

All specs

General	
Weight	5.8 oz
Width	3.3 in
Depth	0.9 in
Height	2.2 in
Body material	Stainless steel
Main Features	
Sensor resolution	3.2 megapixels
Optical sensor type	CCD
Effective sensor resolution	3,200,000 pixels
Gross sensor resolution	3,300,000 pixels
Optical sensor size	1/2.7 in
Light sensitivity	ISO 50, ISO 100, ISO 200, ISO 400
Digital zoom	3.2 x
Shooting modes	Frame movie mode
Shooting programs	Macro, Landscape, Stitch assist
Special effects	Sepia, Vivid, Neutral, Black & White, Low Sharpening
Max shutter speed	1/1500 sec
Min shutter speed	15 sec

Updated version:  
July 23, 2006

# Table Extraction Using Spatial Reasoning on the CSS2 Visual Box Model

AAAI-06

*Boston, July 18, 2006*

Wolfgang Gatterbauer

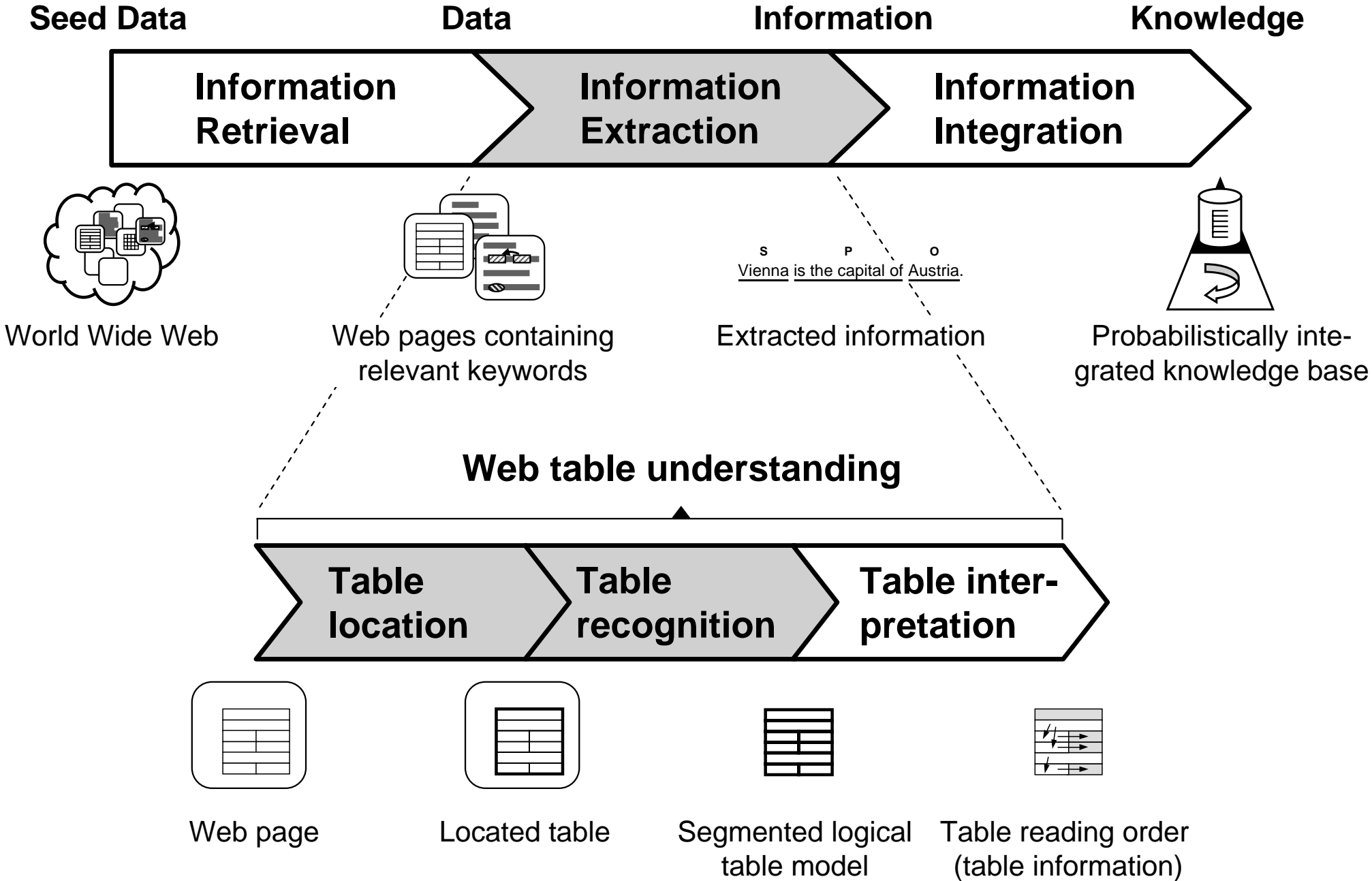
Paul Bohunsky



Database & Artificial Intelligence Group  
Vienna University of Technology

# KNOWLEDGE ACQUISITION PROCESS FROM THE WEB

 Table extraction



# EXAMPLE WEB INFORMATION IN TABLES

Government Austria [Top of Page](#)

Country name

## People @ DBAI

**Administrators**

[Ilse Epper](#)

[Eva Nedetzka](#)

[Therese Schuster](#)

**Professors**

[Georg G. Gruber](#)

[Jürgen Döllinger](#)

[Reinhard Schuster](#)

**Faculty**

[Robert B. Stammers](#)

[Oliver Franke](#)

[Thomas J. Hayes](#)

[Marcus Hutter](#)

[Nysret Memari](#)

[Katrin Schwaninger](#)

[Fang Wei](#)

**Details**

General

Product Type: Digital camera

Width

Depth

Height

Weight

Body Material

Miscellaneous

Cables Included

Included Accessories

Min Operating Temperature

Max Operating Temperature

Power

Power Device

Software

Software

Display

Type

Display Form Factor

Display Format

Battery

Type

Included Qty

Capacity

Max Recharge Cycles

Lens System

Type

Focal Length

Focal Length Equivalent to Camera

Camera

Focus Adjustment

Auto Focus

Auto Focus Points (Zones)

Min Focus Range

Macro Focus Range

Lens Aperture

Optical Zoom

Zoom Adjustment

Lens Construction

Features

Viewfinder

Viewfinder Type

Viewfinder Frames


LED Information

Main Features

Resolution

Try it **risk free** for 30 days
**WACOM**  
intuos<sub>3</sub>

### Canon PowerShot SD110 digital camera specifications

	Canon PowerShot SD110
Image	
More information	<a href="#">Announced 09-Feb-04</a> <a href="#">All Canon products</a>
Discussion	<a href="#">Canon Talk Forum</a> <a href="#">Find related discussion</a>
Owners opinions	<span style="color: red;">★★★★★</span> <a href="#">Read owners opinions (5)</a> <a href="#">Post / Edit your opinion</a>
Support this site by purchasing from our affiliate merchants	<a href="#">Click here to check price / order</a>
Format	Ultra Compact
Price (street)	US\$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (W)	35 mm

# PROBLEM OF CODE-BASED TABLE RECOGNITION

Web page with interesting tabular information

Try it **risk free** for 30 days

**WACOM**  
**intuos.3**

## Canon PowerShot SD110 digital camera specifications


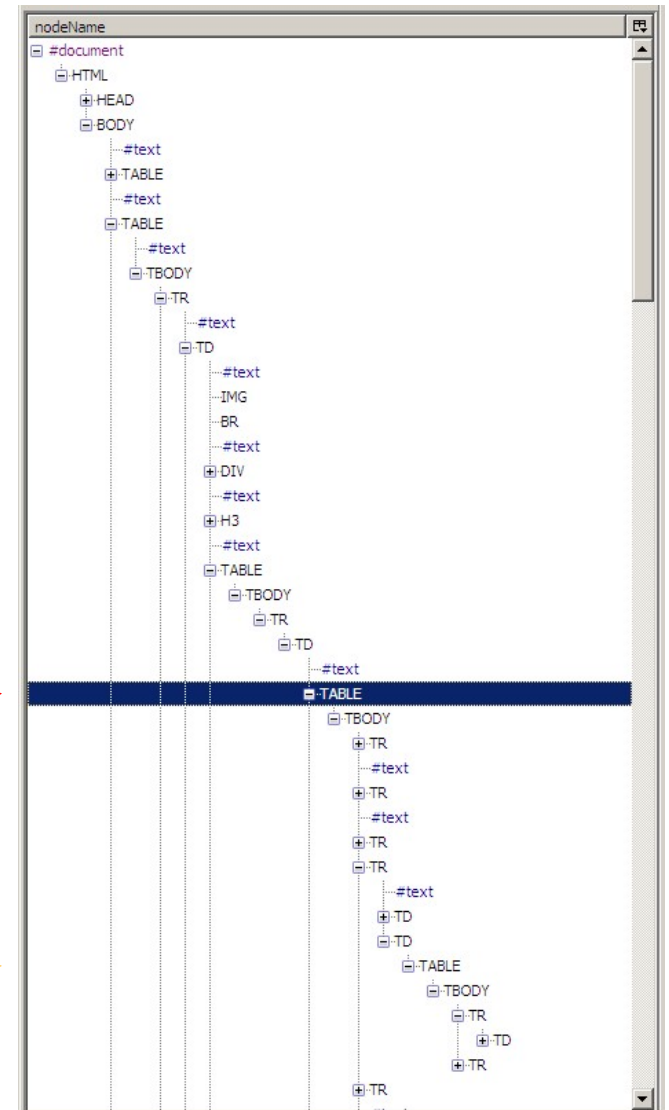
Image	
More information	<a href="#">Announced 09-Feb-04</a> <a href="#">All Canon products</a>
Discussion	<a href="#">Canon Talk Forum</a> <a href="#">Find related discussion</a>
Owners opinions	<p>★★★★☆</p> <a href="#">Read owners opinions (5)</a> <a href="#">Post / Edit your opinion</a>
Support this site by purchasing from our affiliate merchants	<a href="#">Click here to check price / order</a>
Format	Ultra Compact
Price (street)	US\$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (W)	35 mm

Table in the DOM tree



# PROBLEM OF CODE-BASED TABLE RECOGNITION

Nested non-leaf <TABLE> tables


Digital Photography Review™  
dpreview.com

- News
- Reviews
- Cameras
- Timeline
- Buying Guide
- Galleries
- Forums
- Search
- Learn
- Glossary
- Feedback
- Newsletter
- Links
- Support Us
- About

Try it **risk free** for 30 days

**WACOM**  
**intuos.3**

### Canon PowerShot SD110 digital camera specifications


Image	
More information	<ul style="list-style-type: none"> <li>➤ <a href="#">Announced 09-Feb-04</a></li> <li>➤ <a href="#">All Canon products</a></li> </ul>
Discussion	<ul style="list-style-type: none"> <li>➤ <a href="#">Canon Talk Forum</a></li> <li>➤ <a href="#">Find related discussion</a></li> </ul>
Owners opinions	<p>★★★★★</p> <p><a href="#">Read owners opinions (5)</a> <a href="#">Post / Edit your opinion</a></p>
Support this site by purchasing from our affiliate merchants	<p><a href="#">Click here to check price / order</a></p>
Format	Ultra Compact
Price (street)	US\$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (W)	35 mm

Sequentially aligned <TABLE> tables

Details

General	
Product Type	Digital camera
Width	8.5 cm
Depth	2.4 cm
Height	5.6 cm
Weight	165 g
Body Material	Stainless steel
Miscellaneous	
Cables Included	1 x A/V cable 1 x USB cable
Included Accessories	Wrist strap
Min Operating Temperature	0 °C
Max Operating Temperature	40 °C
Power	
Power Device	Battery charger - external
Software	
Software	Drivers & Utilities, Canon PhotoStitch, Canon ZoomBrowser EX, ArcSoft PhotoImpression, ArcSoft VideoImpression, Canon ImageBrowser
Display	
Type	LCD display - TFT active matrix - 1.5" - colour
Display Form Factor	Built-in
Display Format	118,000 pixels
Battery	
Type	1 x camera battery - rechargeable - Lithium Ion
Included Qty	1
Capacity	790 mAh
Max Recharge Cycles	300
Lens System	
Type	Zoom lens
Focal Length	5.4 mm - 10.8 mm
Focal Length Equivalent to 35mm Camera	35 - 70mm
Focus Adjustment	Automatic
Auto Focus	TTL contrast detection
Auto Focus Points (Zones)	9
Min Focus Range	47 cm
Macro Focus Range	10-47cm
Lens Aperture	F/2.8-3.9
Optical Zoom	2 x
Zoom Adjustment	Motorised drive
Lens Construction	6 group(s) / 6 element(s)
Features	Built-in lens shield, aspherical lens
Viewfinder	
Viewfinder Type	Optical - real-image zoom
Viewfinder Frames	Autofocus frame
LED Information	Flash ready, autofocus ready
Main Features	
Resolution	3.2 Megapixel

# VENTrec: Visualized Element Nodes Table RECOgnition

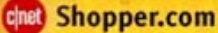


Digital  
Photography  
Review™  
dpreview.com


## SHOP TIL YOU PRICE DROP

Track daily price decreases on your favorite digital cameras

Digital cameras  
cameras Digital cam



**Canon PowerShot SD110 digital camera specifications**

	Canon PowerShot SD110
Image	
More information	<a href="#">↗ Announced 09-Feb-04</a> <a href="#">↗ All Canon products</a>
Discussion	<a href="#">↗ Canon Talk Forum</a> <a href="#">↗ Find related discussion</a>
Owners opinions	<p>★★★★★</p> <p><a href="#">Read owners opinions (5)</a>  <a href="#">Post / Edit your opinion</a></p>
Support this site by purchasing from our affiliate merchants	<a href="#">Click here to check price / order</a>
Format	Ultra Compact
Price (street)	US\$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (mm)	35 mm

News  
Reviews  
Cameras  
Timeline  
Buying Guide  
Galleries  
Forums  
Search  
Learn  
Glossary  
Feedback  
Newsletter  
Links  
Support Us  
About

# VENTrec: Visualized Element Nodes Table RECOgnition

[Empty Header Row]	
[Empty Sidebar Row 1]	[Empty Main Table Row 1]
[Empty Sidebar Row 2]	[Empty Main Table Row 2]
[Empty Sidebar Row 3]	[Empty Main Table Row 3]
[Empty Sidebar Row 4]	[Empty Main Table Row 4]
[Empty Sidebar Row 5]	[Empty Main Table Row 5]
[Empty Sidebar Row 6]	[Empty Main Table Row 6]
[Empty Sidebar Row 7]	[Empty Main Table Row 7]
[Empty Sidebar Row 8]	[Empty Main Table Row 8]
[Empty Sidebar Row 9]	[Empty Main Table Row 9]
[Empty Sidebar Row 10]	[Empty Main Table Row 10]
[Empty Sidebar Row 11]	[Empty Main Table Row 11]
[Empty Sidebar Row 12]	[Empty Main Table Row 12]
[Empty Sidebar Row 13]	[Empty Main Table Row 13]
[Empty Sidebar Row 14]	[Empty Main Table Row 14]
[Empty Sidebar Row 15]	[Empty Main Table Row 15]
[Empty Sidebar Row 16]	[Empty Main Table Row 16]
[Empty Sidebar Row 17]	[Empty Main Table Row 17]
[Empty Sidebar Row 18]	[Empty Main Table Row 18]
[Empty Sidebar Row 19]	[Empty Main Table Row 19]
[Empty Sidebar Row 20]	[Empty Main Table Row 20]
[Empty Sidebar Row 21]	[Empty Main Table Row 21]
[Empty Sidebar Row 22]	[Empty Main Table Row 22]
[Empty Sidebar Row 23]	[Empty Main Table Row 23]
[Empty Sidebar Row 24]	[Empty Main Table Row 24]
[Empty Sidebar Row 25]	[Empty Main Table Row 25]
[Empty Sidebar Row 26]	[Empty Main Table Row 26]
[Empty Sidebar Row 27]	[Empty Main Table Row 27]
[Empty Sidebar Row 28]	[Empty Main Table Row 28]
[Empty Sidebar Row 29]	[Empty Main Table Row 29]
[Empty Sidebar Row 30]	[Empty Main Table Row 30]
[Empty Sidebar Row 31]	[Empty Main Table Row 31]
[Empty Sidebar Row 32]	[Empty Main Table Row 32]
[Empty Sidebar Row 33]	[Empty Main Table Row 33]
[Empty Sidebar Row 34]	[Empty Main Table Row 34]
[Empty Sidebar Row 35]	[Empty Main Table Row 35]
[Empty Sidebar Row 36]	[Empty Main Table Row 36]
[Empty Sidebar Row 37]	[Empty Main Table Row 37]
[Empty Sidebar Row 38]	[Empty Main Table Row 38]
[Empty Sidebar Row 39]	[Empty Main Table Row 39]
[Empty Sidebar Row 40]	[Empty Main Table Row 40]
[Empty Sidebar Row 41]	[Empty Main Table Row 41]
[Empty Sidebar Row 42]	[Empty Main Table Row 42]
[Empty Sidebar Row 43]	[Empty Main Table Row 43]
[Empty Sidebar Row 44]	[Empty Main Table Row 44]
[Empty Sidebar Row 45]	[Empty Main Table Row 45]
[Empty Sidebar Row 46]	[Empty Main Table Row 46]
[Empty Sidebar Row 47]	[Empty Main Table Row 47]
[Empty Sidebar Row 48]	[Empty Main Table Row 48]
[Empty Sidebar Row 49]	[Empty Main Table Row 49]
[Empty Sidebar Row 50]	[Empty Main Table Row 50]

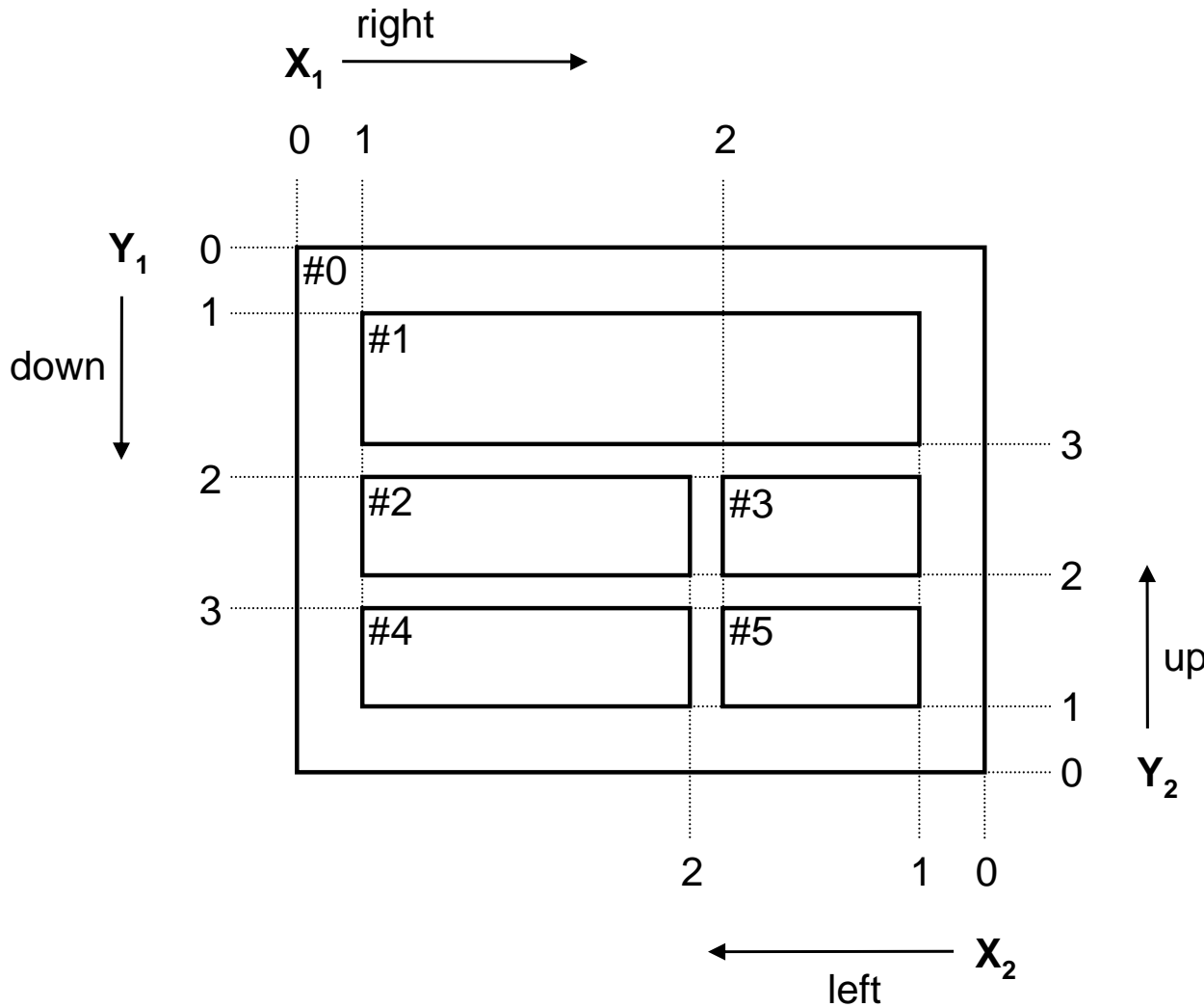
# VENTrec: Visualized Element Nodes Table RECOgnition




# VENTrec: Visualized Element Nodes Table RECOgnition

	Canon PowerShot SD110
Image	
More information	Announced 09-Feb-04 All Canon products
Discussion	Canon Talk Forum Find related discussion
Owners opinions	Read owners opinions (5) Post / Edit your opinion
Support this site by purchasing from our affiliate merchants	Click here to check price / order
Format	Ultra Compact
Price (street)	US\$100
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	R G B
Sensor manufacturer	Unknown

# SUPERIMPOSED MINIMAL DOUBLE TOPOGRAPHICAL GRID (DTG)



## Visualized Element Nodes

VEN	$X_1$	$Y_1$	$X_2$	$Y_2$
#0	0	0	0	0
#1	1	1	1	3
#2	1	2	2	2
#3	2	2	1	2
#4	1	3	2	1
#5	2	3	1	1

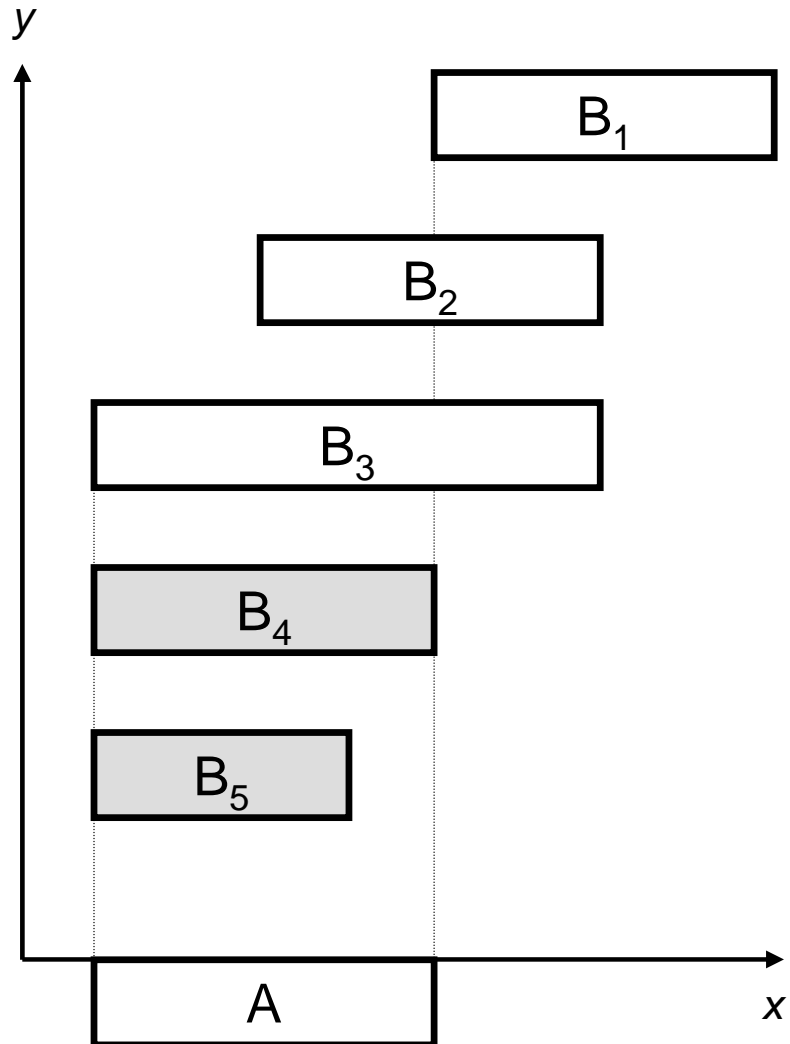
## Grid structure

$X_1$	coord.	$X_2$	coord.
0	20	0	860
1	100	1	780
2	540	2	500

$Y_1$	coord.	$Y_1$	coord.
0	20	0	660
1	100	1	580
2	300	2	420
3	460	3	260

# ADJACENCY & ALIGNMENT

For each of the 4 dimensions, adjacent boxes are categorized according to 5 alignment relationships



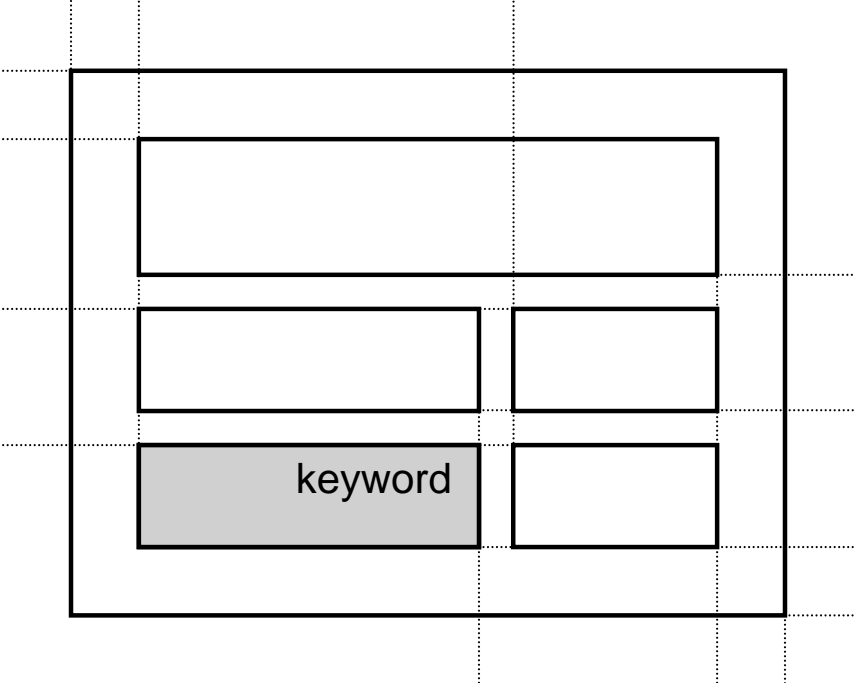
**X-neighbor relationship of B(A)**

**Allen's interval relations**  
(Allen 1983; Aiello 2002)

$B_1 = \text{no neighbor (A)}$	13 A before B 12 B before A 11 A meets B 10 B meets A
$B_2 = \text{step neighbor (A)}$	9 A overlaps B 8 B overlaps A
$B_3 = \text{bigger neighbor (A)}$	7 A during B 6 A starts B 5 A finishes B
$B_4 = \text{twin neighbor (A)}$	4 B equal A
$B_5 = \text{smaller neighbor (A)}$	3 B finishes A 2 B starts A 1 B during A

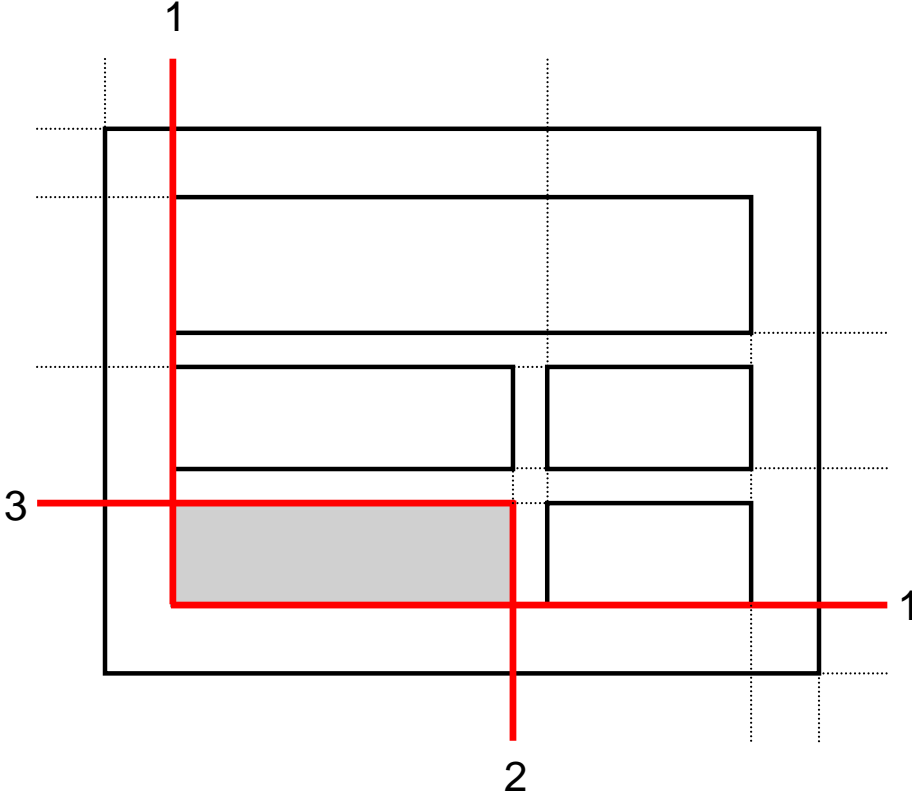
2 neighbor relationships are of relevance for the expansion algorithm

# WORKING OF THE EXPANSION ALGORITHM



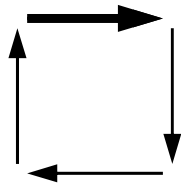
Keyword projected into  
Element Nodes

# WORKING OF THE EXPANSION ALGORITHM

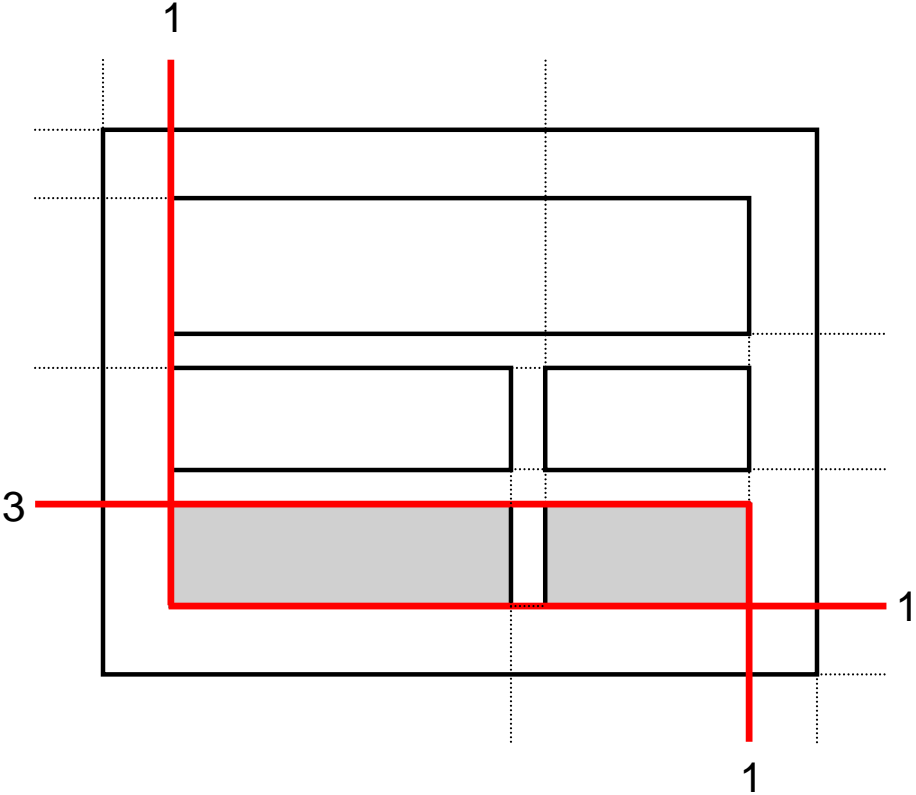


Keyword projected into  
Element Nodes

Circulating HyperBox  
Expansion Algorithm

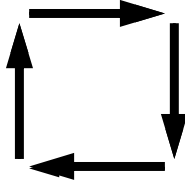


# WORKING OF THE EXPANSION ALGORITHM

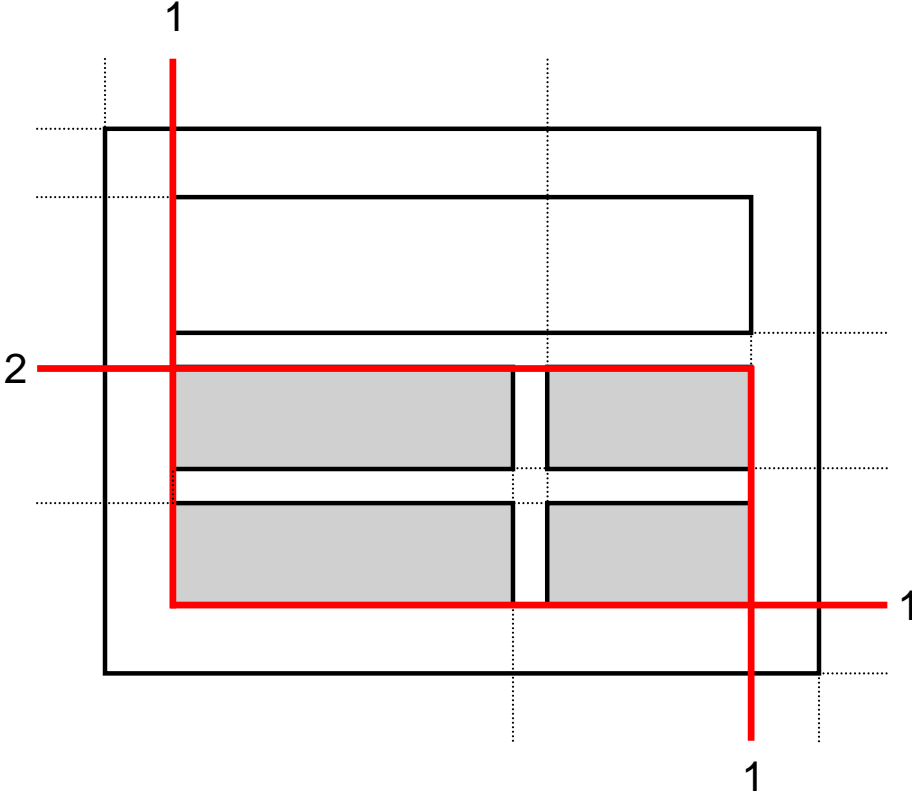


Keyword projected into  
Element Nodes

Circulating HyperBox  
Expansion Algorithm

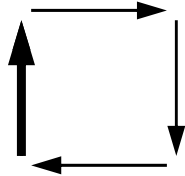


# WORKING OF THE EXPANSION ALGORITHM

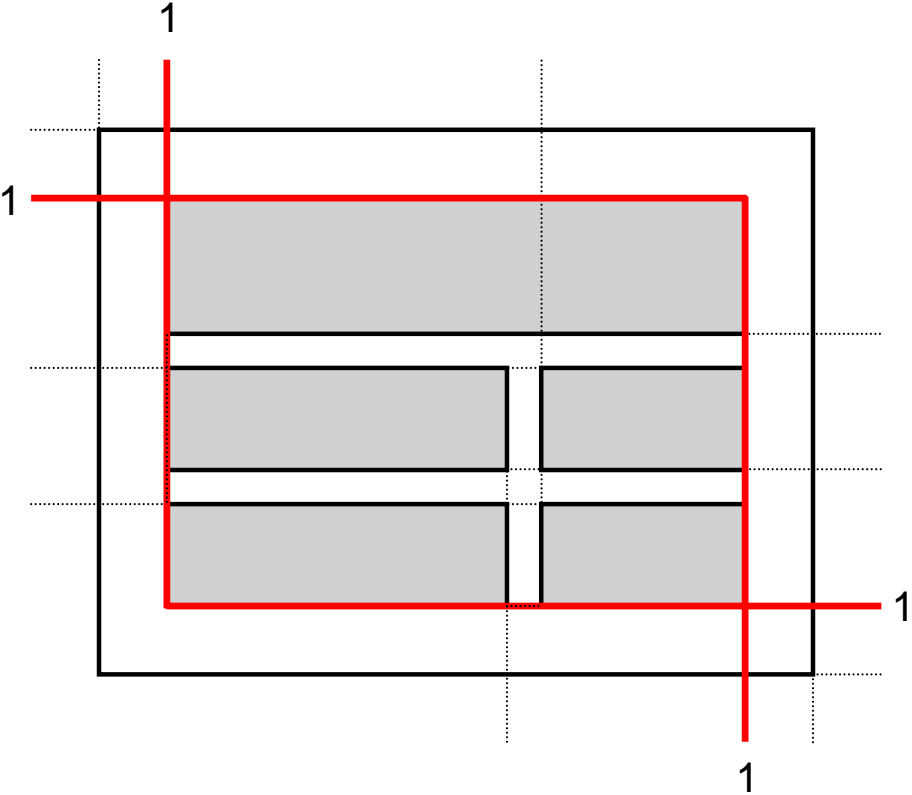


Keyword projected into  
Element Nodes

Circulating HyperBox  
Expansion Algorithm

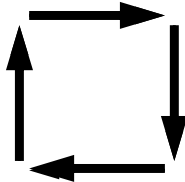


# WORKING OF THE EXPANSION ALGORITHM



Keyword projected into  
Element Nodes

Circulating HyperBox  
Expansion Algorithm



STOP



# TIME COMPLEXITY

n ... # of element nodes  
k ... # of keyword appearances

Positional data gathering

$$n$$

Circulating Expansion  
Algorithm

$$k\sqrt{n}$$

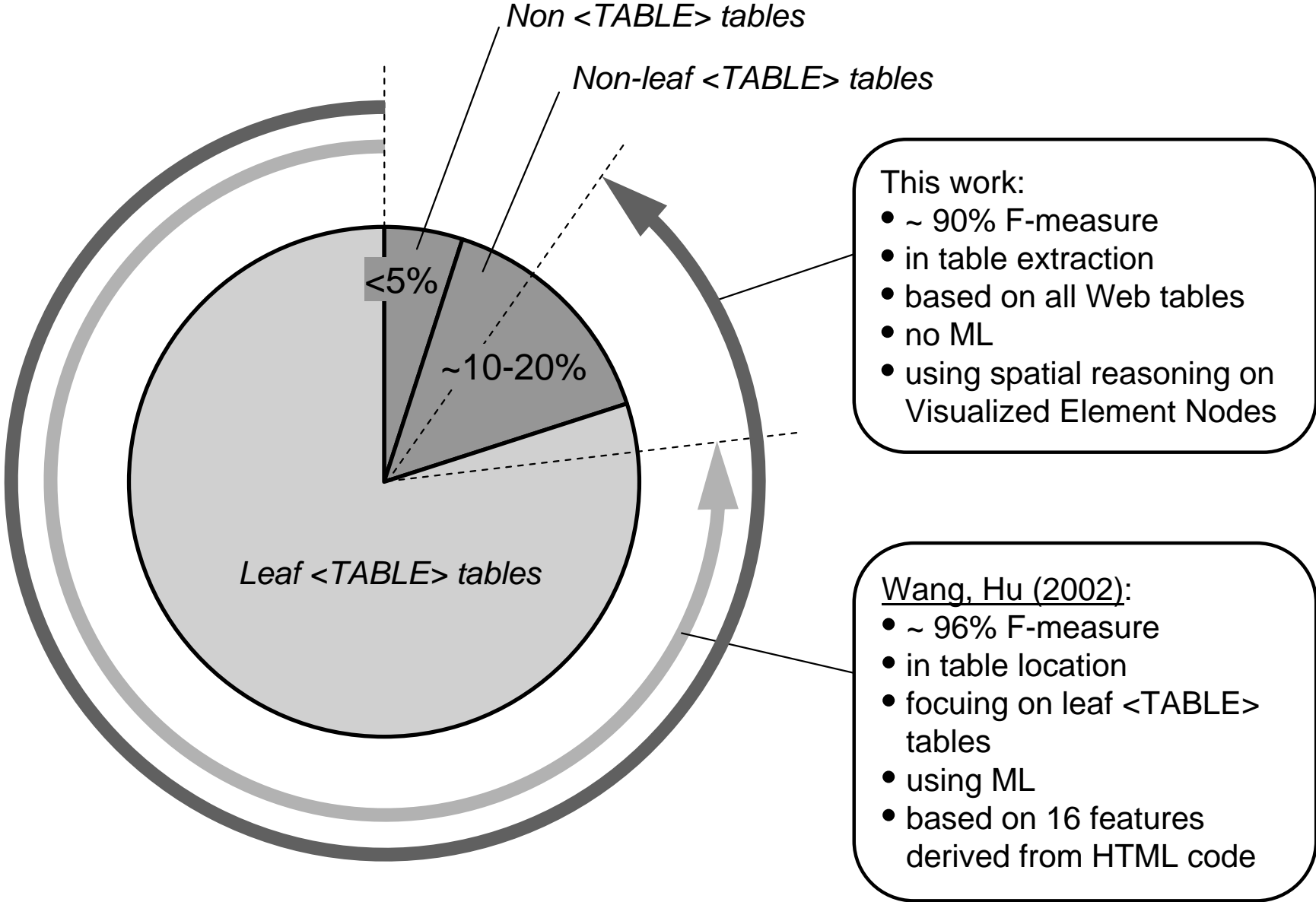
---

System

$$n + k\sqrt{n}$$

# PERFORMANCE COMPARISON

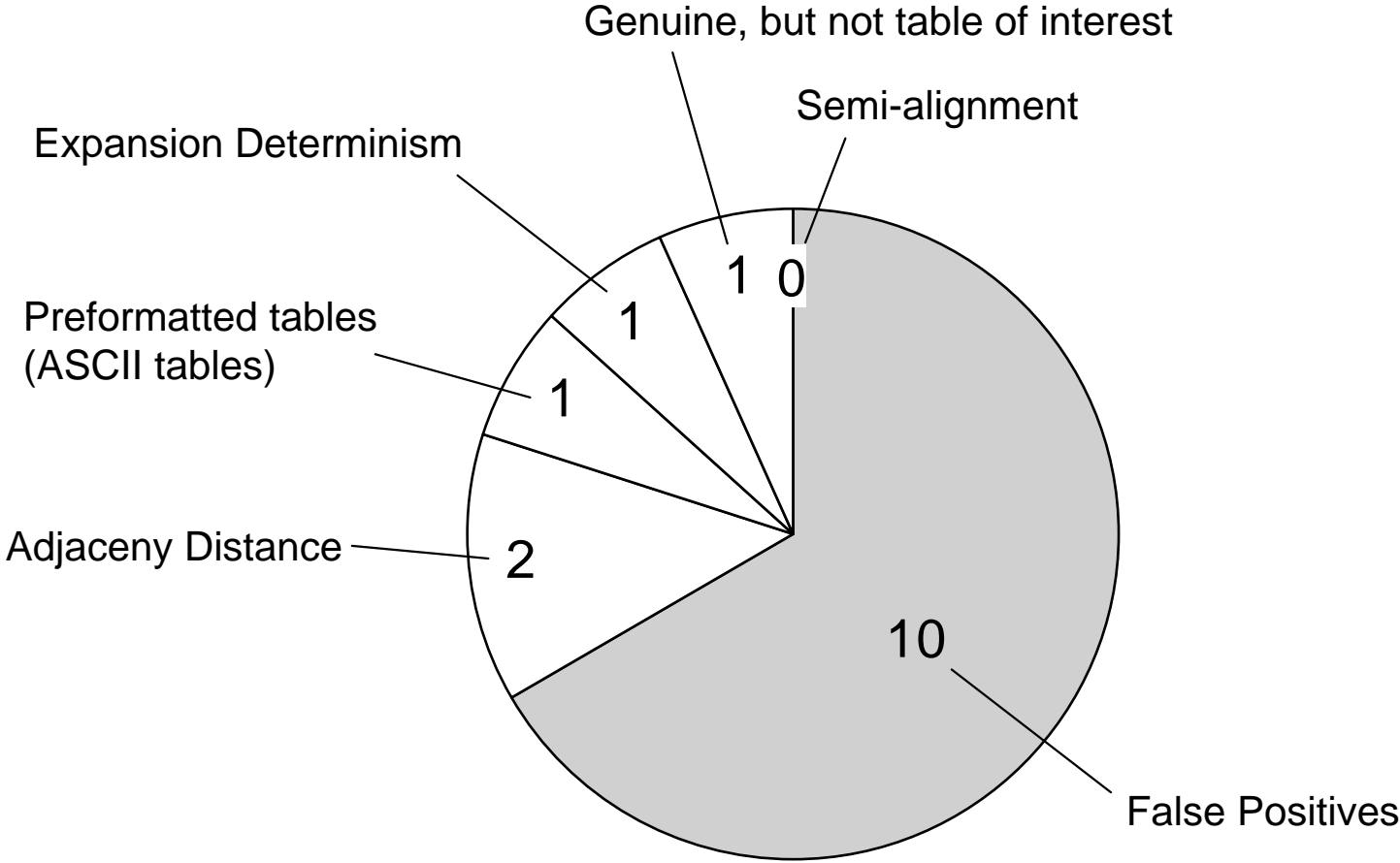
In Percent of Web tables



# DOMINANT REASONS FOR WRONG RESULTS

Total = 15

~10% of quantitative test set



# ONLINE VENTrec

Test it: <http://education.dbai.tuwien.ac.at/ventrec/>

Introduction  
Publication  
Evaluation  
**Test it !**  
About us  
Contact

## Online VENTrec

### Robust Table Extraction from Arbitrary Web pages



© 2006  
Team VENTrec

### Test it !

Online VENTrec

Here you can test whether the key logical table model


URL:

Keyword:

Introduction  
Publication  
Evaluation  
**Test it !**  
About us  
**Contact**

## Online VENTrec

### Robust Table Extraction from Arbitrary Web pages



### Contact Team VENTrec

Your email address:

Your message:

# SUMMARY

## Web table understanding

- As one approach for automated knowledge acquisition from the Web
- Table extraction from HTML code, however, is difficult
- Alternative: extracting tables from *rendered Web pages*

## Our approach

- Reasons on *spatial relationships*
- Between *Visualized Element Nodes*
- Working upon a *Double Topographical Grid* Structure

## Experiments

- ~90% F-measure yet without any form of learning
- Includes ~20% of Web tables missed by previous approaches

# NEXT STEPS

## VENTrec II

- Objectivation of table model
  - Ambiguity resolution
  - Heuristics for non-alignment
  - ML-improved version
  - Methodology for web table ground truthing
- } Theoretical foundations

## Table interpretation

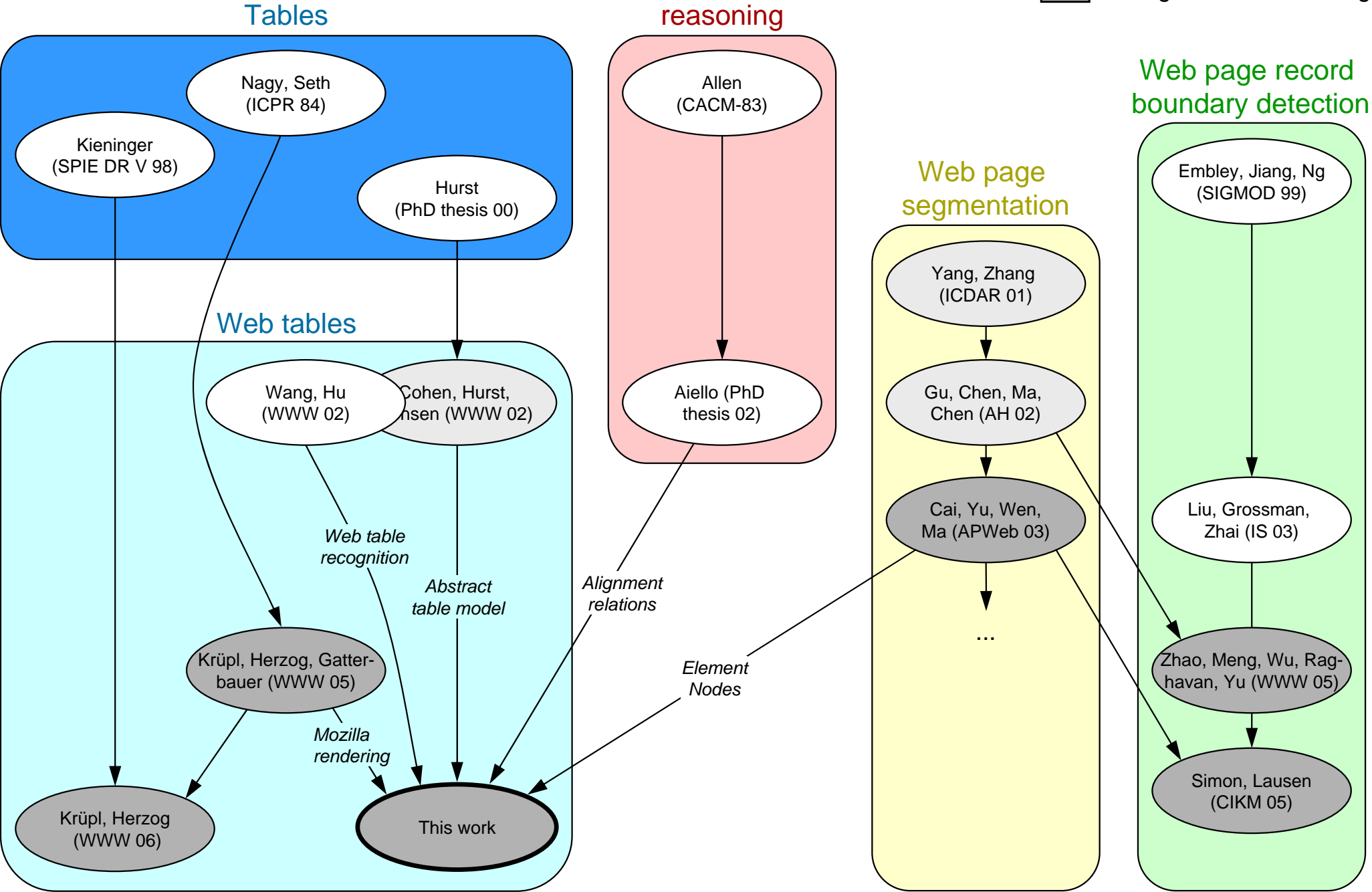
- Agent-based bottom-up learning of reading order = extraction of contained information

Visit us on the Web: <http://education.dbai.tuwien.ac.at/ventrec/>

# Backup

# RELATED LITERATURE

- Using visual concepts
- Using browser rendering





# RELATED LITERATURE

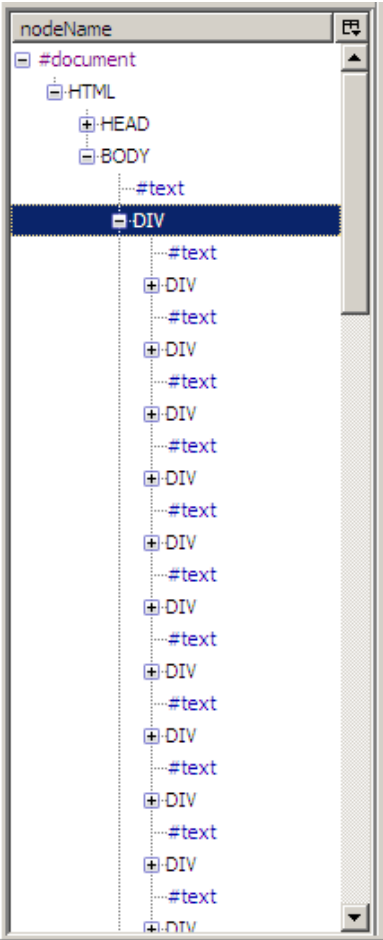
- Aiello, M. 2002. *Spatial Reasoning: Theory and Practice*. Ph.D. thesis, ILLC, University of Amsterdam.
- Allen, J.F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11):832-843.
- Cai, D.; Yu, S.; Wen, J.-R.; and Ma, W.-Y. 2003. Extracting content structure for web pages based on visual representation. In *Proc. APWeb'03*, 406-417. Springer.
- Cohen, W.W.; Hurst, M.; and Jensen, L.S. 2002. A flexible learning system for wrapping tables and lists in HTML documents. In *Proc. WWW'02*, 232-241. ACM.
- Embley, D.W.; Jiang, Y.S.; and Ng, Y.-K. 1999. Record-Boundary Discovery in Web Documents. In *Proc. SIGMOD'99*, 467-478, ACM.
- Gu, X; Chen, J; Ma, W.-Y.; and Chen, G. 2002. Visual Based Content Understanding towards Web Adaptation. In *Proc. AH'02*, 164-173, Springer.
- Hurst, M. 2000. *The Interpretation of Tables in Texts*. PhD. thesis, University of Edinburgh.
- Kieninger, T. G. 1998. Table structure recognition based on robust block segmentation. *Proc. SPIE Vol. 3305, Document Recognition V: 22-32*.
- Krüpl, B.; Herzog, M.; and Gatterbauer, W. 2005. Using visual cues for extraction of tabular data from arbitrary HTML documents. In *Poster Proc. WWW'05*, 1000-1001. ACM.
- Krüpl, B., and Herzog, M. 2006. Visually guided bottom-up table detection and segmentation in web documents. In *Poster Proc. WWW'06*. 933-934. ACM.
- Liu, B; Grossman, R. L.; Zhai, Y. 2003. Mining data records in Web pages. In *Proc. SIGKDD'03*, 601-606, ACM.
- Nagy, G., and Seth, S.C. 1984. Hierarchical representation of optically scanned documents. In *Proc. ICPR'84*, 347-349. IEEE.
- Simon, K., and Lausen, G. 2005. ViPER: augmenting automatic information extraction with visual perceptions. In *Proc. CIKM'05*, 381-388. ACM
- Wang, Y., and Hu, J. 2002. A machine learning based approach for table detection on the Web In *Proc. WWW'02*, 242-250. ACM.
- Yang, Y. and Zang, H. 2001. HTML Page Analysis Based on Visual Cues. In *Proc. ICDAR'99*, 859-864, IEEE.
- Zhao, H.; Meng, W.; Wu, Z.; Raghavan, V.; and Yu, C. 2005. Fully automatic wrapper generation for search engines. In *Proc. WWW'05*, 66-75. ACM

This work:

- Gatterbauer, W., and Bohunsky, P. 2006. Table extraction using spatial reasoning on the CSS2 visual box model. In *Proc. AAAI'06*, 1313-1318. AAAI, MIT Press.

# HTML RENDERING AS NON-INJECTIVE MAPPING

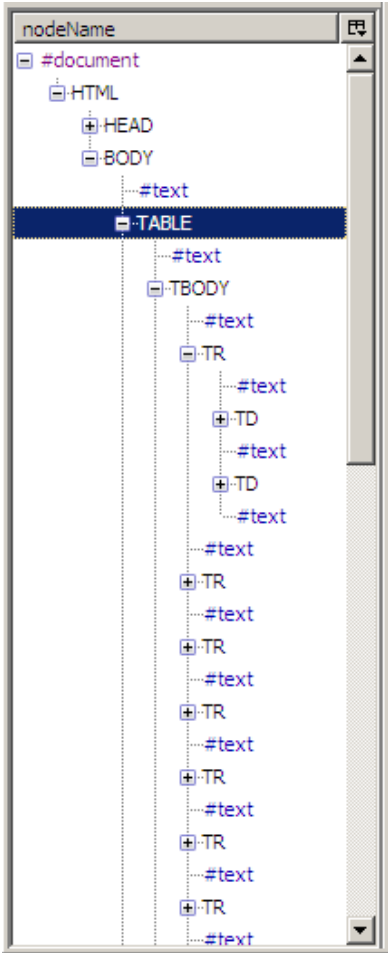
<DIV> table



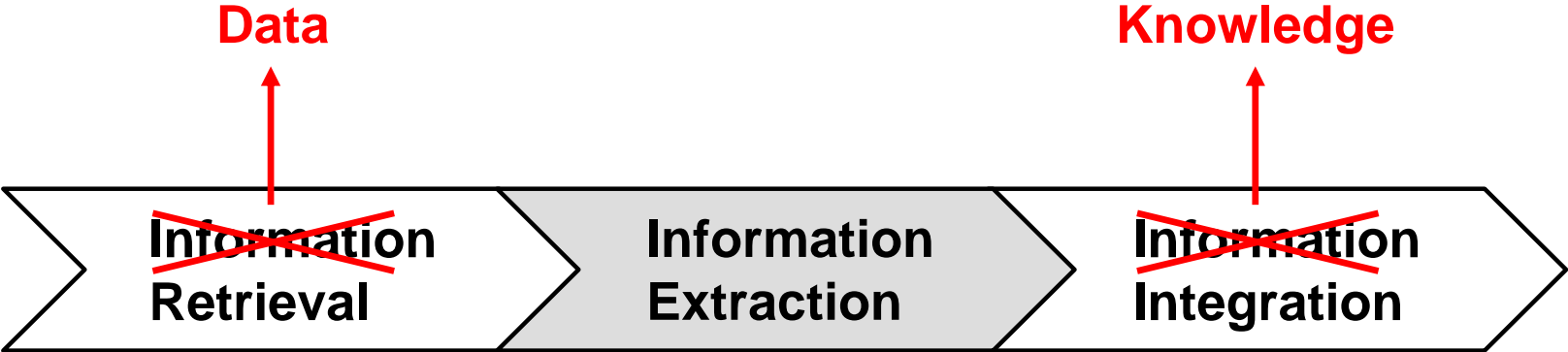
Visual rendering

Athlete	Country
AAMODT Kjetil Andre	NOR
ABRAMASHVILI Iason	GEO
ACTON Brigitte	CAN
AGUIRRE Facundo	ARG
AHUJA Neha	IND
ALBRECHT Daniel	SUI
ALCOTT Chemmy	GBR
ALIEVA Olesja	RUS
ANGUITA Daniela	CHI
ANTOR Alex	AND
ARNHOLD Mirella	BRA
AUFDENBLATTEN Fraenzi	SUI
BABUSIAK Jaroslav	SVK
BANK Ondrej	CZE
BARAHONA Noelle	CHI

<TABLE> table



# KNOWLEDGE ACQUISITION PROCESS



?