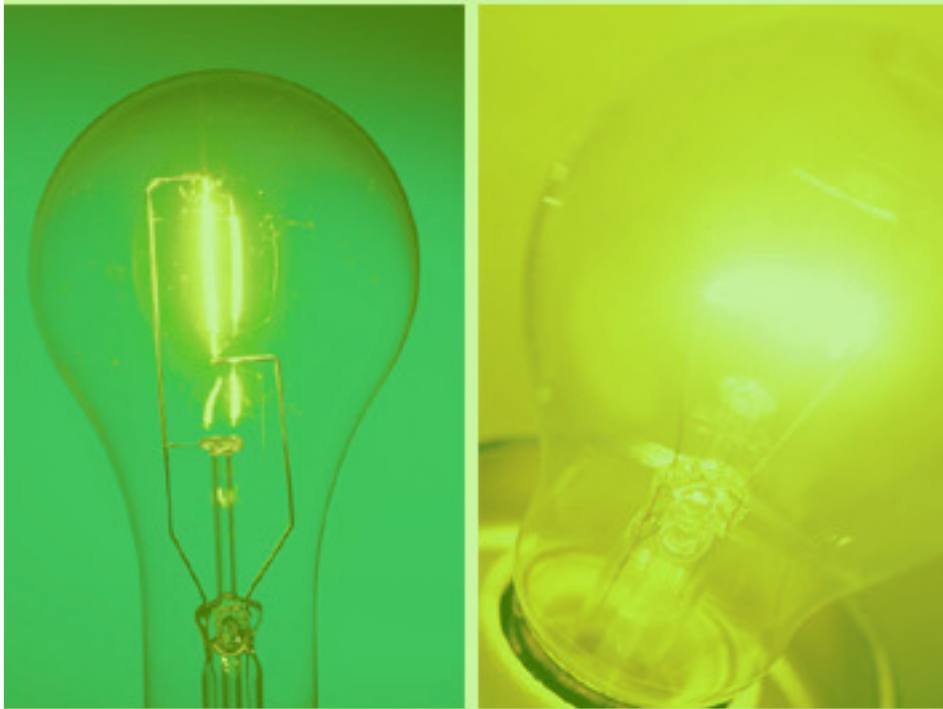


Automating the Extraction of Data from HTML Tables with Unknown Structure



Stefan Rümmele
PSWIE SS 2005



Overview

- **Application range**
- **Problems**
- **Extraction ontologies**
- **Location & extraction solution**



Application range

Source:

- Various websites with HTML tables
- Tables have different and unknown structure
- All tables relate to a specific domain of interest

Target:

- Predefined relational schema



Example: car advertisement

Car	Year	Make	Model	...
0001	1999	Pontiac	Firebird	...
0002	2000	Acura	RL 3.5	...
0003	2002	Honda	Accord EX	...
...

Car	Feature
0001	Blue
...	...
0003	White
0003	Air Conditioning

Viewing and querying web car advertisements
through target schema:

{Car, Year, Make, Model, Mileage, Price, PhoneNr}

{Car, Feature}

Pre-Owned Inventory

To see a list of all our cars, trucks, vans and SUV's, [click here](#).

- Looking for a price quote? Check out our [Quick Quote Form](#).
- Need financing? Try our new [Pre-Approval Form](#).
- Check out our [Internet-Only Specials](#).

To search for a specific vehicle or model, use our easy search engine below. Our inventory changes daily, so drop us an email or give us a call if you don't see the car you want. We will make sure you find your dream car! You searched for:

- All vehicles available.

66 matches found. Vehicles 1 to 25 shown.

Year	Make and Model	Price	Miles	Exterior	Photo
<input type="checkbox"/>	1999 Pontiac Firebird	Contact Us	32,883	Blue	
<input type="checkbox"/>	2000 Acura RL 3.5	\$23,988	36,657	Silver	
<input type="checkbox"/>	2002 Honda Accord EX	\$21,988	13,875	White	
<input type="checkbox"/>	2002 Honda Passport	\$20,998	10,410	Black	
<input type="checkbox"/>	2002 Acura PDX Type-S	\$20,988	14,208	Red	
<input type="checkbox"/>	2000 Chevrolet Cavalry	\$13,995	45,297	White	
<input type="checkbox"/>	2001 Honda Accord Value Package	\$13,995	31,710	Silver	
<input type="checkbox"/>	2001 Chevrolet Silverado C1500	\$13,988	28,022	Pewter	

Show checked vehicles

New search

Show 25 more

All vehicles subject to prior sale. We reserve the right to make changes without notice, and are not responsible for errors.

Bob Howard Honda
14137 Broadway Extension
Oklahoma City, OK 73013

Toll Free: 1-877-944-2842
Phone: 405-936-8666
Fax: 405-936-8674

E-mail: sales@bobhowardauto.dealerspace.com

Pre-Owned Inventory

At Howard Auto Group we have created an Internet sales department to give our customers an alternative buying experience. Once you have found a vehicle you like we will be glad to **give you our lowest no haggle price right up front!** Then if you like that price you can complete the transaction with your Internet manager. He can also quote you a payment, interest rate and if you have a trade in give you an evaluation of your trade in. Remember the Internet department is designed to provide the fastest and friendliest service to the Internet user and to ensure a totally different buying experience! If you have any questions please feel free to give a call at **405-936-8666** or toll free **877-944-2842** or drop us an e-mail. Your Internet sales staff at Howard AutoNet is waiting to help you. [Kyle](#), [Brendan](#), [Shane](#), [Dk](#), [Rory](#), [Mitesh](#), [Nic](#), [Steve](#), [Karrim](#), [Jay](#), [Traci](#), and [Ryan](#).

[Schedule a test drive](#)

[Send Me More Information](#)

2002 Honda Accord EX \$21,988

Features

- Air Conditioning
- Driver Side Air Bag
- Passenger Side Air Bag
- Anti-Lock Brakes
- AM/FM Cassette
- Security Features
- Alloy Wheels
- Automatic Transmission
- Bucket Seats
- Compact Disc Player
- Cruise Control
- Front Wheel Drive
- Intermittent Wipers
- Map Light



www.BobHowardAuto.com

Click on photo to enlarge

Price	\$21,988
Mileage	13,875 miles
Body Type	Car
Body Style	Coupe
Exterior	White
Transmission	Automatic
Engine	3.0L 6 cyl Fuel Injection
Fuel Type	Gas
Stock Number	350291A
VIN	1HGCG22562A018644

Vehicles

[Search New](#)
[Pre-Owned](#)
[Specials](#)
[Get A Quote](#)
[Financing](#)
[Vehicle Pricing](#)

[Finance](#)

[Specials](#)

[Contact](#)

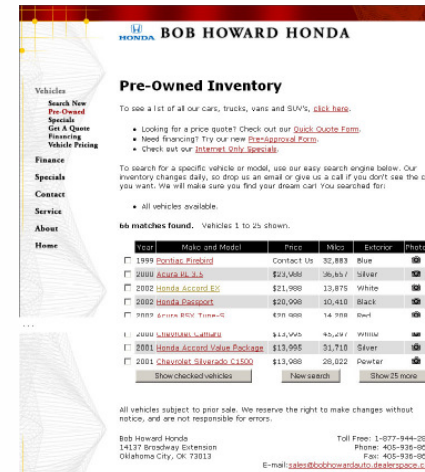
[Service](#)

[About](#)

[Home](#)

Location problems

- Multiple frames
- Tables for layout
- Table rows not part of the data
- Tables displayed piecemeal
- Tables spanning multiple pages
- No <table> tag



BOB HOWARD HONDA

Pre-Owned Inventory

To see a list of all our cars, trucks, vans and SUV's, [click here](#).

- Looking for a price quote? Check out our [Quick Quote Form](#).
- Need financing? Try our new [Financing Form](#).
- Check out our [Internet Only Specials](#).

To search for a specific vehicle or model, use our easy search engine below. Our inventory changes daily, so drop us an email or give us a call if you don't see the car you want. We will make sure you find your dream car! You searched for:

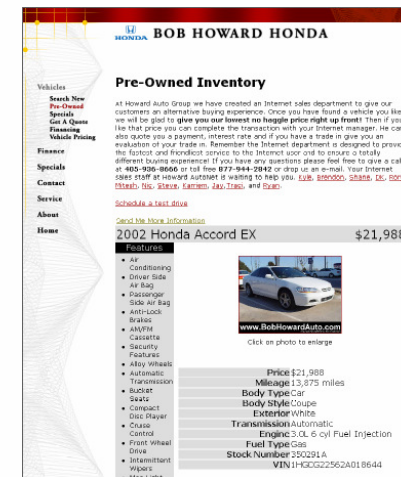
- All vehicles available.

14 matches found. Vehicles 1 to 25 shown.

Year	Make and Model	Price	Miles	Exterior	Color
1999	Porsche Boxster	Contact Us	32,883	Blue	
2008	Acura ILX	\$23,980	36,857	Silver	
2002	Honda Accord EX	\$21,980	13,875	White	
2002	Honda Passport	\$20,990	10,410	Black	
2003	Acura Integra Type-R	\$30,990	14,788	Black	
2001	Honda Accord Value Package	\$13,995	31,710	Silver	
2001	Chevrolet Silverado C1500	\$13,980	20,022	Power	

Bob Howard Honda
14137 Broadway Extension
Oklahoma City, OK 73013

Toll Free: 1-877-844-2042
Phone: 405-938-8006
Fax: 405-938-8074
E-mail: sales@bobhowardhonda.com



BOB HOWARD HONDA

Pre-Owned Inventory

At Howard Auto Group we have created an Internet sales department to give our customers an alternative buying experience. Once you have found a vehicle you like we will be glad to give you our lowest no haggle price right up front! Then if you like that price you can complete the transaction with your internet manager. He can also quote you a payment, interest rate and if you have a trade in give you an evaluation of your trade in. Remember the Internet department is designed to provide the fastest and most direct service to the internet user and to ensure a totally different buying experience! If you have any questions please feel free to give a call at 405-938-8006 or toll free 877-844-2042 or drop us an e-mail. Your internet sales staff at Howard AutoGroup is waiting to help you. [LUB](#), [INTERNO](#), [SHUTTLE](#), [ID](#), [COPY](#), [PRINT](#), [DL](#), [SHARE](#), [SEARCH](#), [ADD TO CART](#), and [GO](#).

[Schedule a test drive](#)

Send Me More Information

2002 Honda Accord EX \$21,980

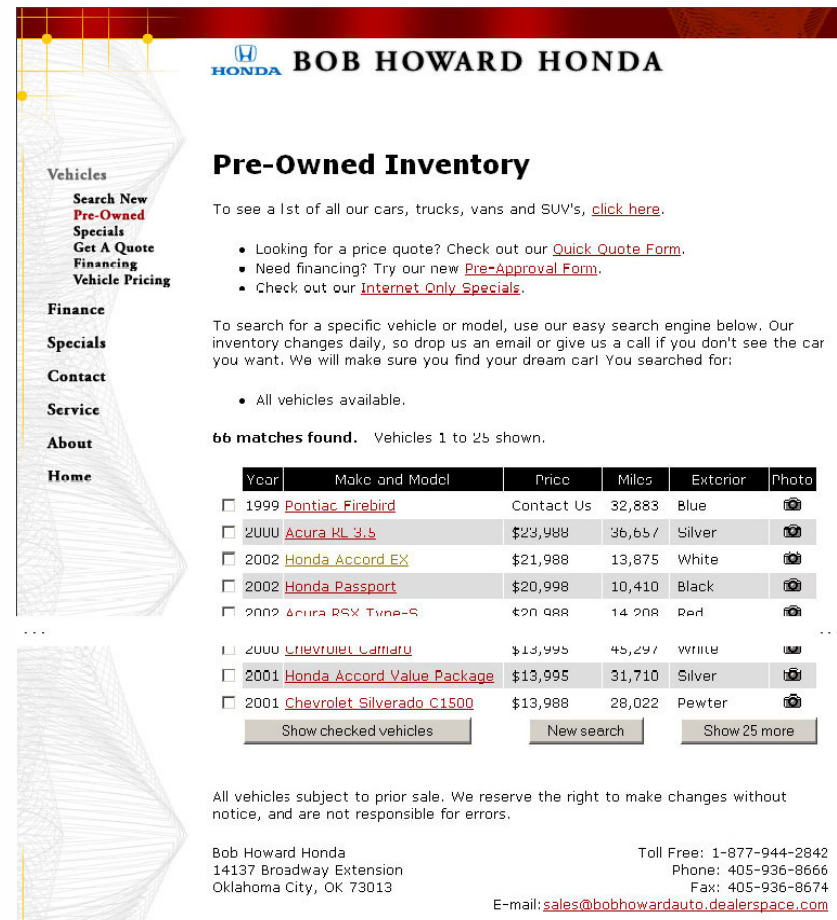
Features:

- Air Conditioning
- Driver Side Air Bag
- Passenger Side Air Bag
- ABS
- Brakes
- Alloy Wheels
- Cassette
- Security Features
- Alloy Wheels
- Automatic Transmission
- Bucket Seats
- Compact Disc Player
- Cruise Control
- Front Wheel Drive
- Intermittent Wipers
- Power Locks

Price \$21,980
Mileage 13,875 miles
Body Type Car
Body Style Coupe
Exterior White
Transmission Automatic
Engine 3.0L 6 cyl Fuel Injection
Fuel Type Gas
Stock Number 350291A
VIN 1HCGG2562A018644

Extraction problems 1/2

- Merged attributes/values
- Subsets
- Synonyms
- Extra information
- Linked information
- Externally factored data
- Unexpected multiple values



The screenshot shows the 'Pre-Owned Inventory' page for Bob Howard Honda. The page features a navigation menu on the left with links for Vehicles, Finance, Specials, Contact, Service, About, and Home. The main content area includes a search engine, a list of 66 matches found, and contact information for the dealership.

HONDA BOB HOWARD HONDA

Pre-Owned Inventory

To see a list of all our cars, trucks, vans and SUV's, [click here](#).

- Looking for a price quote? Check out our [Quick Quote Form](#).
- Need financing? Try our new [Pre-Approval Form](#).
- Check out our [Internet Only Specials](#).

To search for a specific vehicle or model, use our easy search engine below. Our inventory changes daily, so drop us an email or give us a call if you don't see the car you want. We will make sure you find your dream car! You searched for:

- All vehicles available.

66 matches found. Vehicles 1 to 25 shown.

Year	Make and Model	Price	Miles	Exterior	Photo
<input type="checkbox"/>	1999 Pontiac Firebird	Contact Us	32,883	Blue	
<input type="checkbox"/>	2000 Acura HL 3.5	\$23,988	36,657	Silver	
<input type="checkbox"/>	2002 Honda Accord EX	\$21,988	13,875	White	
<input type="checkbox"/>	2002 Honda Passport	\$20,998	10,410	Black	
<input type="checkbox"/>	2002 Acura RSX Type-S	\$20,988	14,208	Red	
<input type="checkbox"/>	2000 Chevrolet Camaro	\$13,995	45,297	White	
<input type="checkbox"/>	2001 Honda Accord Value Package	\$13,995	31,710	Silver	
<input type="checkbox"/>	2001 Chevrolet Silverado C1500	\$13,988	28,022	Pewter	

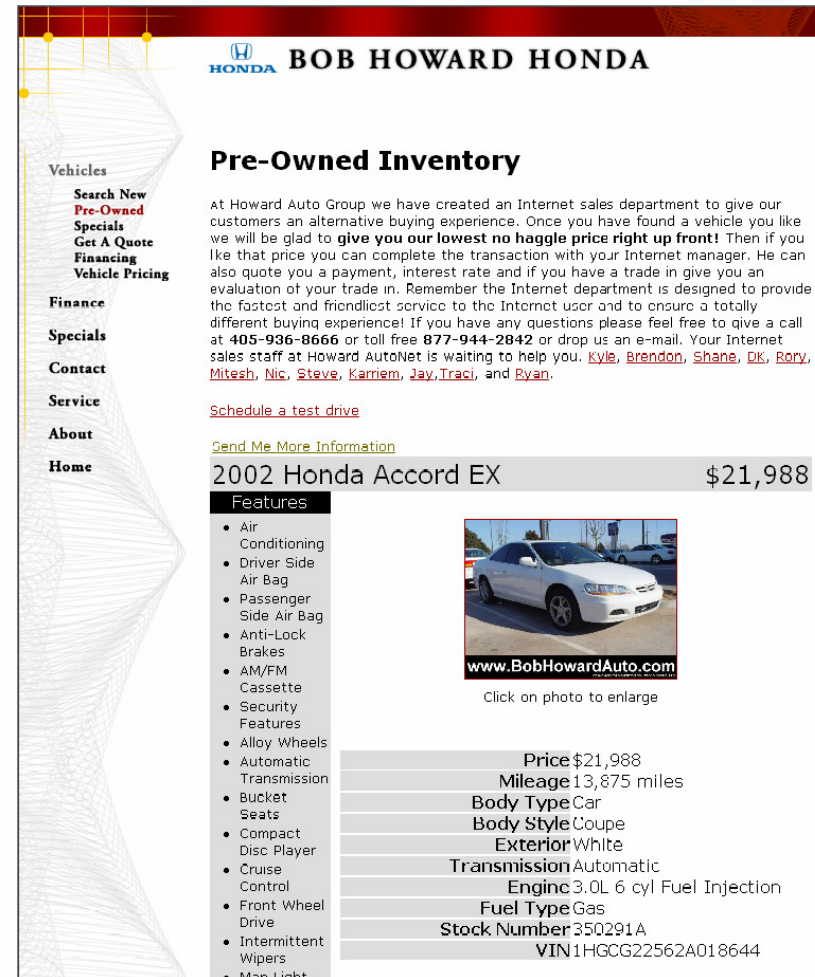
All vehicles subject to prior sale. We reserve the right to make changes without notice, and are not responsible for errors.

Bob Howard Honda
14137 Broadway Extension
Oklahoma City, OK 73013

Toll Free: 1-877-944-2842
Phone: 405-936-8666
Fax: 405-936-8674
E-mail: sales@bobhowardauto.dealerspace.com

Extraction problems 2/2

- List table
- Position of attributes
- Duplicate data
- Missing information
- Attribute as value



HONDA BOB HOWARD HONDA

Pre-Owned Inventory

At Howard Auto Group we have created an Internet sales department to give our customers an alternative buying experience. Once you have found a vehicle you like we will be glad to **give you our lowest no haggle price right up front!** Then if you like that price you can complete the transaction with your Internet manager. He can also quote you a payment, interest rate and if you have a trade in give you an evaluation of your trade in. Remember the Internet department is designed to provide the fastest and friendliest service to the Internet user and to ensure a totally different buying experience! If you have any questions please feel free to give a call at **405-936-8666** or toll free **877-944-2842** or drop us an e-mail. Your Internet sales staff at Howard AutoNet is waiting to help you. [Kyle](#), [Brendon](#), [Shane](#), [DK](#), [Rory](#), [Mitesh](#), [Nic](#), [Steve](#), [Karnem](#), [Jay](#), [Traci](#), and [Ryan](#).


[Schedule a test drive](#)

[Send Me More Information](#)

2002 Honda Accord EX \$21,988

Features

- Air Conditioning
- Driver Side Air Bag
- Passenger Side Air Bag
- Anti-Lock Brakes
- AM/FM Cassette
- Security Features
- Alloy Wheels
- Automatic Transmission
- Bucket Seats
- Compact Disc Player
- Cruise Control
- Front Wheel Drive
- Intermittent Wipers
- Map Light



www.BobHowardAuto.com

Click on photo to enlarge

Price	\$21,988
Mileage	13,875 miles
Body Type	Car
Body Style	Coupe
Exterior	White
Transmission	Automatic
Engine	3.0L 6 cyl Fuel Injection
Fuel Type	Gas
Stock Number	350291A
VIN	1HGCG22562A018644



Extraction ontologies

Definition:

“An ontology is a specification of a conceptualization.” (Tom Gruber)

Content:

- Object/Relationship-model instance
- Data frame for each object set

Purpose:

Formal defined system that serves as a wrapper for a narrow domain of interest.



Extraction ontology - example

```
01. Car [-> object];
02. Car [0:1] has Year [1:*];
03. Car [0:1] has Make [1:*];
04. Car [0:1] has Model [1:*];
05. Car [0:1] has Mileage [1:*];
06. Car [0:*] has Feature [1:*];
07. Car [0:1] has Price [1:*];
08. PhoneNr [1:*] is for Car [0:1];
09. Year matches [4]
10.     constant {extract "\d{2}";
11.         context "\b'[4-9]\d\b";
12.         substitute "^" -> "19"; },
13.     ...
14. Mileage matches [8]
15.     ...
16.     keyword "\bmiles\b", "\bmi\.", "\bmi\b",
17.         "\bmileage\b", "\bodometer\b",
18.     ...
```



Solution (1/3)

1. Locate the table of interest.

- Table on the main page
- Tables on linked pages
- Achieved using a heuristic with several rules based on the ontology

2. Form attribute-value pairs.

```
{<Make: ACURA>, <Model: legend>, <Year: 1992>,  
<Colour: grey>, <Price: $9500>, <Auto: Yes>,  
<Air Cond.: No>, <AM/FM: Yes>, <CD: No>}
```



Solution (2/3)

3. Adjust attribute-value pairs.

- Attribute-value pairs from linked tables
- Process boolean indicators

```
Make: ACURA; Model: legend; Year: 1992;  
Colour: grey; Price: $9500; Auto; AM/FM;
```

4. Analyze extraction patterns.

- Applying the extraction ontology

```
{<Car: 0011>, <Year: 1992>, <Make: ACURA>,  
  <Model: legend>, <Mileage: >,  
  <Price: $9500>, <PhoneNr: >},  
{<Car: 0011>, <Feature: grey>},  
{<Car: 0011>, <Feature: Auto>},  
{<Car: 0011>, <Feature: AM/FM>}
```



Solution (3/3)

5. Infer Mappings.

- Transformations needed in steps 2-4 are recorded**
- Mapping is produced out of this information**
- Queries on the target schema can be translated to a query on the source**
- Result contains additional values not recognized by the ontology**