A Journey of Tabular Information from Unstructured to Structured Data World Using a Rule Engine

Presentation · October 2014

CITATIONS
0

READS
782

1 author:

Alexey Shigarov
Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences (ISDCT SB RAS)
60 PUBLICATIONS **215** CITATIONS

SEE PROFILE

A Journey of Tabular Information from Unstructured to Structured Data World Using a Rule Engine*

Alexey Shigarov[1]
shigarov@icc.ru

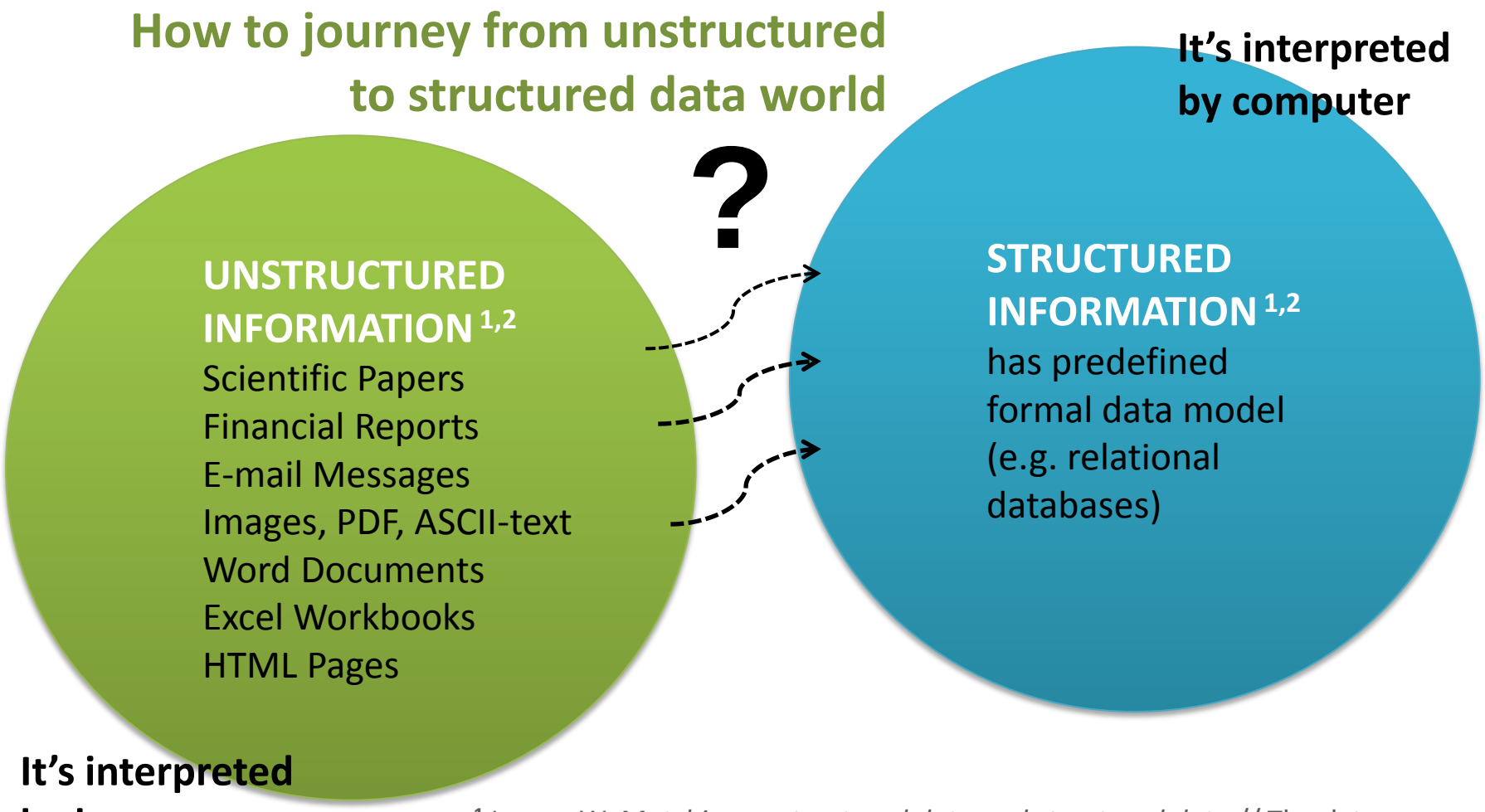
[1] *Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences*

HARBIN

October 18, 2014

* The work was partially supported by the Russian Foundation for Basic Research (Grant No 14-07-00166) and the Council for grants of the President of the Russian Federation (Scholarship No SP-3387.2013.5)

What's the problem with using data?

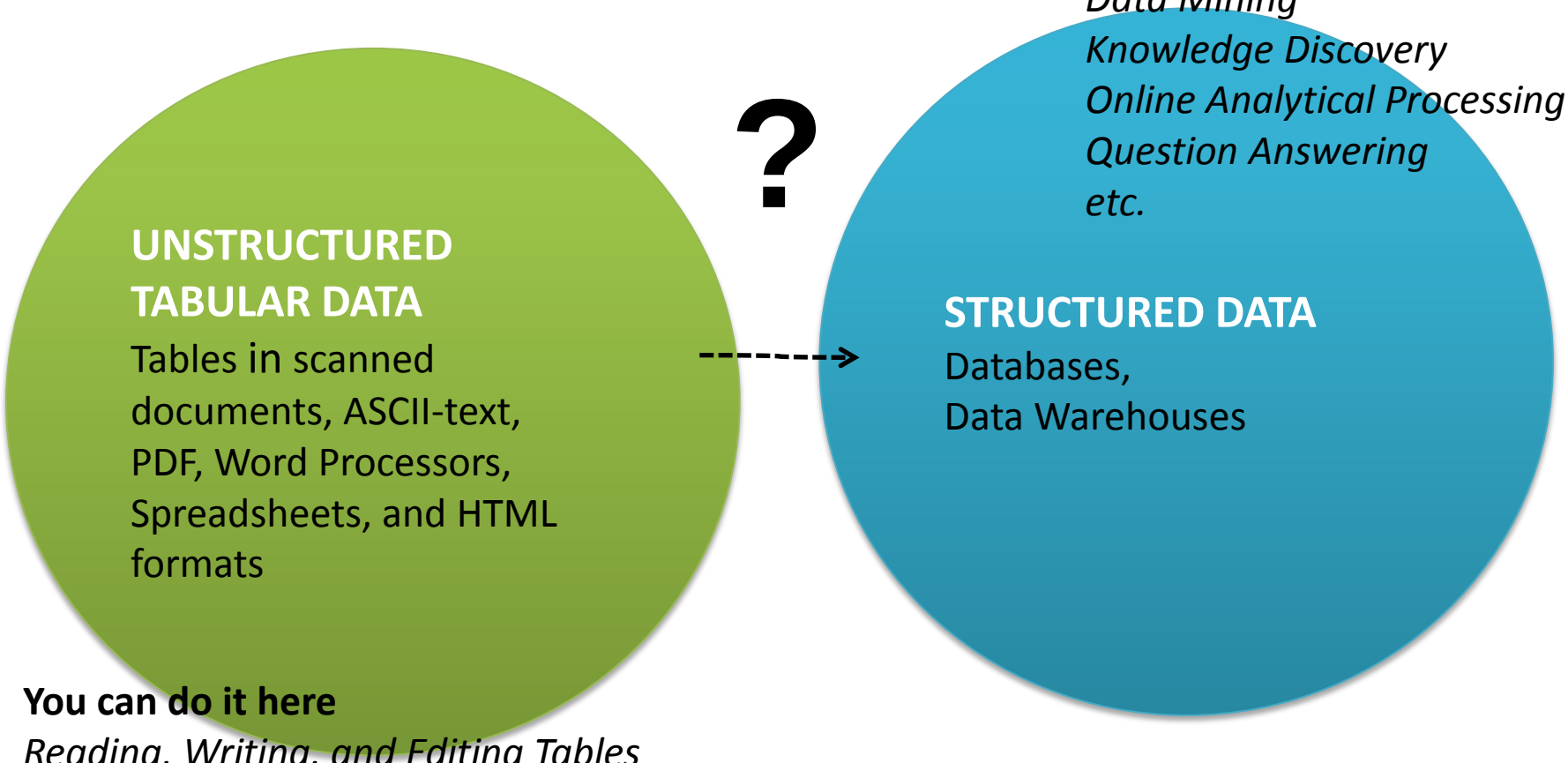


[1] Inmon W. *Matching unstructured data and structured data* // The data administration newsletter. 2006. http://www.tdan.com/view-articles/5009

[2] Blumberg R., Atre S. *The problem with unstructured data* // DM Review, 2003. http://soquelgroup.com/Articles/dmreview_0203_problem.pdf

One of the most important issues is

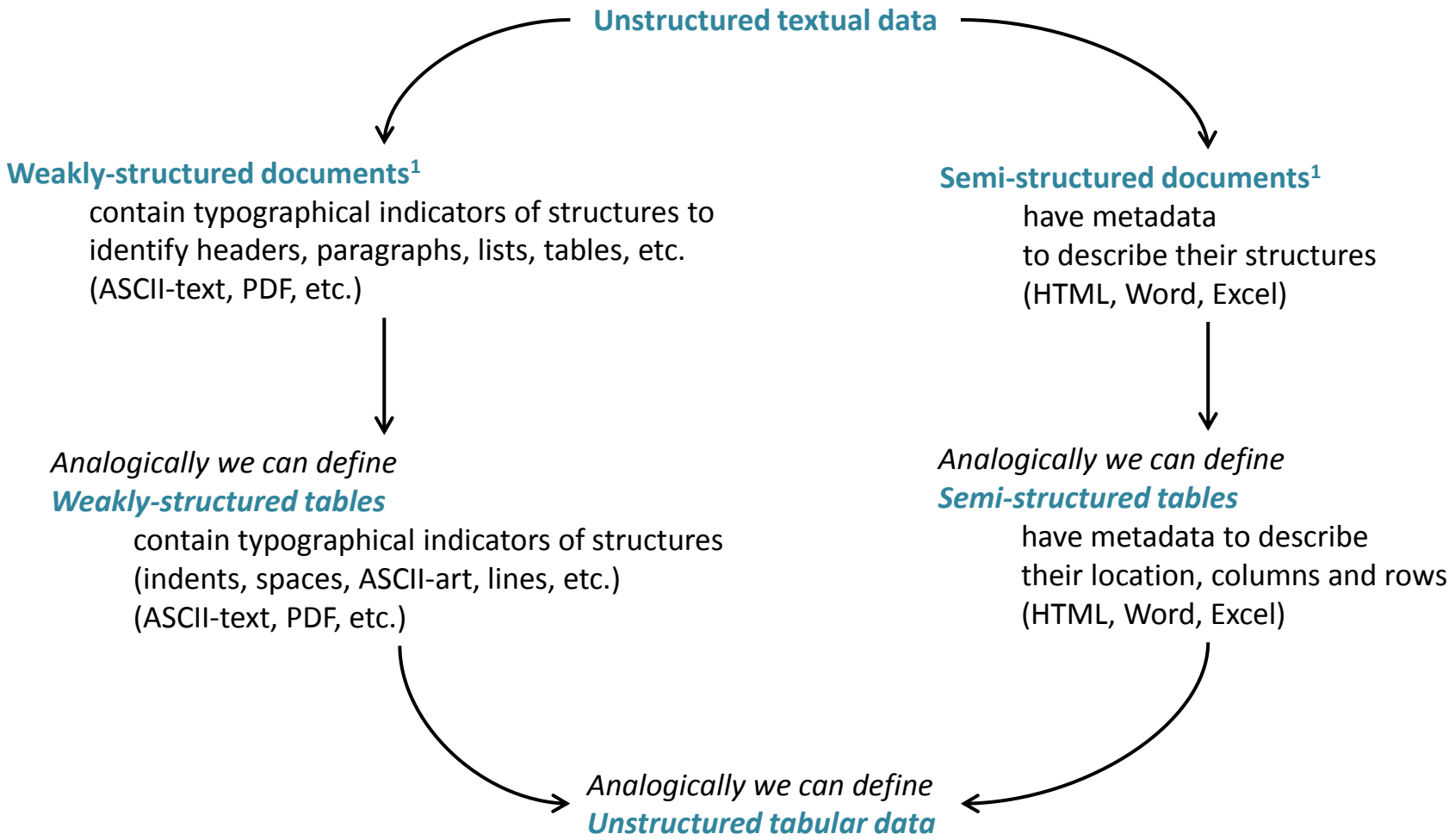
How to convert tabular data from unstructured to structured form?



You can do it here
Structured Queries
Data Mining
Knowledge Discovery
Online Analytical Processing
Question Answering
etc.

You can do it here
Reading, Writing, and Editing Tables

A bit about terminology



[1] The terms from the paper: Feldman R., Sanger J. ***The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*** // Cambridge University Press. 2006. 422 p.

Names for the conversion form unstructured to structured tabular data

- **Table Understanding**[1]

  includes the following subtasks: (1) table location, (2) table recognition, (3) functional and (4) structural analysis, and (5) table interpretation

  [1] Hurst M. ***Layout and Language: Challenges for Table Understanding on the Web*** // In Proc. of the 1st Int. Workshop on Web Document Analysis. 2001. pp. 27-30.

- **Table Understanding**[2]

  consists in the recovering of relationships between labels (headers) and data values as well as between labels and dimensions (domains)

- **Information Extraction from Tables**[2]

  "Information extraction from tables is perhaps analogous to the task of the same name applied to sentential text. The narrow definition requires a target schema and requires that arbitrary input (generally of some standard encoding) be transformed into instances of the schema" [2]

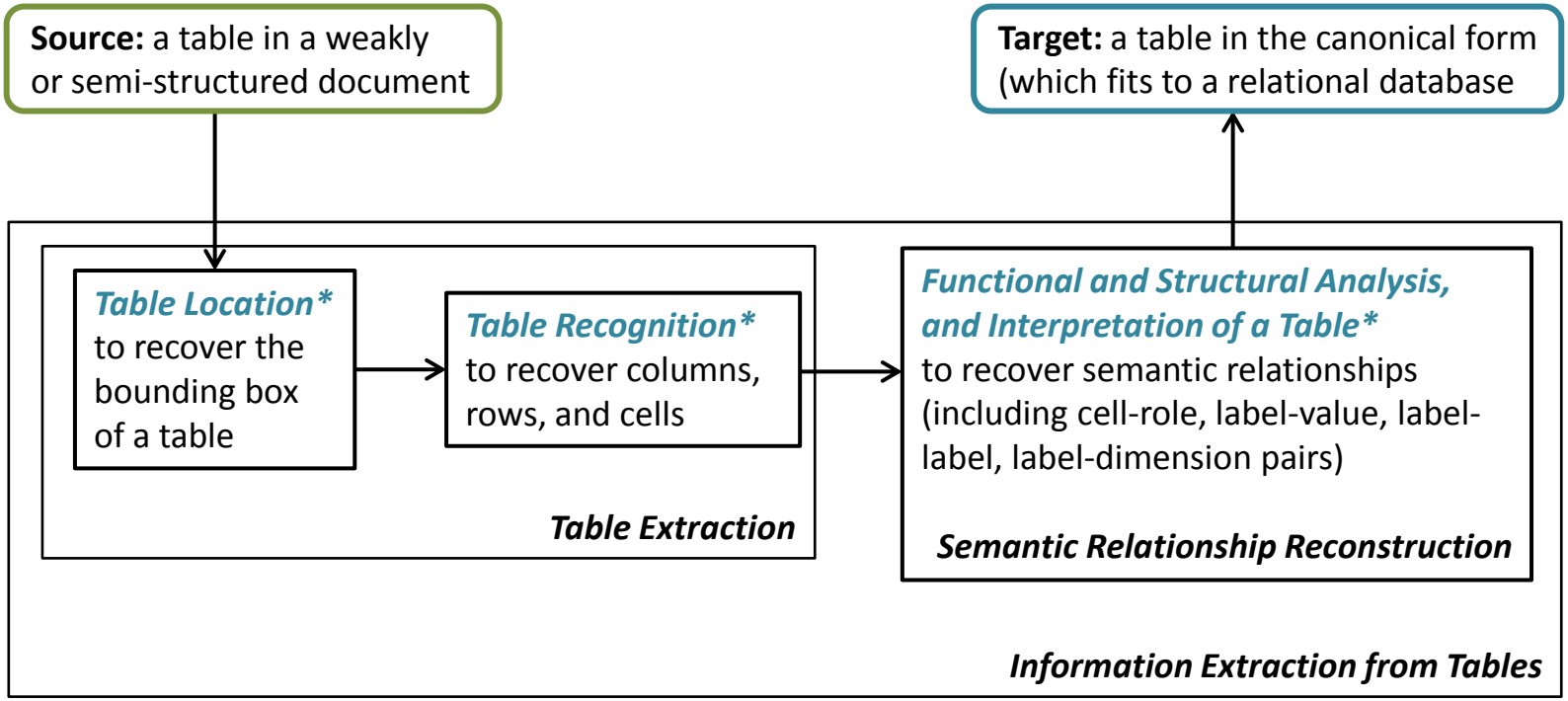
  [2] Embley D.W., Hurst M., Lopresti D., Nagy G. ***Table-Processing Paradigms: a Research Survey*** // Int. J. on Document Analysis and Recognition. 2006. Vol. 8, No. 2. pp. 66-86.

- **Table Canonicalization**[3]

  is transformation of a table to the canonical form that fits into the table of relational database

  [3] Tijerino Y., Embley D., Lonsdale D., Nagy G. ***Towards Ontology Generation from Tables*** // World Wide Web: Internet and Web Information Systems. 2005. Vol. 8, No 3. pp. 261-285.

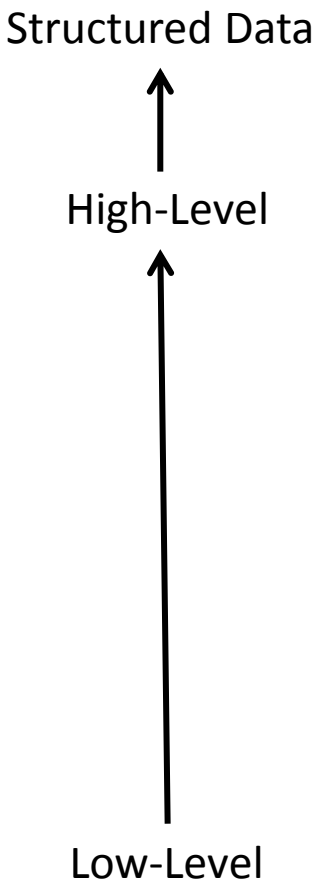
The way from unstructured to structured tabular data as the recovering of unknown information



* Термины из работы Hurst M. *Layout and Language: Challenges for Table Understanding on the Web* // In Proc. Of the Int. Workshop on Web Document Analysis. 2001. pp. 27-30.

Answer to question how to extract information from tables depends on how they are presented

Format	What we have initially	What we need to extract data
Excel Workbooks Word Documents HTML Pages	Cells with their positions, styles, and content (text, images)	**Semantic Relationship Reconstruction** (Functional and Structural Analysis, and Interpretation) Table Extraction (Location and Recognition)
PDF Documents	Characters with their positions, font metrics, as well as graphics	
Plain-Text	Characters with their positions	
Scanned Documents	Bitmaps	Character Recognition And so on



State-of-the-art in Information Extraction from Tables

Tasks & Formats	Research & Development, Software	Area
Semantic Relationship Reconstruction	Douglas S. (1995), Hurst M. (2000), e Silva A.C. (2004), Embley D. (2005), Tijerino Y. (2005), Pivk A. (2006), Gatterbauer W. (2007), Tao C. (2009) et al.	**Information Extraction**
Table Location and Recognition in HTML	Chen H.-H. (2000), Hurst M. (2001) , Yoshida M. (2001), Cohen W.W. (2002), Wang Y. (2002), Lerman K. (2004), Tengli A. (2004), Embley D.W. (2005), Tijerino Y. (2005), Krüpl B. (2006), Gatterbauer W. (2006, 2007), Weizsäcker L. (2008), et al.	
Table Location and Recognition in PDF, PS, or EMF	Ramel J.-Y. (2003), Hassan T. (2007), Hirano T. (2007) , Liu Y. (2007), Shigarov A. (2009, 2010), et al., **PDF to Word/Excel Converters**	**Document Analysis and Recognition**
Table Location and Recognition in Plain-text	Rus D. (1994), Douglas S. (1995), Tupaj S. (1996), Pyreddy P. (1997), Hurst M. (1997, 2001, 2003), Kieninger T. (1998, 1999, 2001), Ng H.T. (1999) , Hu J. (2000, 2001), Klein B. (2001), Pinto D. (2003) , e Silva A.C. (2004), Li J. (2006), et al.	
Table Location and Recognition in Scanned Documents	Kojima H. (1990), Chandran S. (1993), Itonori K. (1993), Green E. (1995), Hirayama Y. (1995), Watanabe T. (1995), Zuev K. (1997), Shamillian J.H. (1997), Tersteegen W.T. (1998), Handley J.C. (1999), Cesarini F. (2002), Wang Y. (2002), Wasserman H. (2002) , Mandal S. (2004, 2006), et al., **OCR Systems**	
Character Recognition	**OCR Systems**	

Challenges in table understanding

- A huge amount of ways to portray a table
  - Features originate from typographical standards, corporative practice, ad hoc software, data formats, and human inventiveness
- Assumptions about tables serve to reduce the complexity of table understanding
  - Usually those assumptions are entirely embedded in algorithms of existed systems
  - It constrains a range of tables which are successfully understood by these systems
- Today, no recognized corpus of test tables to evaluate a table understanding system

Our purpose

is to develop software for the conversion of tabular data from unstructured sources (like Excel workbooks, Word documents, HTML pages) to databases

Input unstructured tabular data

	Sent			Received	
	FY2010	FY2011	2011/2010 (%)	FY2010	FY2011
	Letters				
EU					
Spain	462.9	469.4	101.4	556.3	576.4
Cyprus	82.9	89.7	108.2	97.1	101.7
Belgium	352.3	341.1	96.82	387.2	366.1
Middle East					
Lebanon	21.1	21.5	101.9	19.8	19.5
Israel	353.8	483.0	136.5	365.8	376.0
	Parcels				
EU					
Spain	102.2	109.3	106.9	134.2	145.4
Middle East					
Lebanon	12.3	13.1	106.5	11.7	11.3

Output structured tabular data

DATA	OPERA-TION	YEAR	MAIL TYPE	REGION	COUNTRY
462.9	Sent	2010	Letters	EU	Spain
82.9	Sent	2010	Letters	EU	Cyprus
...	...	...	...	...	...
12.3	Sent	2010	Parcels	Middle East	Lebanon
469.4	Sent	2011	Letters	EU	Spain
89.7	Sent	2011	Letters	EU	Cyprus
341.1	Sent	2011	Letters	EU	Belgium
21.5	Sent	2011	Letters	Middle East	Lebanon
483.0	Sent	2011	Letters	Middle East	Israel
109.3	Sent	2011	Parcels	EU	Spain
13.1	Sent	2011	Parcels	Middle East	Lebanon
556.3	Received	2010	Letters	EU	Spain
...	...	...	...	...	...
11.3	Received	2011	Parcels	Middle East	Lebanon

Our approach to the semantic relationship reconstruction

- **We assume that**
  - Tables produced by the same vendor often have similar layout, formatting, and content
  - It allows to define a template describing a class of these tables or how they can be interpreted
- **We propose**
  - To use a set of formalized rules (a knowledge base) for recovering semantic relationships (i.e. cell-role, label-value, label-label, label-dimension pairs) in a table from the class
  - The rules define how we can interpret what we know (i.e. positions, style settings, and content of cells) to recover what we don't know (i.e. semantic relationships)
- **It is expected that**
  - Implementation of rule sets for different table classes provides processing of a wide range of tables having various structures and features

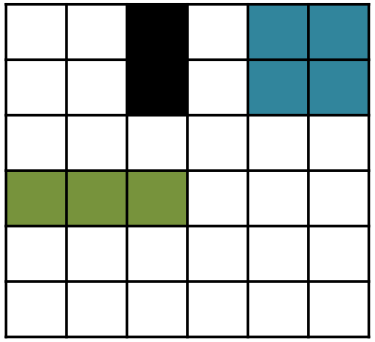
What we mean when we say “Table”

Table structure on the bottom level

- **Cell positions** (row and column coordinates)
- **Merged cells** (as shown in Figure *A*, but not in Figure *B* or *C*)
- **Cell style** (border style, content placement, text metrics, etc.)
- **Cell content** (a text and/or images, but not other tables or cells)

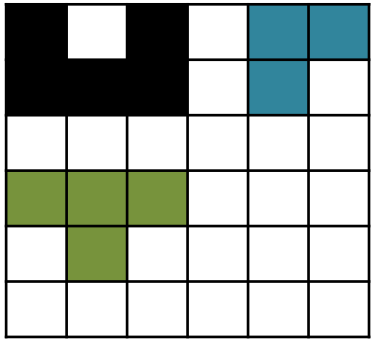
Perhaps, it’s a table generated in Excel, Word, HTML, etc.

A



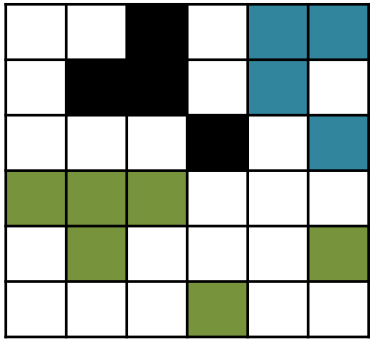
A cell can include multiple tiles in Excel, Word, HTML, or LaTeX

B



A cell can visually include a few tiles for human reading using graphical lines

C



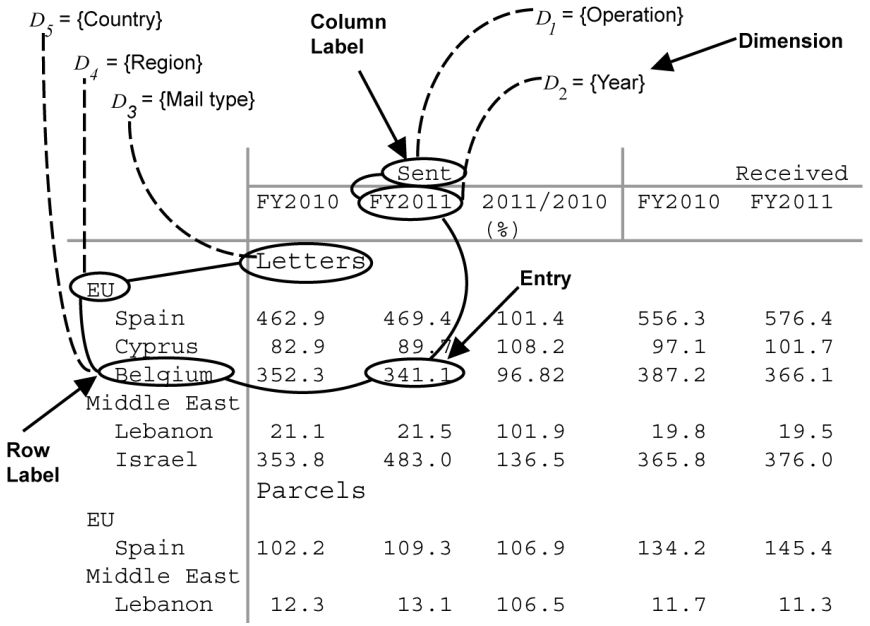
Perhaps nobody presents a cell like this

What we mean when we say “Table”

Table structure on the top level

- A cell serves as either entry* or label*
- An entry represents data value
- A label describes (addresses) entries
- A label can address entries and other labels either in rows or in columns where it spans them
- A label can be a value of a dimension

Perhaps, it’s a pivot table generated by an OLAP system



* The terms “entry” and “label” correspond to the meaning that was suggested in the paper Wang X. ***Tabular Abstraction, Editing, and Formatting***. PhD Thesis. Waterloo, Ontario, Canada. 1996.

CELLS table model, Bottom level

Known facts about a table

$Tb = (Sr, Sc, C)$, where

Sr — a set of rows

Sc — a set of columns

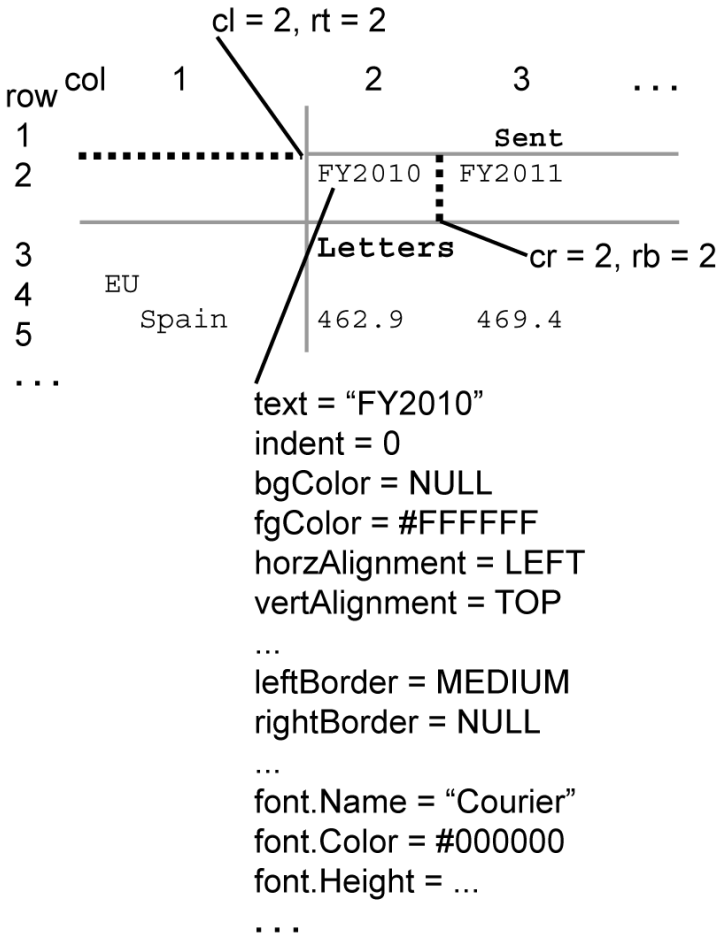
C — a set of cells

A cell — $c = (p, S, c')$, where

$p = (cl, rt, cr, rb)$ — positions in the row-column coordinate system (Sr and Sc sets)

S — style settings (including colors, font metrics, adjustment, styles of borders etc.)

c' — a content (text, images)



CELLS table model, Top level

Unknown facts about a table

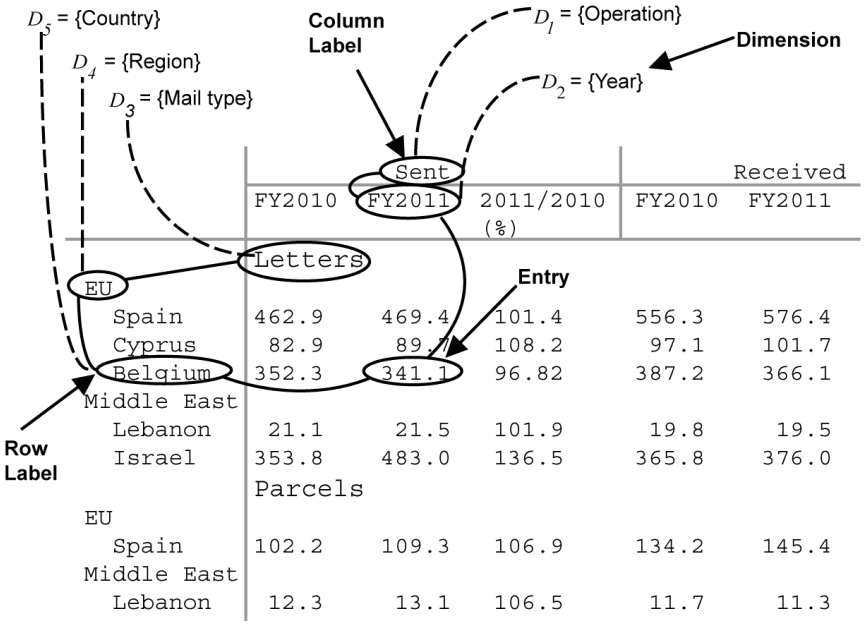
$Tt = (\ D, Lr, Lc, E\ )$, where

D — a set of domains

Lr — a tree of row labels

Lc — a tree of column labels

E — a set of entries



A label — $l = (\ l'\ )$, where

l' — a content, which is not a value of any domain from D

An entry — $e = (\ D', L', e'\ )$, where

D' — a subset of values of domains from D related with this entry

L' — a set of labels from trees Lr and Lc related with this entry

e' — a content

Proposed schema for the rule-based extraction information from tables

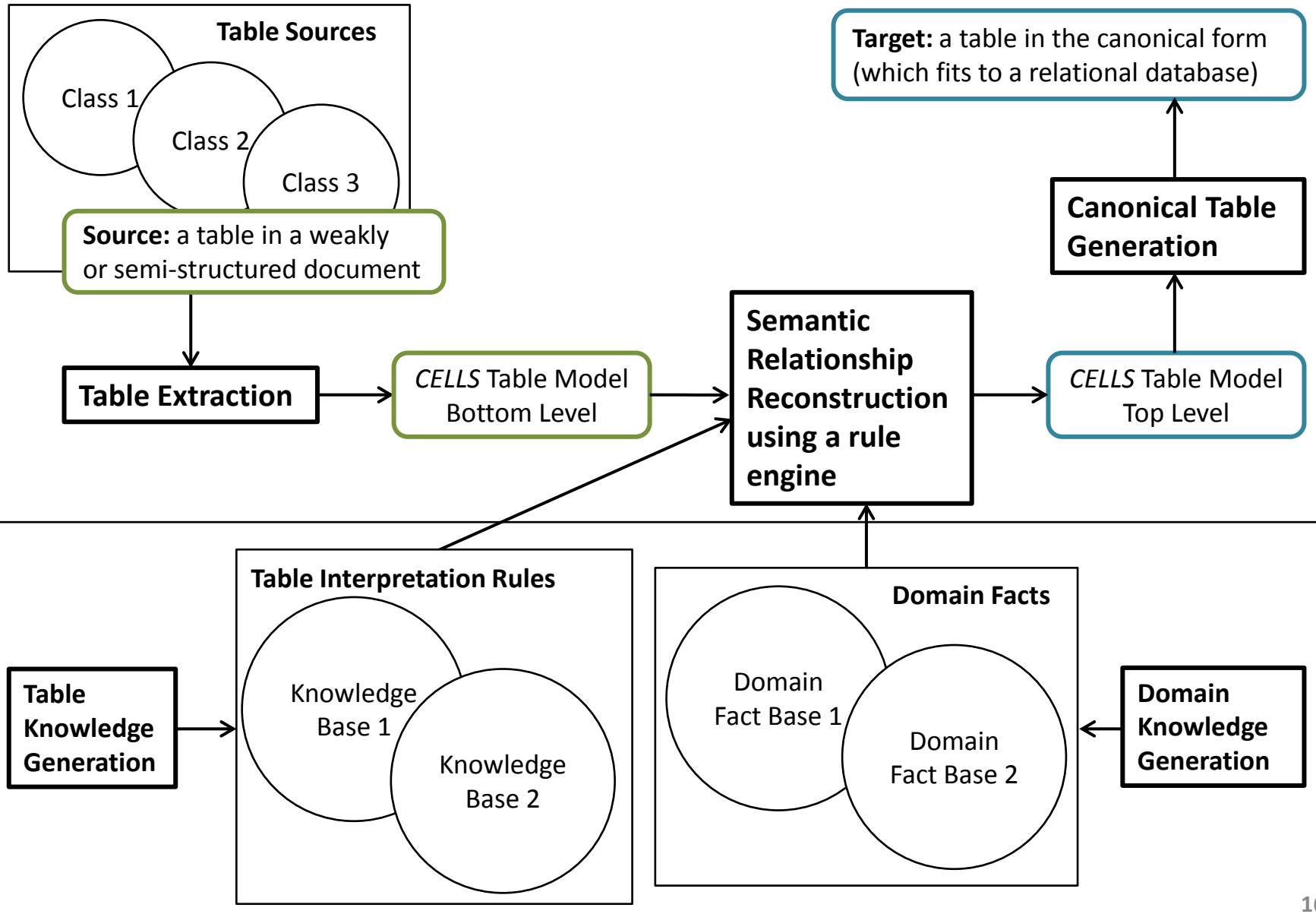
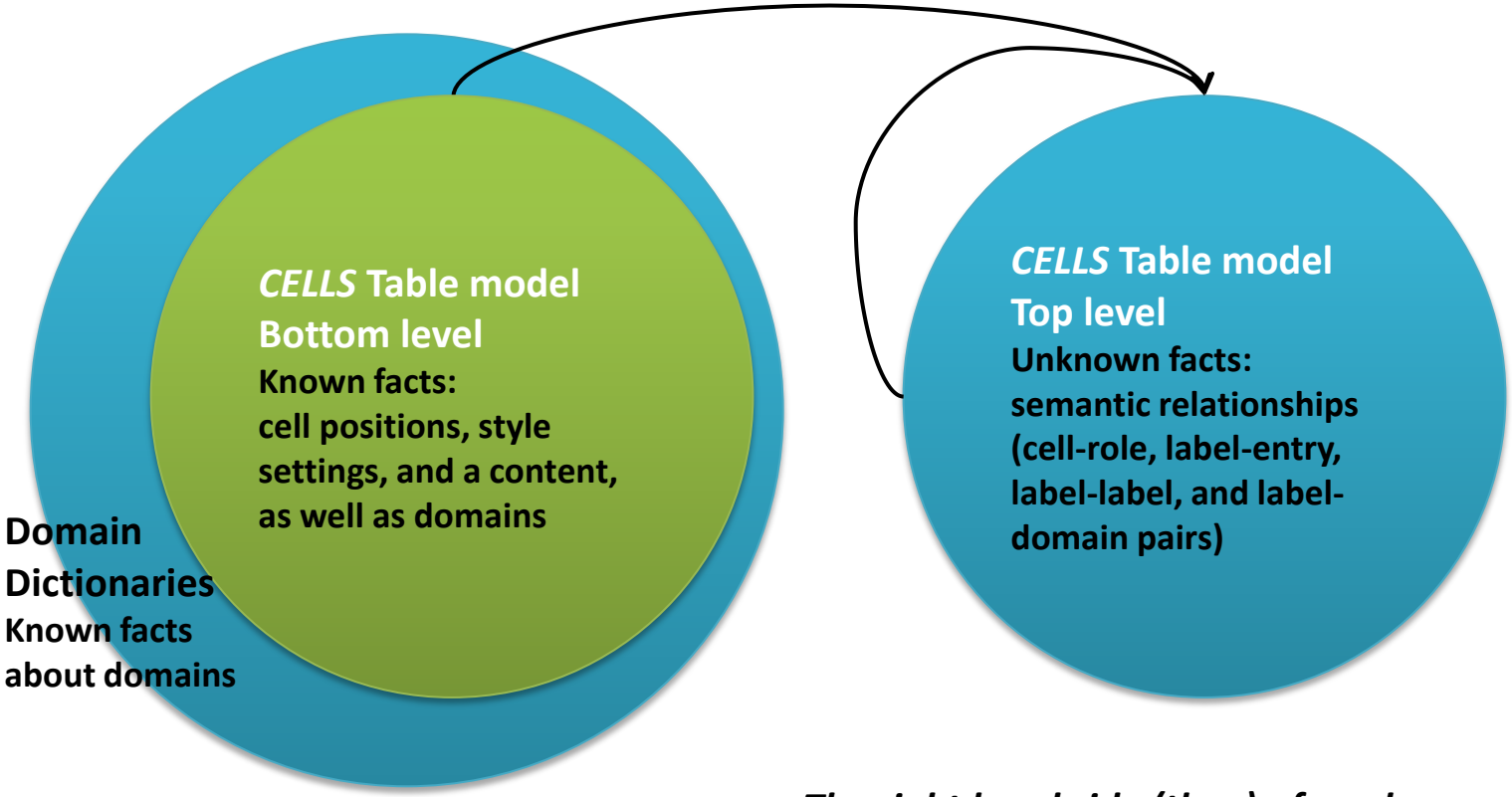


Table interpretation rules

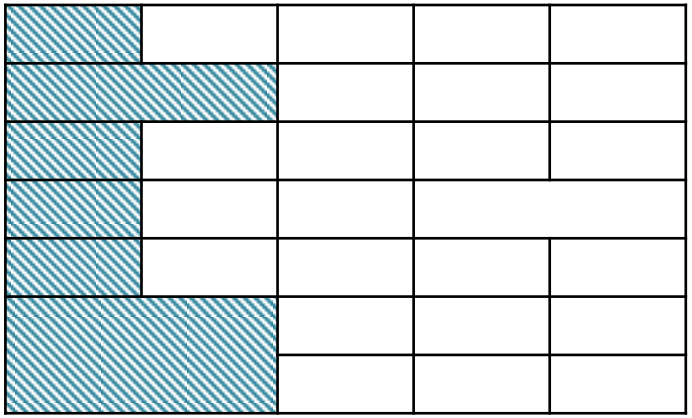


The left hand side (when) of a rule
defines conditions using known facts about cells and domains

The right hand side (then) of a rule
recovers unknown facts about a table, including assignment cell roles (label or entry), binding cells (i.e. creating label-entry, label-label, and label-domain pairs),
etc.

Rules for table structure analysis*
Sample 1

```
...
when
    $c : CCell( cl == 1 )
then
    modify ( $c ) { setRole( Role.ROWLABEL ) }
...
```

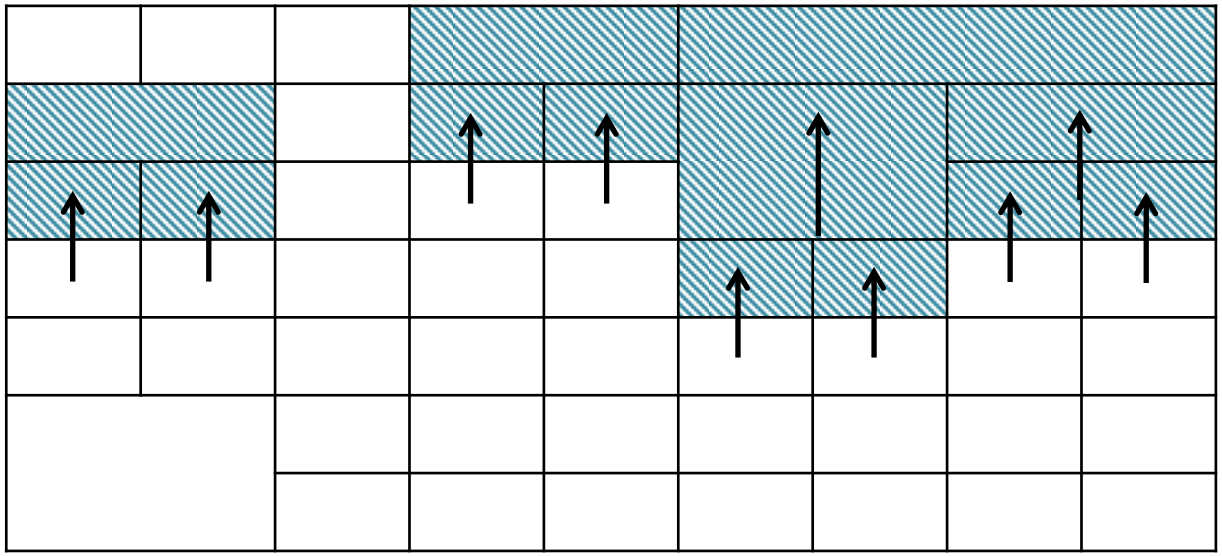


*The rules are written by the expression language *MVEL* http://mvel.codehaus.org
for the rule engine *Drools Expert* http://www.jboss.org/drools

Rules for table structure analysis

Sample 2

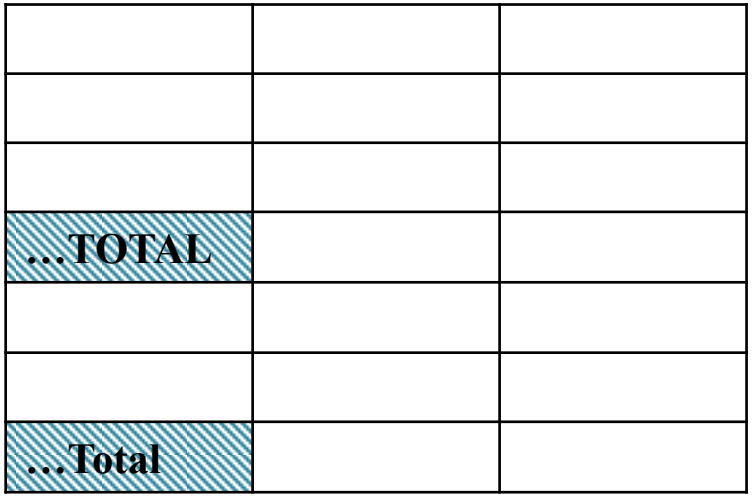
```
...
when
    $c1 : CCell()
    $c2 : CCell( rt == $c1.rb + 1,
                 ( $c1.cl <= cl && cr < $c1.cr ) ||
                 ( $c1.cl < cl && cr <= $c1.cr ) )
then
    $c1.addConnectedCell( $c2 )
...
```



Rules for table structure analysis

Sample 3

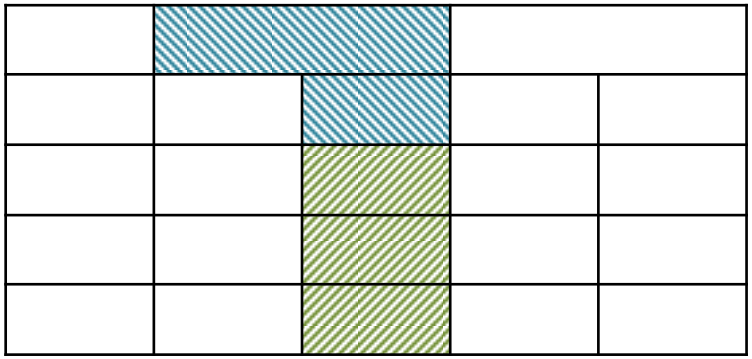
```
...
when
    $c : CCell( cl == 1, cl == cr, text matches "(?i).*(total)" )
then
    modify ( $c ) { setIgnored( true ) }
...
```



Rules for table structure analysis

Sample 4

```
...
when
    $l : CCell( role == Role.COLLABEL )
    $e : CCell( role == Role.ENTRY, cl == $l.cl, cr == $l.cr )
then
    $e.addConnectedCell( $l )
...
```



Rules for table structure analysis

Sample 5

```
...
when
    $d : CDimension( name == "Religion" )
    $c : CCell ( cl == 1, rt > 1,
                 text != null, role == null,
                 style.getFont().getColor() == "#ff0000" )
    not ( exists CCell ( cl > $c.cr, rt == $c.rt, text != null ) )
then
    $c.setDimension( $d )
...
```

More samples of rules are available at address http://cells.icc.ru/test

	Text	Text	Text
Text			
Text	Text	Text	Text
Text	Text	Text	Text
Text	Text	Text	Text
Text			
Text	Text	Text	Text
Text	Text	Text	Text

Optional preprocessing for cells

- **Elimination of empty rows and columns**
- **Cell border enhancement**

  Visual borders of a cell are not always its physical borders

  They can be visually formed by borders of its neighbor cells (***a***)

rightBorder = *MEDIUM*	**leftBorder** = *NONE*

a

  Unknown border style settings of a cell are recovered by border style settings of its neighbor cells (***b***)

rightBorder = *MEDIUM*	**leftBorder** = *MEDIUM*

b

  It allows simplifying table interpretation rules

Optional pre- and post-processing for text

- **Removal of whitespaces and special symbols**

  For example, the expression ␣ ␣ ␣ ␣ **`Total.....`** is converted to **`Total`**

- **Conversion from synonymous to reference expressions using reference dictionaries**

  For example, the following synonyms: **`2014`**, **`FY2014`**, **`2014 год`**, **`Current year`**
  can have the same meaning **`Year 2014`**

  A reference dictionary is a set of pairs (R_s, R_t), where

  R_s — a source natural language or regular expression
  R_t — a target natural language or regular expression

  For example, the pair (**`FY[2][0][0-1][0-4]`**, **`[2][0][0-1][0-4]`**) allows converting all synonyms sort of **`FY2000`**, … , **`FY2014`** to the following reference expressions **`2000`**, … , **`2014`** correspondingly

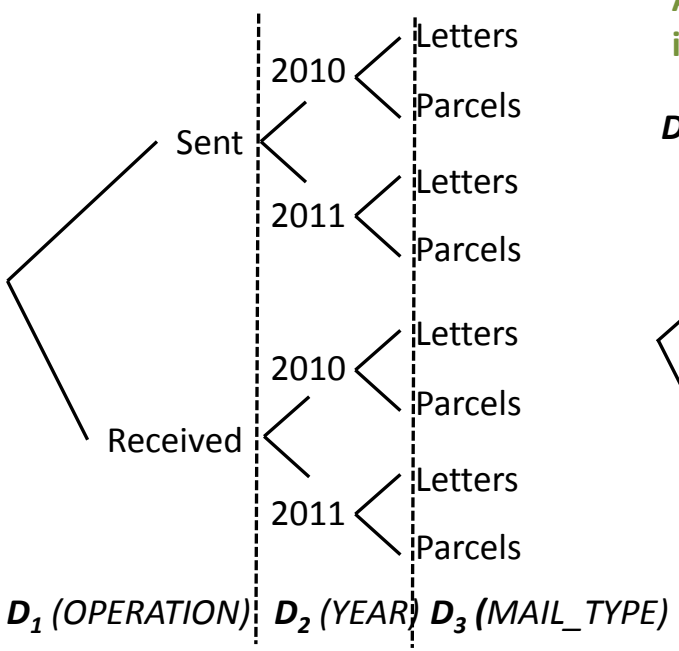
Optional post-processing for label trees

- Optionally, labels in trees can be assigned to domains using domain dictionaries
- A Domain Dictionary is a set of pairs (R, D_i), where

  R — a natural language or regular expression
  D_i — a domain

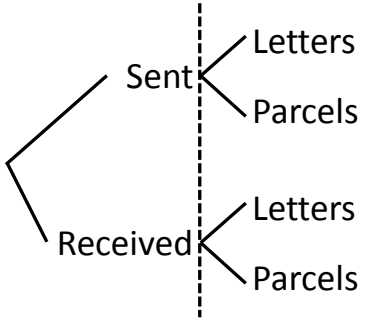
- In the result label trees can be reduced or completely degenerate

Before post-processing



After post-processing in the common case

D₂ (YEAR) = {2010, 2011}



After post-processing in the perfect case

D₁ (OPERATION) = {Sent, Received}
D₂ (YEAR) = {2010, 2011}
D₃ (MAIL_TYPE) = {Letters, Parcels}

Generation of a table in the canonical form

- A generated table in the canonical form consists of the following fields

  DATA contains entries

  ROW_LABEL contains paths from leaves to roots in the non-degenerate tree of row labels Lr

  COL_LABEL contains paths from leaves to roots in the non-degenerate tree of column labels Lc

  D_1,…,D_N present values of the corresponding domains Di from the set D

- Generated tables in the canonical form can be exported into a relational database using standard tools of database management systems

The instance of an output table in the canonical form

DATA	OPERA-TION	YEAR	MAIL TYPE	REGION	COUNTRY
462.9	Sent	2010	Letters	EU	Spain
82.9	Sent	2010	Letters	EU	Cyprus
...	...	...	...	...	...
12.3	Sent	2010	Parcels	Middle East	Lebanon
469.4	Sent	2011	Letters	EU	Spain
89.7	Sent	2011	Letters	EU	Cyprus
341.1	Sent	2011	Letters	EU	Belgium
21.5	Sent	2011	Letters	Middle East	Lebanon
483.0	Sent	2011	Letters	Middle East	Israel
109.3	Sent	2011	Parcels	EU	Spain
13.1	Sent	2011	Parcels	Middle East	Lebanon
556.3	Received	2010	Letters	EU	Spain
...	...	...	...	...	...
11.3	Received	2011	Parcels	Middle East	Lebanon

Test data

We collected a corpus (~**100** tables in *Excel* spreadsheet format)

They are available at address http://cells.icc.ru/test

- Any test table has precise layout of their cells
- Additionally we use special tags to locate a test table in the corresponding *Excel* sheet
- It allows to avoid steps of table detection and table structure recognition (i.e. identifying columns and rows)

Tag for identifying the start point of a table

$START							
	Company name	**Place of incorporation and operation**	**Activity**	**Percentage held as of December 31, 2006**	**Percentage held as of December 31, 2005**	
	LLC “Airport Moscow”	Moscow region	Cargo handling	50,00%	50,00%	
	CJSC “Aerofirst”	Moscow region	Trading	33,30%	33,30%	
	CJSC “TZK Sheremetyevo”	Moscow region	Fuel trading company	31,00%	31,00%	
	CJSC “AeroMASH – AB”	Moscow region	Aviation security	45,00%	45,00%	
						$END

Tag for identifying the end point of a table

Sample of test tables

From Statistical Handbook of Japan 2007

Item	Total	National forest	Non-national forest		
			Municipal	Private	Others
Forest land area (1,000 ha)	25 121	7 838	2 796	14 440	46
Forest growing stock (1 mil. m3)	4 040	1 011	433	2 590	5
Planted forests					
Land area (1,000 ha)	10 361	2 411	1 232	6 705	12
Growing stock (1 mil. m3)	2 338	368	255	1 712	3
Natural forests					
Land area (1,000 ha)	13 349	4 770	1 426	7 126	27
Growing stock (1 mil. m3)	1 701	642	178	878	3

Fiscal year	Technology Trade				Exports value / Imports Value
	Exports		Imports		
	Value (billion yen)	Annual increase rate (%)	Value (billion yen)	Annual increase rate (%)	
1990	339,4	3	371,9	12,7	0,91
1995	562,1	21,6	391,7	5,7	1,43
2000	1057,9	10,1	443,3	8	2,39
2002	1386,8	11,2	541,7	-1,2	2,56
2003	1512,2	9	563,8	4,1	2,68
2004	1769,4	17	567,6	0,7	3,12
2005	2028,3	14,6	703,7	24	2,88

Sample of a test table

From USDA NASS Statistical Report 2003

Kind of seed	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
	Price per 100 pounds									
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
Alfalfa, uncertified varieties	152.00	161.00	168.00	185.00	185.00	205.00	184.00	165.00	158.00	280.00
Alfalfa, certified varieties	269	266	274	277	282	288	287	277	278	157
Clover, ladino	324	321	320	318	307	308	298	285	285	280
Clover, red	148	148	134	172	184	194	178	143	132	130
Lespedeza, Korean	132	84,5	66	99	90	89	76,15	77,5	160	98
Sunflower	300	297	297	313	355	380	400	395	407	407
Cottonseed, all	62,7	63,5	68,2	73	74,9	79,3	82,4	128	154	213
Biotech [1]	.....	.....	.....	.....	.....	.....	.....	.....	217	271
Non-biotech	.....	.....	.....	.....	.....	.....	.....	.....	87	94
Grain sorghum, hybrid	74,5	82,1	78,7	84	92	96	97,6	93	93	96
	Price per bushel									
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
Corn, hybrid, all [2]	72,7	73,4	77,1	77,7	83,5	86,9	88,1	87,5	92,2	92
Biotech [1]	.....	.....	.....	.....	.....	.....	.....	.....	110	113
Non-biotec	.....	.....	.....	.....	.....	.....	.....	.....	85,3	85,8
Wheat (spring)	5,98	7,37	7,12	8,1	7,3	6,85	6,1	6,1	6,2	6,5
Wheat (winter)	7,73	7,9	7,8	8,5	10	8,25	7,35	7,05	7,2	7,7
Rice	15,4	22	15,1	17,5	19	19,5	19,1	17,25	15,7	14,9
Barley (spring)	5	5,18	5,37	6,49	6,13	6,04	5,8	5,8	5,8	5,8
Soybeans for seed, all	12,4	13,6	13,4	14,8	16,1	17,15	17	17,1	20,7	22,5
Biotech [1]	.....	.....	.....	.....	.....	.....	.....	.....	23,9	27
Non-biotec	.....	.....	.....	.....	.....	.....	.....	.....	17,9	15
Flaxseed	7,37	7,74	8	8,14	9,31	10	8,5	7,9	7,6	7,6

Sample of a test table

From China Statistical Yearbook 2003

线路名称	Name	客运量（万人）Passenger Traffic (10 000 persons)	旅客周转量（百万人公里）Passenger-kilometers (million passenger-km)	线路名称	Name	货运量（万吨）Freight Traffic (10 000 tons)	货物周转量（百万吨公里）Freight Ton-kilometers (million ton-km)
京沪线	Beijing-Shanghai	5496	32975	京沈线	Beijing-Shenyang	3438	82790
新石线	Xinjiang-Rizhao			哈大线	Harbin-Dalian	3233	60717
沪杭	Shanghai-Hangzhou	654	6188	津沪线	Tianjin-Shanghai	5304	100909
浙赣线	Hangzhou-Ganzhou	3785	33028	沪杭线	Shanghai-Hangzhou	202	4939
鹰厦线	Yingtan-Xiamen	10869	88717	京广线	Beijing-Guangzhou	7187	131196
京九线	Beijing-Kowloon	814	1906	南北同蒲线	Datong-Taiyuan-Fenglingdu	11168	30412
京广线	Beijing-Guangzhou	491	1708	太焦柳线	Taiyuan-Jiaozuo-Liuzhou	8206	56729
石太线	Shijiazhuang-Taiyuan	1800	9364	京九线	Beijing-Kowloon	2644	61919
石德线	Shijiazhuang-Dezhou	1575	6452	兰新线	Lanzhou-Urumqi	3366	63348
焦柳线	Jiaozuo-Liuzhou	655	2512	滨洲线	Harbin-Manzhouli	3137	21181
京包线	Bingjing-Baotou	541	1288	滨绥线	Harbin-Suifenhe	1178	16384
包兰线	Baotou-Lanzhou	1245	3615	京包线	Bingjing-Baotou	5881	57077
北同蒲线	Taiyuan-Datong	4759	33838	石太线	Shijiazhuang-Taiyuan	3760	21301
南同蒲线	Fenglingdu-Taiyuan	74	1459	石德线	Shijiazhuang-Dezhou	379	11664
陇海线	Lianyungang-Lanzhou	1055	16149	浙赣线	Hangzhou-Ganzhou	2464	45035
宝中线	Baoji-Zhongwei	413	1865	陇海线	Lianyungang-Lanzhou	6357	100027

Sample of a test table

From China Statistical Yearbook 2003

地　区	Region	铁　路 营业里程 Length of Railways in Operation	内河航道里程 Length of Navigable Inland Waterways	公路里程 Total Length of Highways	等级路 Expressway and Class I to IV Highway	# 高　速 Express -way	# 一　级 First Class	# 二　级 Second Class	等外路 Highway Below Class IV
全　国	**National Total**	**71897,5**	**121557**	**1765222**	**1382926**	**25130**	**27468**	**197143**	**382296**
北　京	Beijing	1138,1		14359	13940	463	331	1822	419
天　津	Tianjin	681,6	443	9696	9126	331	404	1408	570
河　北	Hebei	4585,7	75	63079	53995	1591	2050	9835	9084
山　西	Shanxi	3050,5	305	59611	57250	1070	734	8851	2361
内蒙古	Inner Mongolia	6192,6	1188	72673	63000	252	330	6069	9673
辽　宁	Liaoning	3799,8	813	48051	47769	1637	987	10770	282
吉　林	Jilin	3561,8	1787	41095	38408	542	1120	4918	2687
黑龙江	Heilongjiang	5502,8	5057	63046	57882	413	707	5821	5164
上　海	Shanghai	256,5	2037	6286	6024	240	442	1203	262
江　苏	Jiangsu	1340,4	23899	60141	49959	1704	3085	10637	10182
浙　江	Zhejiang	1300,1	10408	45646	42759	1307	2070	5777	2887
安　徽	Anhui	2219,7	5611	67547	61406	866	300	7480	6141
福　建	Fujian	1453,9	3701	54155	41220	583	278	5573	12935
江　西	Jiangxi	2368,6	5537	60696	36070	666	314	6731	24626
山　东	Shandong	2855,4	2552	74029	73884	2411	3521	20251	145

Experimental evaluation

Source	Number of						Inference time (ms) ***
	Tables	Cells	Entries	Labels	Relationships (label-label)*	Rules **	
JAPAN_STAT[1]	15	1088	734	257	102	10	417
AEROFLOT[2]	13	2047	727	321	167	16	526
BOEING[3]	21	2156	964	470	196	14	663
CHINA_STAT[4]	18	7216	4180	862	551	12	964
CHEVRON[5]	7	812	268	141	89	12	283
USDA_NASS[6]	7	1553	1175	313	174	16	638
TOBACCO[7]	16	2844	2195	508	335	10	730

[1]Statistical Handbook of Japan 2007. Chapter 5, 8. Statistics Bureau of Japan.

[2]OJSC "Aeroflot – Russian Airlines" Consolidated Financial Statements For the Year Ended December 31, 2006. Aeroflot.

[3]Boeing Co, Annual Report 2010. Boeing Co.

[4]China statistical yearbook 2003

[5]Chevron Corp. News Release November 2, 2012

[6]Agricultural Statistics Annual. Chapter VI Statistics of hay, seeds, and minor field. USDA NASS. 2003

[7]Tobacco: World Markets and Trade 2005. USDA (U.S. Department of Agriculture). Foreign Agricultural Service

* Excluding relationships from roots of label trees
** Rules and results are available at http://cells.icc.ru/test
*** For the processor *Intel Core 2 Quad*, 2,66 ГГц and the rule engine Drools Expert (5.4.0.Final), http://www.jboss.org/drools

Related work

- Hurst M. ***The Interpretation of Tables in Texts****. PhD. Thesis*. School of Cognitive Science, Informatics, The University of Edinburgh. UK, 2000.

- e Silva A.C., Jorge A.M., Torgo L. ***Design of an End-to-End Method to Extract Information from Tables*** // Int. J. on Document Analysis and Recognition. Springer-Verlag. 2006. Vol. 8, No. 2. pp. 144–171.

- Embley D.W., Tao C., Liddle S.W. ***Automating the Extraction of Data from HTML Tables with Unknown Structure*** // Data & Knowledge Engineering. Elsevier. 2005. Vol. 54, No 1, pp. 3–28.

- Tijerino Y., Embley D., Lonsdale D., Nagy G. ***Towards Ontology Generation from Tables*** // World Wide Web: Internet and Web Information Systems. 2005. Vol. 8, No 3. pp. 261–285.

- Pivk A., Cimiano P., Sure Y., Gams M., Rajkovi˘c V., Studer R. ***Transforming Arbitrary Tables into Logical Form with TARTAR*** // Data & Knowledge Engineering. 2007. Vol. 60 , pp. 567–595.

- Gatterbauer W., Bohunsky P., Herzog M., Krüpl B., Pollak B. ***Towards Domain-Independent Information Extraction from Web Tables*** // In Proc. of the 16th Int. Conf. on World Wide Web. ACM New York, NY, US, 2007. pp. 71–80.

Contribution

- Implementation of the rule-based approach to table understanding
- Using both domain independent (spatial and style) information and domain-specific (natural-language) information for table analysis and interpretation
- Dealing with table features like cut-ins (headers in a table body), non-numerical data values, the duplication of multilingual labels, label columns which are alternated by data columns, and so on

Application

- Unstructured tabular data integration in business intelligence
- Populating databases with statistical information
- Information extraction from financial reports

Conclusion

- Our approach is oriented

  1. to use in table-processing all or nearly all of tabular data available in a source (spatial structure, styles, and natural language)

  2. to be applied to conversion of tabular data from unstructured to structured form as part of information integration

- Now, our system provides information extraction from a wide range of tables in *Excel* spreadsheet format

- Perhaps, further development of the proposed model, data structures, as well as post-processing and preprocessing algorithms allows to simplify the writing of rules

- Each original class of tables produced by the same vendor potentially requires developing a suitable knowledge base

- Perhaps, development of an unified knowledge base for heterogeneous sources from various vendors is too expensive or even impossible since they often are contradictory

Collaboration

- If you interested in using our technologies for your tasks of large-volume conversion of tabular data from unstructured sources to databases
- If you interested in a cooperative research project
- Please, e-mail us at shigarov@icc.ru
- We are interested in the development and use of our technologies both in research and practice

Thanks!

Alexey Shigarov
mailto:shigarov@icc.ru
http://cells.icc.ru