dbai.tuwien.ac.at</td> <td>18449</td> </tr>
</table>
</body> </html>
```
| Logic heaven |
| DB theory heaven |
| DB programming heaven |
| Application design heaven |

| MSO |
| Monadic Datalog |
| Elog |
| Lixto Visual Wrapper |
| LiXto Suite |
Monadic Datalog as a Wrapping Language

entry(X) :- root(R), firstchild(R,U), label[html](U),
            firstchild(U,V), label[body](V),
            firstchild(V,W), label[table](W),
            firstchild(W,X), label[tr](X).

entry(X):- entry(Y), nextsibling(Y,X).

name(X) :- entry(E), firstchild(E, X), label[td](X).

eemail(X) :- name(N), nextsibling(N, X), label[td](X).

phone(X) :- email(M), nextsibling(M, X), label[td](X).
entry(X) :- root(R), firstchild(R,U), label[html](U), firstchild(U,V), label[body](V), firstchild(V,W), label[table](W), firstchild(W,X), label[tr](X).

entry(X):- entry(Y), nextsibling(Y,X).

name(X) :- entry(E), firstchild(E, X), label[td](X).

email(X) :- name(N), nextsibling(N, X), label[td](X).

phone(X) :- email(M), nextsibling(M, X), label[td](X).
entry(X) :- root(R), firstchild(R,U), label[html](U),
           firstchild(U,V), label[body](V),
           firstchild(V,W), label[table](W),
           firstchild(W,X), label[tr](X).
entry(X):- entry(Y), nextsibling(Y,X).

name(X) :- entry(E), firstchild(E, X), label[td](X).
email(X) :- name(N), nextsibling(N, X), label[td](X).
phone(X) :- email(M), nextsibling(M, X), label[td](X).
Monadic Datalog as a Wrapping Language

entry(X) :- root(R), firstchild(R,U), label[html](U),
          firstchild(U,V), label[body](V),
          firstchild(V,W), label[table](W),
          firstchild(W,X), label[tr](X).
entry(X):- entry(Y), nextsibling(Y,X).
name(X) :- entry(E), firstchild(E, X), label[td](X).
email(X) :- name(N), nextsibling(N, X), label[td](X).
phone(X) :- email(M), nextsibling(M, X), label[td](X).
Monadic Datalog as a Wrapping Language

entry(X) :- root(R), firstchild(R,U), label[html](U),
           firstchild(U,V), label[body](V),
           firstchild(V,W), label[table](W),
           firstchild(W,X), label[tr](X).

entry(X):- entry(Y), nextsibling(Y,X).

name(X) :- entry(E), firstchild(E, X), label[td](X).

email(X) :- name(N), nextsibling(N, X), label[td](X).

phone(X) :- email(M), nextsibling(M, X), label[td](X).
Monadic Datalog as a Wrapping Language

entry(X) :- root(R), firstchild(R, U), label[html](U),
            firstchild(U, V), label[body](V),
            firstchild(V, W), label[table](W),
            firstchild(W, X), label[tr](X).

entry(X) :- entry(Y), nextsibling(Y, X).

name(X) :- entry(E), firstchild(E, X), label[td](X).

e-mail(X) :- name(N), nextsibling(N, X), label[td](X).

phone(X) :- email(M), nextsibling(M, X), label[td](X).
How complex is Monadic Datalog?

Theorem [G. & Koch 2002]:

Monadic Datalog over trees has combined complexity:

\[ O(|\text{data tree}| \times |\text{program}|) \]
Examples of Special predicates:

- subelem(S,X,Path,…)
- before(X,Y,…,d.)
- after(X,Y,…)
- property(X,Attribute, Op,Value…..)
- document(URL,D)
- getdocumentFromHref(X,D),
  etc.

Additional features:  
Stratified negation,  
string processing  
ontological concepts “phonenumber(X)”  
ranges:  H(S,X) :- body(……..)[1,5]  
object hierarchies
tableseq(S, X) ← document(“www.ebay.com/”, S), subsq(S, (.body, []), (.table, []), (table, []), X), before(S, X, (.table, [(elementtext, item, [])], 1, 1, _, _), after(S, X, (.hr, []), 1, 1, _, _)

record(S, X) ← tableseq(_, S), subelem(S, .table, X)

itemnum(S, X) ← record(_, S), subelem(S, *.td, X), notbefore(S, X, (.td, []), maxint)

itemdes(S, X) ← record(_, S), subelem(S, (*td.*content, [(a, 0)]), X)

price(S, X) ← record(_, S), subelem(S, (*td, [(elementtext, Y, 1)]), X), valuta(Y)

bids(S, X) ← record(_, S), subelem(S, *.td, X), before(S, X, (.td, []), 1, 30, Y, _), price(S, Y)

currency(S, X) ← price(_, S), subtext(S, Y, X), valuta(Y)

pricewc(S, X) ← price(_, S), subtext(S, [0-9]+, X)
Lixto Visual Developer (VD)

Navigation Steps

Mozilla Web Browser

Extraction Configuration
## Comparison of Prices

### Prices by Different Sources

<table>
<thead>
<tr>
<th>Week</th>
<th>EAN/UPC</th>
<th>Product Name</th>
<th>You</th>
<th>Source 1</th>
<th>Source 2</th>
<th>Source 3</th>
<th>Source 4</th>
<th>Source 5</th>
<th>Source 6</th>
<th>Source 7</th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Meantime</th>
<th>Diff</th>
<th>Diff %</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>4045827387809</td>
<td>Product #132</td>
<td>1070€</td>
<td>799€</td>
<td>798€</td>
<td>796€</td>
<td>798€</td>
<td>798€</td>
<td>798€</td>
<td>1070€</td>
<td>889€</td>
<td>798€</td>
<td>0%</td>
<td>+50€</td>
<td>+7%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827401063</td>
<td>Product #137</td>
<td>799€</td>
<td>749€</td>
<td>798€</td>
<td>796€</td>
<td>798€</td>
<td>798€</td>
<td>798€</td>
<td>1070€</td>
<td>889€</td>
<td>798€</td>
<td>0%</td>
<td>+50€</td>
<td>+7%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>40458274040035</td>
<td>Product #138</td>
<td>890€</td>
<td>700€</td>
<td>780€</td>
<td>780€</td>
<td>780€</td>
<td>780€</td>
<td>800€</td>
<td>800€</td>
<td>836€</td>
<td>800€</td>
<td>0%</td>
<td>-34€</td>
<td>-4%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827408969</td>
<td>Product #142</td>
<td>999€</td>
<td>1005€</td>
<td>1005€</td>
<td>1032€</td>
<td>1000€</td>
<td>1000€</td>
<td>1000€</td>
<td>1000€</td>
<td>1000€</td>
<td>1000€</td>
<td>100%</td>
<td>-80€</td>
<td>-8%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827407010</td>
<td>Product #144</td>
<td>899€</td>
<td>1300€</td>
<td>1300€</td>
<td>1100€</td>
<td>1100€</td>
<td>1100€</td>
<td>1100€</td>
<td>1100€</td>
<td>100%</td>
<td>-401€</td>
<td>-31%</td>
<td>+203€</td>
<td>+26%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827409533</td>
<td>Product #145</td>
<td>1128€</td>
<td>1445€</td>
<td>1128€</td>
<td>1445€</td>
<td>1287€</td>
<td>1287€</td>
<td>1287€</td>
<td>1287€</td>
<td>1287€</td>
<td>1287€</td>
<td>100%</td>
<td>-401€</td>
<td>-31%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827417096</td>
<td>Product #146</td>
<td>832€</td>
<td>849€</td>
<td>832€</td>
<td>832€</td>
<td>832€</td>
<td>832€</td>
<td>832€</td>
<td>832€</td>
<td>866€</td>
<td>832€</td>
<td>0%</td>
<td>+37€</td>
<td>+8%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827422715</td>
<td>Product #160</td>
<td>899€</td>
<td>914€</td>
<td>932€</td>
<td>1299€</td>
<td>1299€</td>
<td>1299€</td>
<td>1299€</td>
<td>1299€</td>
<td>1299€</td>
<td>1299€</td>
<td>67%</td>
<td>+57€</td>
<td>+8%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827439509</td>
<td>Product #162</td>
<td>999€</td>
<td>796€</td>
<td>999€</td>
<td>998€</td>
<td>998€</td>
<td>998€</td>
<td>998€</td>
<td>998€</td>
<td>998€</td>
<td>998€</td>
<td>0%</td>
<td>+203€</td>
<td>+26%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827449553</td>
<td>Product #177</td>
<td>1426€</td>
<td>923€</td>
<td>1099€</td>
<td>929€</td>
<td>1426€</td>
<td>1426€</td>
<td>1426€</td>
<td>1426€</td>
<td>1151€</td>
<td>1151€</td>
<td>0%</td>
<td>+203€</td>
<td>+26%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827453604</td>
<td>Product #509</td>
<td>594€</td>
<td>474€</td>
<td>700€</td>
<td>700€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>0%</td>
<td>+0€</td>
<td>+0%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827466666</td>
<td>Product #540</td>
<td>690€</td>
<td>1199€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>690€</td>
<td>0%</td>
<td>+16€</td>
<td>+2%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827466697</td>
<td>Product #545</td>
<td>713€</td>
<td>1599€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>713€</td>
<td>0%</td>
<td>+0€</td>
<td>+0%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827478481</td>
<td>Product #547</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>1499€</td>
<td>0%</td>
<td>+0€</td>
<td>+0%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827482764</td>
<td>Product #581</td>
<td>683€</td>
<td>691€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>683€</td>
<td>0%</td>
<td>+16€</td>
<td>+2%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827492771</td>
<td>Product #582</td>
<td>689€</td>
<td>698€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>689€</td>
<td>0%</td>
<td>+16€</td>
<td>+2%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827492798</td>
<td>Product #586</td>
<td>790€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>775€</td>
<td>0%</td>
<td>+24€</td>
<td>+3%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827493150</td>
<td>Product #587</td>
<td>649€</td>
<td>631€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>785€</td>
<td>0%</td>
<td>-136€</td>
<td>-17%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827505690</td>
<td>Product #582</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>498€</td>
<td>0%</td>
<td>+1€</td>
<td>+0%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4045827511090</td>
<td>Product #842</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>594€</td>
<td>0%</td>
<td>+1€</td>
<td>+0%</td>
<td></td>
</tr>
</tbody>
</table>
McKinsey & Company acquires Lixto

Periscope, a McKinsey Solution, has enhanced its suite of revenue management solutions with McKinsey & Company’s acquisition of Lixto. McKinsey …

Periscope, a McKinsey Solution, has enhanced its suite of revenue management solutions with McKinsey & Company’s acquisition of Lixto.

McKinsey & Company, a global management consulting firm, has announced the acquisition of Lixto Software, a Vienna-based SaaS solutions company in the field of online competitive intelligence.

This acquisition will enhance Periscope, an integrated suite of solutions within McKinsey Solutions designed to deliver sustainable ROS improvement through better pricing, promotions, assortment and performance management.

Lixto will continue to offer services for automated web data extraction, data mapping and matching, providing frequent and reliable competitive online data on pricing, promotions, assortment range, terms/conditions and product attributes, delivering a competitive advantage to a variety of clients in Retail, Travel & Hospitality, High Tech, Consumer Products, Distribution, and Manufacturing industries. These capabilities will support clients with market intelligence and actionable insights, powered by Periscope Performance Vision market analytics and Periscope Price Advisor dynamic pricing solutions.
This [Lixto] is great technology, but …
... too many people needed for generating and maintaining the wrappers!
From the tourism, real estate, and retail domains, we understood independently that automated wrapper-generation would be useful.

But how can this be achieved?
Need for Fully Automated Extraction Technology

Example: Real Estate UK > 15000 sites
  many not covered by aggregators
  list of all agencies easy to get (source discovery)
  but: manual or semi-automatic wrapping too expensive
    wrapper construction
    testing
    tracking changes

No existing tool or methodology could do it fully automatically
This talk is about extracting data from the Web…
…and about our (ad)ventures in:

- Theory
- Technology
- Business
- Commercialization
DIADEM

a €2.5M project started in April '10

Domain-centric Intelligent Automated Data Extraction Methodology

Application domain with thousands of websites

Application-relevant Structured data (XML or RDF)
rule-based reasoning
knowledge-based
page exploration strategy
navigation planning
global decision making
plausibility checks

machine learning
e.g. neural networks
page classification
visual clue recognition
text classification
small entity identification

Taught knowledge (expert system)
Self-learned knowledge
Rough Idea: Knowledge via Rules

Use “expert” rules that analyze Web pages and interact with them

- **Ontological rules** (how do entities relate to each other)
  - a flat is a real-estate property
  - a house is a real-estate property
  - a real-estate property has a number of rooms
  - a price consists of a number and a currency

- **Phemomenological rules** (how do entities manifest themselves on the Web?)
  - the text chunk closest to an input field is with high prob. its descriptor.
  - each sales item is described in a “convex” (usual. rectangular) region.

- **Site exploration rules:**
  - before filling a field try to leave it empty
  - rules for handling next-page links

- **Other types of rules**
DIADEM PROJECT

DIADEM lab at Oxford University

2010 2011 2012 2013 2014 2015

spin-out start-up Wrapidity

Funding so far > $7M
New knowledge-based technology combining formalised knowledge with machine learning
Evaluation on 10k+ Sites

- 10,493 Sites from real-estate and used-car
- 45 Node Amazon EC2 cluster running 2.1 days
- 92% Effective wrappers for more than 92% of sites on average
- 97% Precision of extracted primary attributes
- 100 Domain-dependent concepts and relations
- 20 Days (one expert) to adjust system to a new domain
Domains considered so far (since 2014)

- Real estate UK
- Real estate US
- Used cars
- Consumer electronics
- Restaurant chains
- Restaurants in the ‘Open Web’
- Jobs (from company Web sites)
- News
Commercial Impact

ERC Advanced Grant DIADEM + ERC Proof of Concept Grant EXTRALYTICS

2 possibilities:
• Build up company with large client portfolio
• Sell technology, IP & software to strategic partner
Meltwater Acquires Wrapidity to Add AI Capabilities into Media Intelligence Platform

By Sudipto Ghosh  
Posted on February 21, 2017

Meltwater, the leading B2B data analytics company, has acquired London-based web data extraction company Wrapidity for an undisclosed amount. The AI-startup that spun out of Oxford University in 2015 will be a separate entity in Meltwater’s existing platform. By beefing up its “media intelligence” platform, Meltwater will now offer AI-powered automation tools for data analytics and media monitoring from unstructured web-based content.

In the era of specialized AI for MarTech, Wrapidity offers tailor-made solutions to content-specific problems arising in image recognition, Natural Language Processing, and machine learning. By acquiring Wrapidity, Meltwater will be able to automate its data extraction processes to reach out to a wide range of online customers based on accurate analytics of historical and real-time data. Meltwater is expected to further improve Wrapidity’s AI capabilities for content discovery and data asset management, enabling marketers to interrogate data for diverse purposes, including sales enablement, social media monitoring and so on.
New Application: Personalised Ads
New Application: Personalised Ads

Reverse Geocoding

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/11/2008</td>
<td>4:35:44 PM</td>
<td>-70.959850</td>
<td>42.646133</td>
<td></td>
</tr>
<tr>
<td>3/12/2008</td>
<td>4:42:52 PM</td>
<td>-70.959853</td>
<td>42.646133</td>
<td></td>
</tr>
<tr>
<td>3/13/2008</td>
<td>4:43:14 PM</td>
<td>-70.957533</td>
<td>42.646133</td>
<td></td>
</tr>
<tr>
<td>3/14/2008</td>
<td>4:44:14 PM</td>
<td>-70.958849</td>
<td>42.646538</td>
<td></td>
</tr>
<tr>
<td>3/15/2008</td>
<td>4:44:52 PM</td>
<td>-70.957519</td>
<td>42.646133</td>
<td></td>
</tr>
<tr>
<td>3/16/2008</td>
<td>4:45:23 PM</td>
<td>-70.956917</td>
<td>42.646153</td>
<td></td>
</tr>
<tr>
<td>3/17/2008</td>
<td>4:56:50 PM</td>
<td>-70.957059</td>
<td>42.646132</td>
<td></td>
</tr>
<tr>
<td>3/18/2008</td>
<td>5:00:14 PM</td>
<td>-70.957118</td>
<td>42.646323</td>
<td></td>
</tr>
<tr>
<td>3/19/2008</td>
<td>5:02:14 PM</td>
<td>-70.956953</td>
<td>42.646138</td>
<td></td>
</tr>
<tr>
<td>3/20/2008</td>
<td>5:02:48 PM</td>
<td>-70.957052</td>
<td>42.646164</td>
<td></td>
</tr>
<tr>
<td>3/21/2008</td>
<td>5:05:49 PM</td>
<td>-70.956950</td>
<td>42.646108</td>
<td></td>
</tr>
</tbody>
</table>

Käppelestr. 5, 76131 Karlsruhe

Herrenstraße 21, 76133 Karlsruhe

Full street addresses
New Application: Personalised Ads

Reverse Geocoding

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/11/2008</td>
<td>4:30 PM</td>
<td>7.155455</td>
<td>49.665369</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:31 PM</td>
<td>7.155448</td>
<td>49.665372</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:32 PM</td>
<td>7.155433</td>
<td>49.665373</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:33 PM</td>
<td>7.155421</td>
<td>49.665379</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:34 PM</td>
<td>7.155413</td>
<td>49.665337</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:35 PM</td>
<td>7.155409</td>
<td>49.665339</td>
<td>0.00</td>
</tr>
<tr>
<td>3/11/2008</td>
<td>4:36 PM</td>
<td>7.155408</td>
<td>49.665339</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Käppelestr. 5, 76131 Karlsruhe

Herrenstraße 21, 76133 Karlsruhe

Full street addresses

Web Data Extraction

Fressnapf
Was Tiere lieben

Hergard
Freche Kindermode
New Application: Personalised Ads

Reverse Geocoding

User profile is created and later refined. Ads related to animals and children will be sent...
CRIME HEATMAPS
Scenario ①: Electronics retailer

- electronics retailer: online market intelligence
  - comprehensive overview of the market
    - daily information on price, shipping costs, trends, product mix
    - by product, geographical region, or competitor
  - thousands of products
  - hundreds of competitors

- nowadays: specialized companies
  - mostly manual, sampling
  - large cost
Scenario ②: Hotel Agency

- online travel agency
- best price guarantee
- prices of competing agencies
- average market price
Scenario 3: Hedge Fund

- house price index
  - published in regular intervals by national statistics agency
  - affects share values of various industries
- hedge fund:
  - online market intelligence to predict the house price index
Scenario ④: Tenders for construction company

- tenders from all over the world
- existing aggregators
- expensive, often incomplete
- yet need to be published (online) by law in most countries