# My Adventures with Datalog

## Walking the Thin Line Between Theory and Practice

Georg Gottlob

University of Oxford & TU Wien

Talk at the University of Linz
03 November 2020

# Cultural Background

H. Hahn, M. Schlick, O. Neurath, R. Carnap, K. Gödel

Ludwig Wittgenstein

Vienna was a hotspot of logic in the early 20th century.

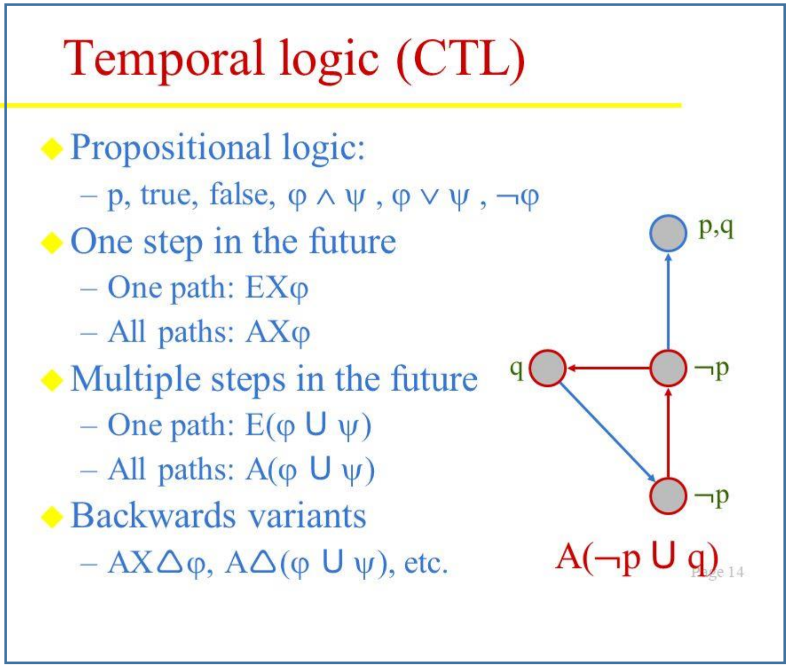
This was stopped by the rise of the Nazis in the late 1930ies.

**Goal**: without any pretention, help reestablishing a tradition of logic in Austria: **Logic in CS – our Mission**

**Computer Science is the continuation of logic by other means.**

# New logics for specific applications have been designed

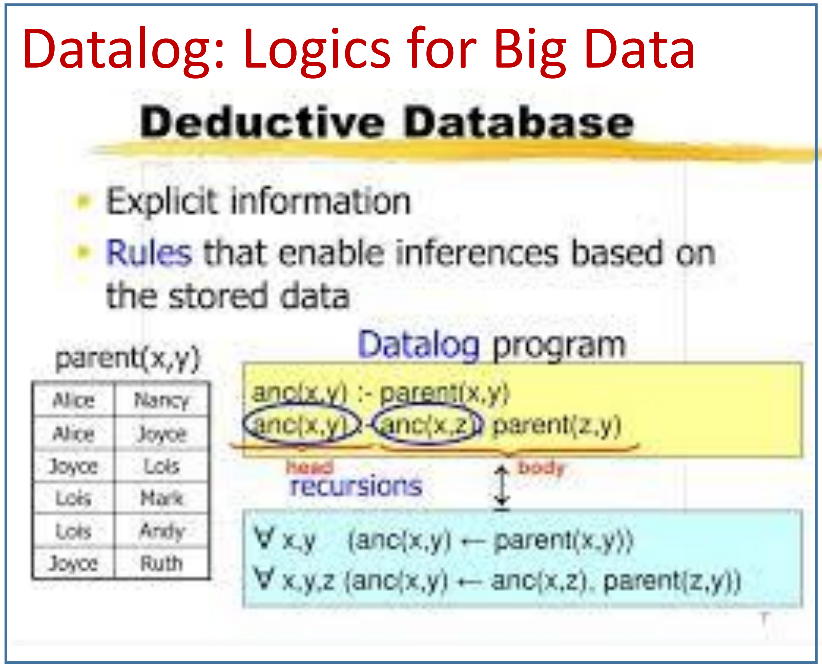
Examples:



## Temporal logic (CTL)

- Propositional logic:
  - p, true, false, $\varphi \wedge \psi$ , $\varphi \vee \psi$ , $\neg\varphi$
- One step in the future
  - One path: $EX\varphi$
  - All paths: $AX\varphi$
- Multiple steps in the future
  - One path: $E(\varphi \cup \psi)$
  - All paths: $A(\varphi \cup \psi)$
- Backwards variants
  - $AX\triangle\varphi$, $A\triangle(\varphi \cup \psi)$, etc.

$A(\neg p \cup q)$



## Datalog: Logics for Big Data

### Deductive Database

- Explicit information
- Rules that enable inferences based on the stored data

parent(x,y)

| | |
|---|---|
| Alice | Nancy |
| Alice | Joyce |
| Joyce | Lois |
| Lois | Mark |
| Lois | Andy |
| Joyce | Ruth |

Datalog program

```
anc(x,y) :- parent(x,y)
anc(x,y) :- anc(x,z), parent(z,y)
```

recursions

$\forall$ x,y (anc(x,y) ← parent(x,y))

$\forall$ x,y,z (anc(x,y) ← anc(x,z), parent(z,y))

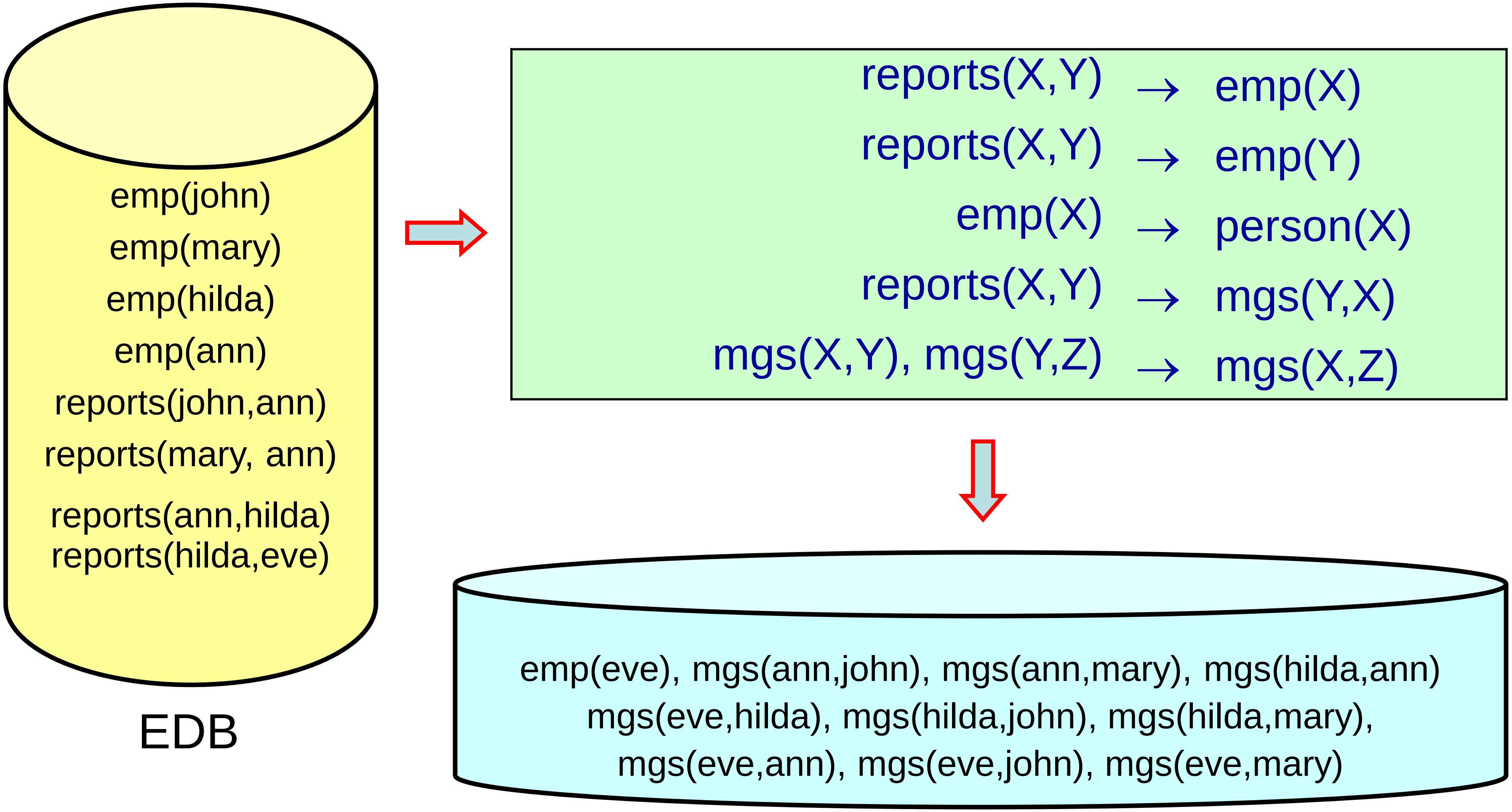
Reasoning with these logics is decidable, their exact complexity is known, and they can be directly implemented as computer programs.

→ Suitable for industrial applications.

## Make Logic Work!

Many variants and many other logics/algorithms (e.g. **SAT-Solving, OWL**) have been designed.

# Datalog - An Example



emp(john)
emp(mary)
emp(hilda)
emp(ann)
reports(john,ann)
reports(mary, ann)
reports(ann,hilda)
reports(hilda,eve)

reports(X,Y) → emp(X)
reports(X,Y) → emp(Y)
emp(X) → person(X)
reports(X,Y) → mgs(Y,X)
mgs(X,Y), mgs(Y,Z) → mgs(X,Z)

emp(eve), mgs(ann,john), mgs(ann,mary), mgs(hilda,ann)
mgs(eve,hilda), mgs(hilda,john), mgs(hilda,mary),
mgs(eve,ann), mgs(eve,john), mgs(eve,mary)

EDB

IDB

Unique Name Assumption
implicitly made, i.e., john ≠ mary etc.

# Logic in Industry…

… is dramatically gaining momentum, and more jobs in LogiCS are becoming available.

SIEMENS
Austria

LogicBlox

Predictix

Automatic configuration of:

- Railway safety systems, telecommunication systems, etc.
- Model-based methods, answer-set programming, SAT solving.

- Based on **Datalog**
- Software for business applications
- about 100 employees including 20-30 computational logicians
- $40m investment
- Recently sold to INFOR

# Datalog Startup

semmle

- Based on **Datalog**
- Software engineering, particularly **code analysis** and **code querying**
- More than fifty employees
- Founded 2006 by **Oege de Moor** as spin-out of Oxford University (total investment $21M)
- Now headquarter in San Francisco and offices in Oxford, New York City, and Copenhagen.
- Eager to hire people whose skills include:
  - Database theory
  - Declarative/logic programming (particularly **Datalog**)
  - Formal logic, lattice theory, abstract interpretation

*Recently sold to Github/Microsoft*

# Logic in the Industry

diffblue

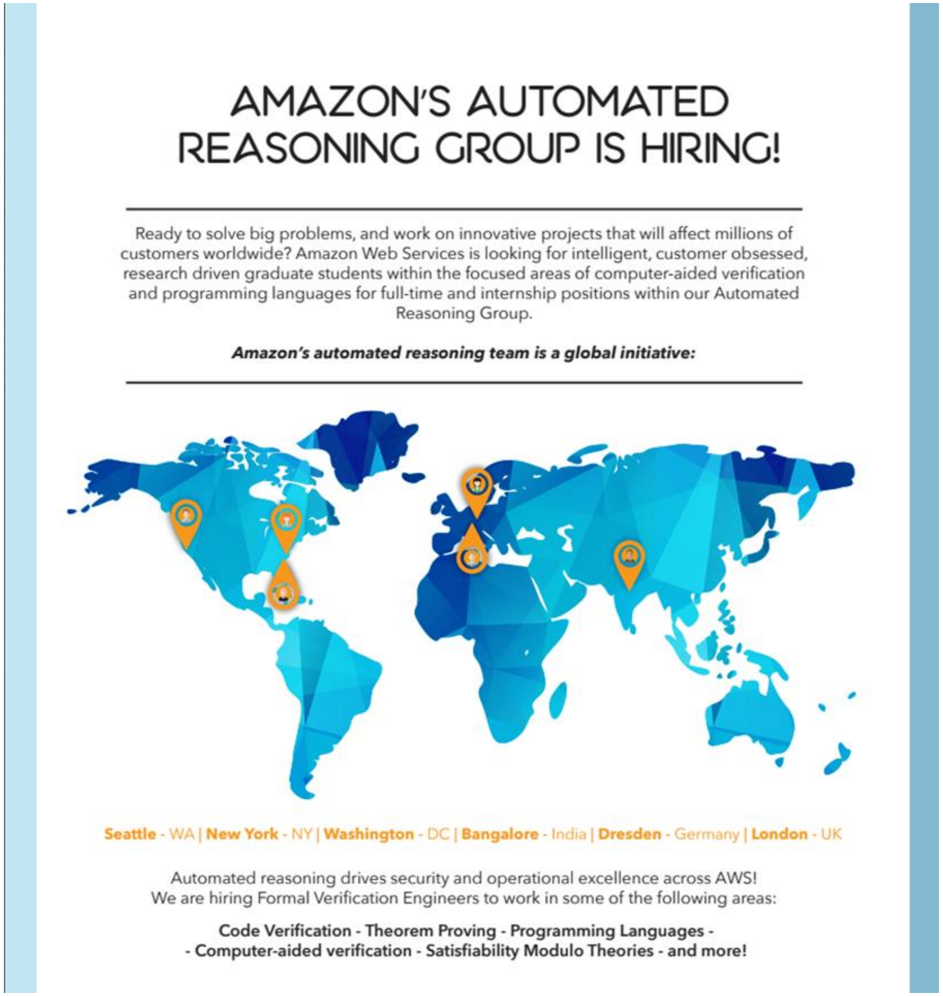
We write tests for you. Automatically

Founded 2016 by **Daniel Kroening** as spin-out of Oxford University.

Logic-based **verification tools** and **software test automation**.

Funding: $22M by four investors;

Eagerly looking for LogiCS PhDs.



AMAZON'S AUTOMATED REASONING GROUP IS HIRING!

Ready to solve big problems, and work on innovative projects that will affect millions of customers worldwide? Amazon Web Services is looking for intelligent, customer obsessed, research driven graduate students within the focused areas of computer-aided verification and programming languages for full-time and internship positions within our Automated Reasoning Group.

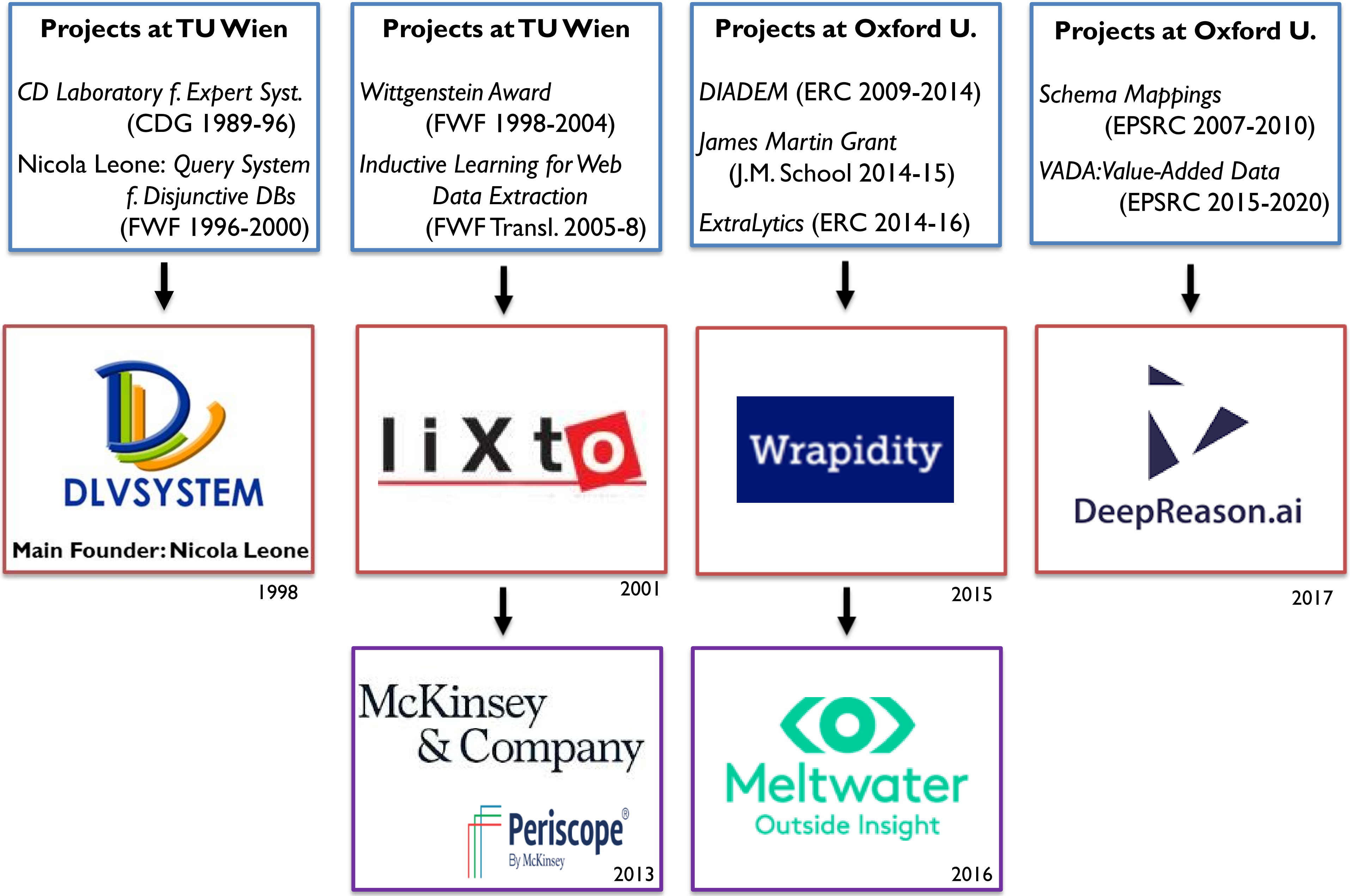
***Amazon's automated reasoning team is a global initiative:***

**Seattle** - WA | **New York** - NY | **Washington** - DC | **Bangalore** - India | **Dresden** - Germany | **London** - UK

Automated reasoning drives security and operational excellence across AWS!
We are hiring Formal Verification Engineers to work in some of the following areas:

**Code Verification - Theorem Proving - Programming Languages -**
**- Computer-aided verification - Satisfiability Modulo Theories - and more!**

# My own Experience as Founder: Co-founded 4 Companies



**Projects at TU Wien**

*CD Laboratory f. Expert Syst.* (CDG 1989-96)

Nicola Leone: *Query System f. Disjunctive DBs* (FWF 1996-2000)

DLVSYSTEM

Main Founder: Nicola Leone

1998

**Projects at TU Wien**

*Wittgenstein Award* (FWF 1998-2004)

*Inductive Learning for Web Data Extraction* (FWF Transl. 2005-8)

lixto

2001

McKinsey & Company

Periscope® By McKinsey

2013

**Projects at Oxford U.**

*DIADEM* (ERC 2009-2014)

*James Martin Grant* (J.M. School 2014-15)

*ExtraLytics* (ERC 2014-16)

Wrapidity

2015

Meltwater

Outside Insight

2016

**Projects at Oxford U.**

*Schema Mappings* (EPSRC 2007-2010)

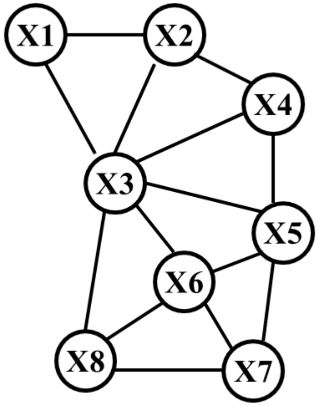
*VADA:Value-Added Data* (EPSRC 2015-2020)

DeepReason.ai

2017

# 1. From Disjunctive Datalog Research to DLVSYSTEM

# 1. From Disjunctive Datalog Research to DLV-System



**edge(X,Y) → edge(Y,X)**

**edge(X,Y) → node(X)**

**edge(X,Y) → node(Y)**

**node(X) → color(X, red) ∨ color(X, blue) ∨ color(X, green)**

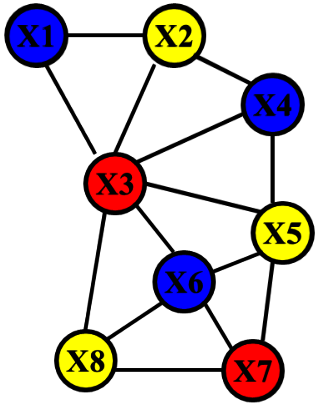
**color(X, red) ∧ color(X, blue) → false**

**color(X, red) ∧ color(X, green) → false**

**color(X, green) ∧ color(X, blue)→ false**

**edge(X,Y) ∧ color(X,C) ∧ color(Y,C) → false**

# 1. From Disjunctive Datalog Research to DLV-System



edge(**X**,**Y**) → edge(**Y**,**X**)

edge(**X**,**Y**) → node(**X**)

edge(**X**,**Y**) → node(**Y**)

node(**X**) → color(**X**, red) ∨ color(**X**, blue) ∨ color(**X**, green)

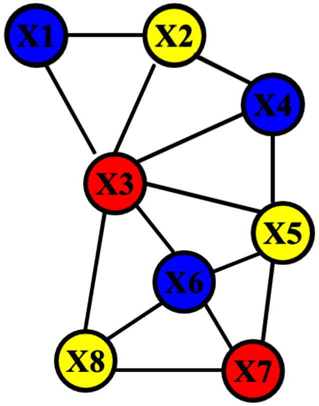
color(**X**, red) ∧ color(**X**, blue) → false

color(**X**, red) ∧ color(**X**, green) → false

color(**X**, green) ∧ color(**X**, blue)→ false

edge(**X**,**Y**) ∧ color(**X**,**C**) ∧ color(**Y**,**C**) → false

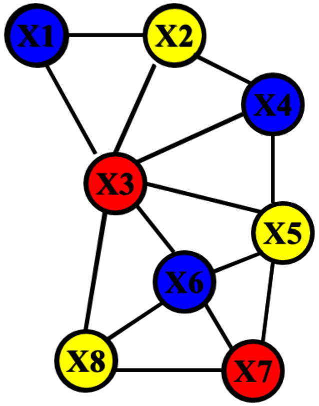
# 1. From Disjunctive Datalog Research to DLV-System



**edge(X,Y) → edge(Y,X)**
**edge(X,Y) → node(X)**
**edge(X,Y) → node(Y)**
**node(X) → color(X, red) ∨ color(X, blue) ∨ color(X, green)**
**color(X, red) ∧ color(X, blue) → false**
**color(X, red) ∧ color(X, green) → false**
**color(X, green) ∧ color(X, blue)→ false**
**edge(X,Y) ∧ color(X,C) ∧ color(Y,C) → false**

Example of reasoning question:
Are there necessarily two nodes of every color?

# 1. From Disjunctive Datalog Research to DLV-System



edge(**X**,Y) → edge(Y,**X**)

edge(**X**,Y) → node(**X**)

edge(**X**,Y) → node(Y)

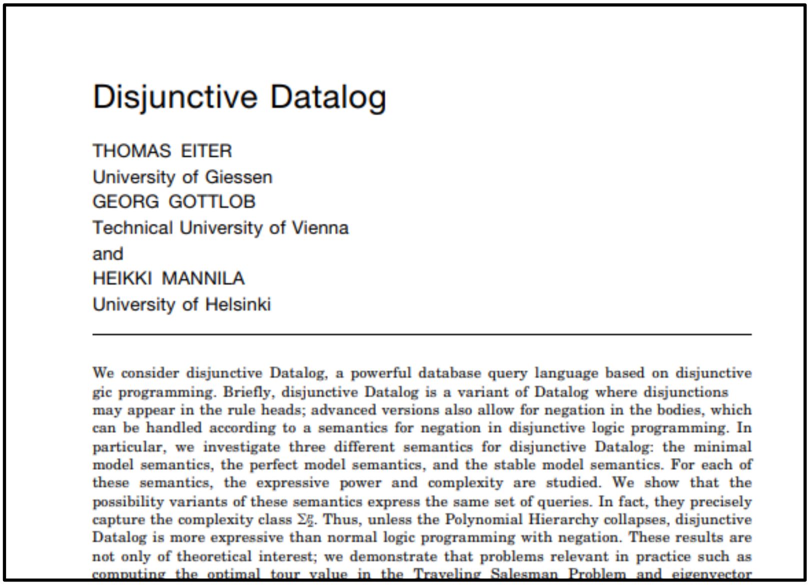
node(**X**) → color(**X**, red) ∨ color(**X**, blue) ∨ color(**X**, green)

color(**X**, red) ∧ color(**X**, blue) → false

color(**X**, red) ∧ color(**X**, green) → false

color(**X**, green) ∧ color(**X**, blue)→ false

edge(**X**,Y) ∧ color(**X**,C) ∧ color(Y,C) → false



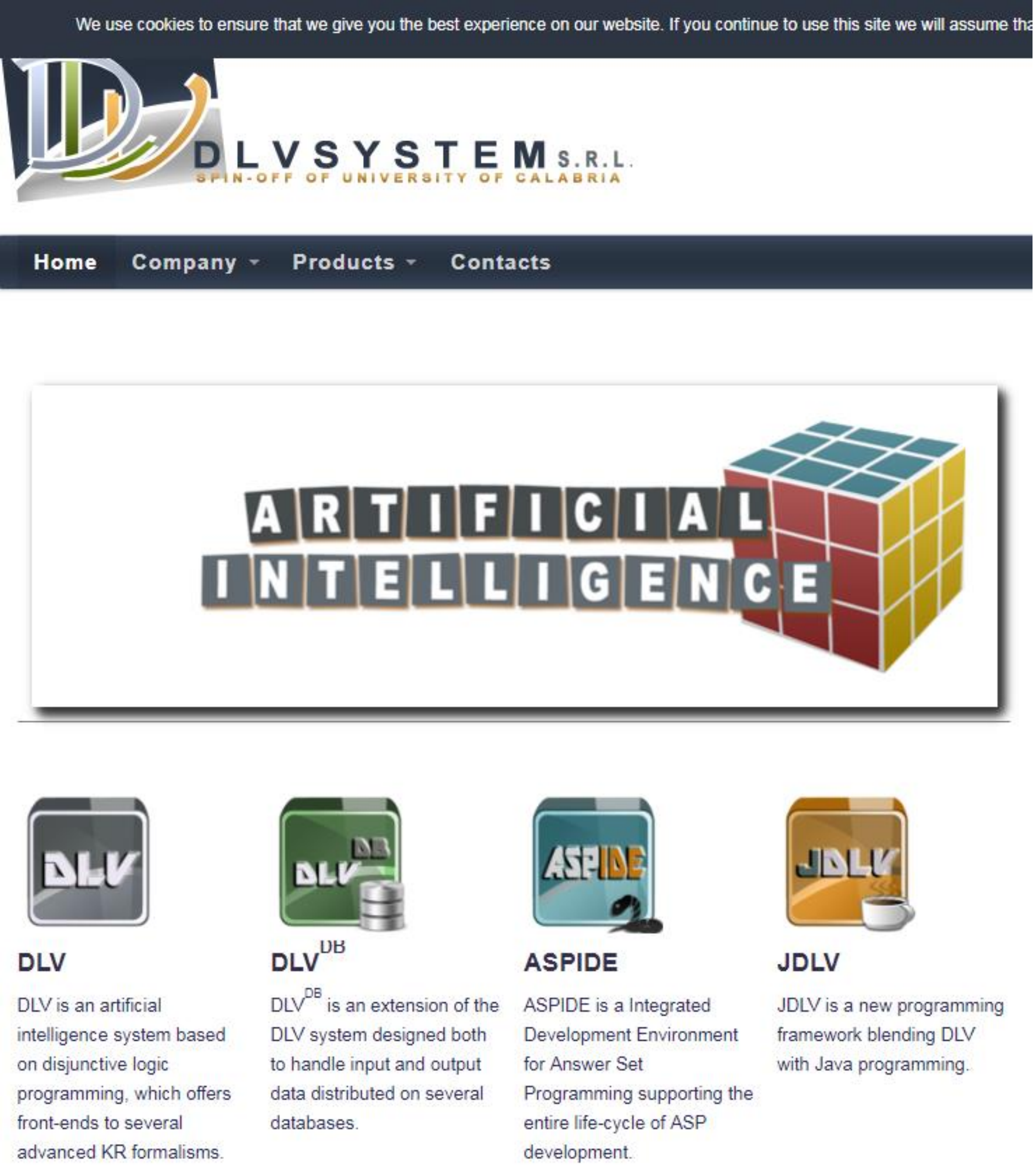
## Disjunctive Datalog

THOMAS EITER
University of Giessen
GEORG GOTTLOB
Technical University of Vienna
and
HEIKKI MANNILA
University of Helsinki

We consider disjunctive Datalog, a powerful database query language based on disjunctive gic programming. Briefly, disjunctive Datalog is a variant of Datalog where disjunctions may appear in the rule heads; advanced versions also allow for negation in the bodies, which can be handled according to a semantics for negation in disjunctive logic programming. In particular, we investigate three different semantics for disjunctive Datalog: the minimal model semantics, the perfect model semantics, and the stable model semantics. For each of these semantics, the expressive power and complexity are studied. We show that the possibility variants of these semantics express the same set of queries. In fact, they precisely capture the complexity class $\Sigma_2^P$. Thus, unless the Polynomial Hierarchy collapses, disjunctive Datalog is more expressive than normal logic programming with negation. These results are not only of theoretical interest; we demonstrate that problems relevant in practice such as computing the optimal tour value in the Traveling Salesman Problem and eigenvector

ACM Transact. on Database Systems 22(3): 364-418 (1997)

Theorem:
Disjunctive Datalog reasoning is $\Pi_2^P$ –complete and can express every property in $\Pi_2^P$ .



We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume tha

Home Company Products Contacts

### DLV

DLV is an artificial intelligence system based on disjunctive logic programming, which offers front-ends to several advanced KR formalisms.

### $\text{DLV}^{\text{DB}}$

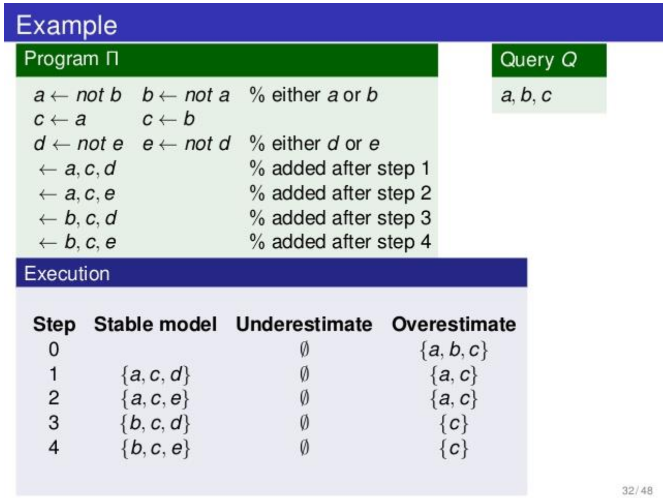
$\text{DLV}^{\text{DB}}$ is an extension of the DLV system designed both to handle input and output data distributed on several databases.

### ASPIDE

ASPIDE is a Integrated Development Environment for Answer Set Programming supporting the entire life-cycle of ASP development.

### JDLV

JDLV is a new programming framework blending DLV with Java programming.



## Example

Program Π

| | | |
|---|---|---|
| $a \leftarrow not\ b$ | $b \leftarrow not\ a$ | % either $a$ or $b$ |
| $c \leftarrow a$ | $c \leftarrow b$ | |
| $d \leftarrow not\ e$ | $e \leftarrow not\ d$ | % either $d$ or $e$ |
| $\leftarrow a, c, d$ | | % added after step 1 |
| $\leftarrow a, c, e$ | | % added after step 2 |
| $\leftarrow b, c, d$ | | % added after step 3 |
| $\leftarrow b, c, e$ | | % added after step 4 |

Query $Q$

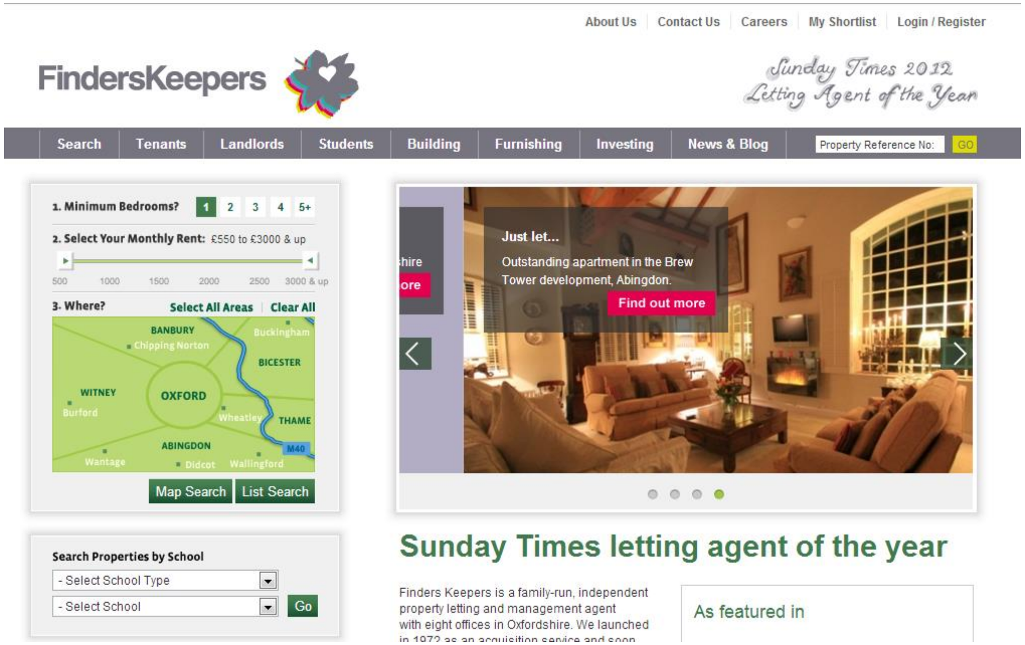
$a, b, c$

Execution

| Step | Stable model | Underestimate | Overestimate |
|---|---|---|---|
| 0 | | $\emptyset$ | $\{a, b, c\}$ |
| 1 | $\{a, c, d\}$ | $\emptyset$ | $\{a, c\}$ |
| 2 | $\{a, c, e\}$ | $\emptyset$ | $\{a, c\}$ |
| 3 | $\{b, c, d\}$ | $\emptyset$ | $\{c\}$ |
| 4 | $\{b, c, e\}$ | $\emptyset$ | $\{c\}$ |

# 2. From Monadic Datalog Research to Lixto

# Web Data Extraction



Outstanding apartment in the Brew Tower development, Abingdon.

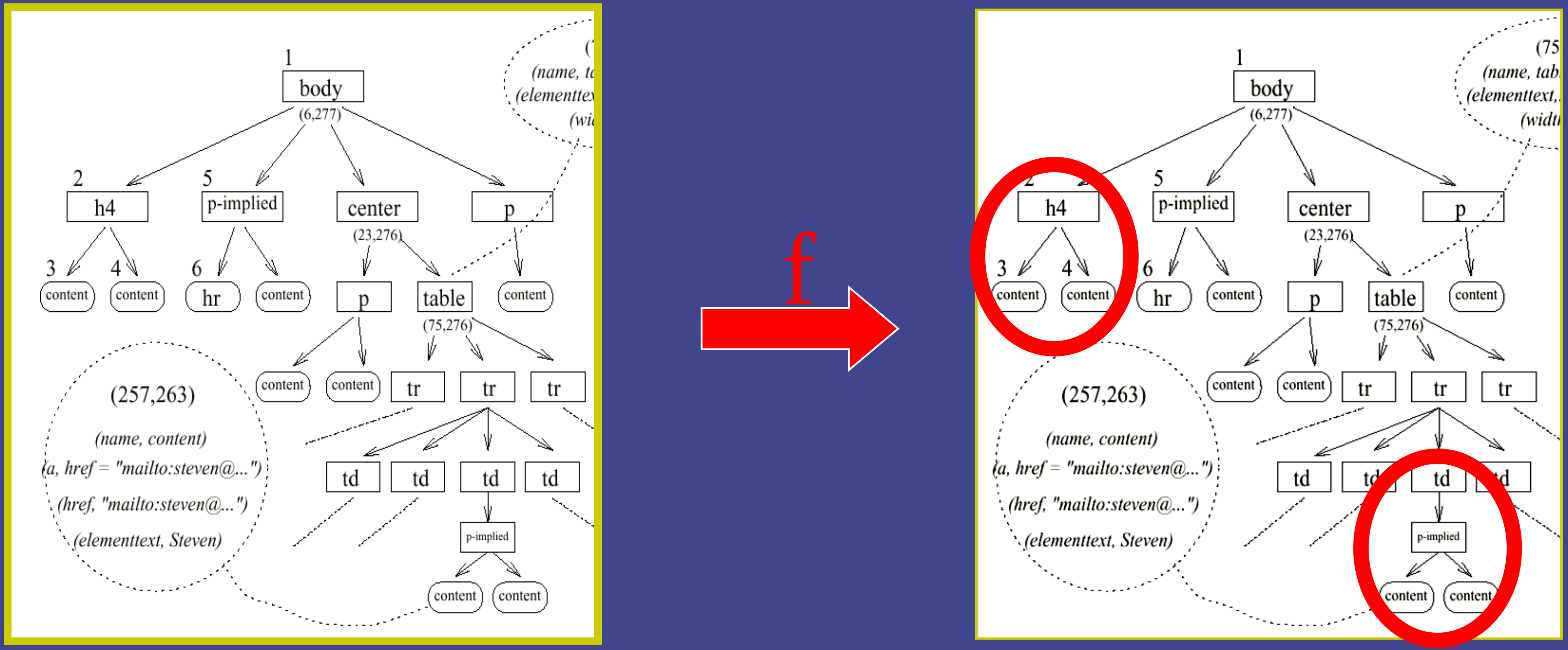
Sunday Times letting agent of the year

Finders Keepers is a family-run, independent property letting and management agent with eight offices in Oxfordshire. We launched in 1972 as an acquisition service and soon

| ref-code | postcode | bedrooms | bathrooms | available | price |
|---|---|---|---|---|---|
| 33453 | OX2 6AR | 3 | 2 | 15/10/2013 | £1280 pcm |
| 33433 | OX4 7DG | 2 | 1 | 18/04/2013 | £995 pcm |

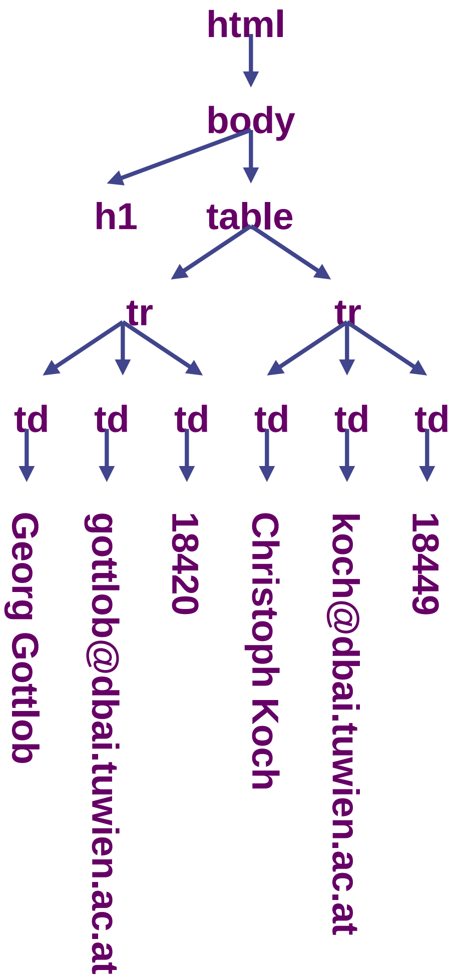
# Function f: HTML Parse tree → Subtrees

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>
```

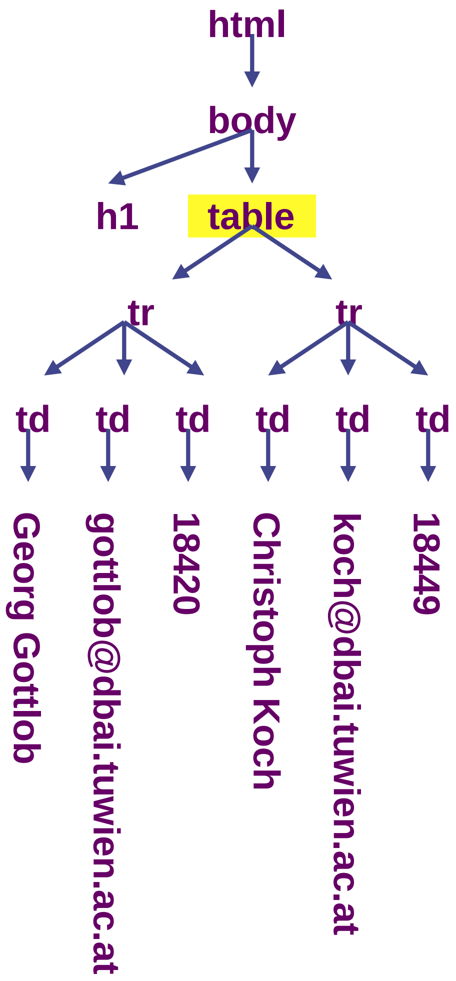


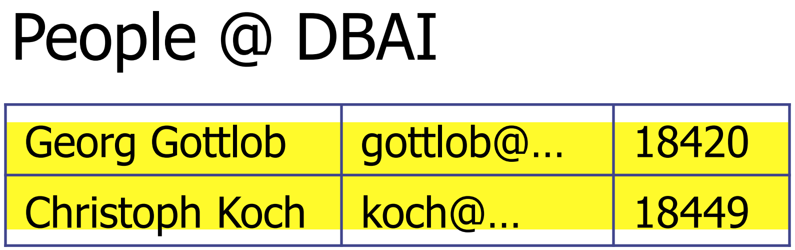
People @ DBAI

| | | |
|---|---|---|
| Georg Gottlob | gottlob@… | 18420 |
| Christoph Koch | koch@… | 18449 |

# Predicate *employeetable*

```
<!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

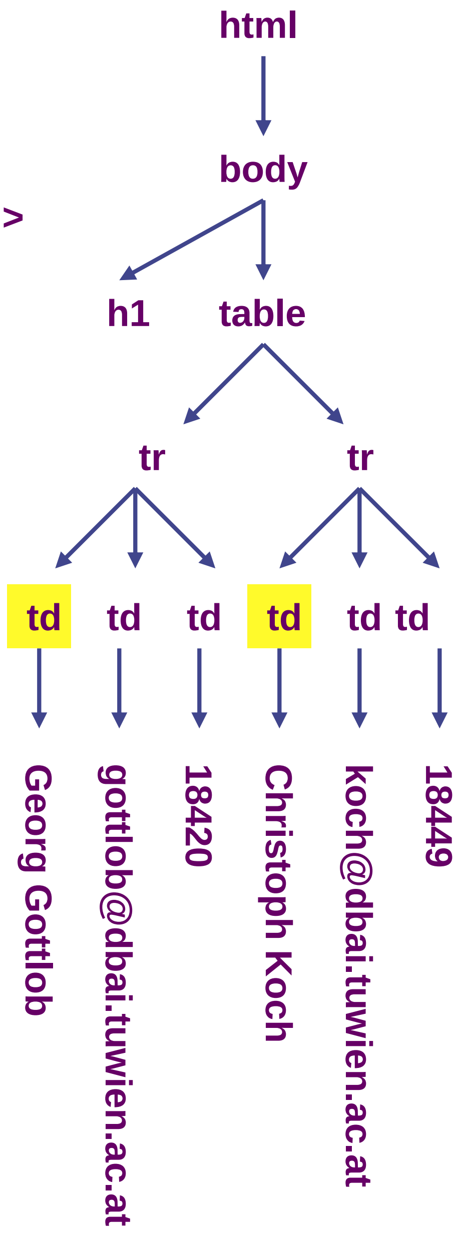
| | | |
|---|---|---|
| Georg Gottlob | gottlob@… | 18420 |
| Christoph Koch | koch@… | 18449 |

# Predicate *employee*

```
<!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>
```

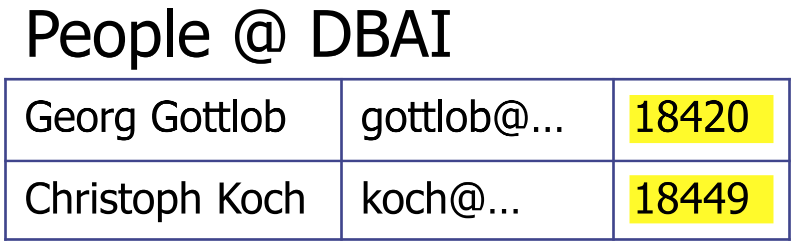
People @ DBAI

| | | |
|---|---|---|
| Georg Gottlob | gottlob@... | 18420 |
| Christoph Koch | koch@... | 18449 |



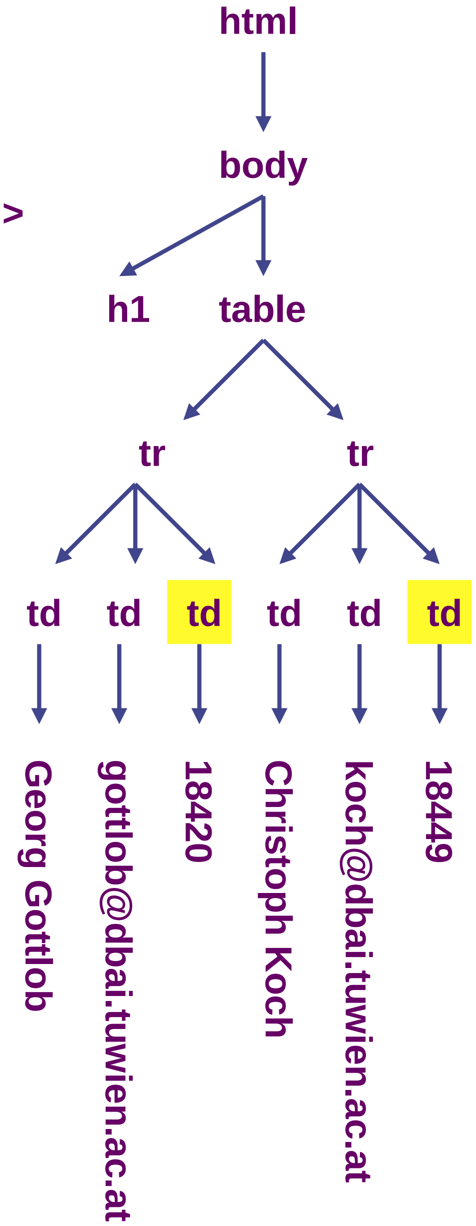
# Predicate *phone*

```
<!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>
```



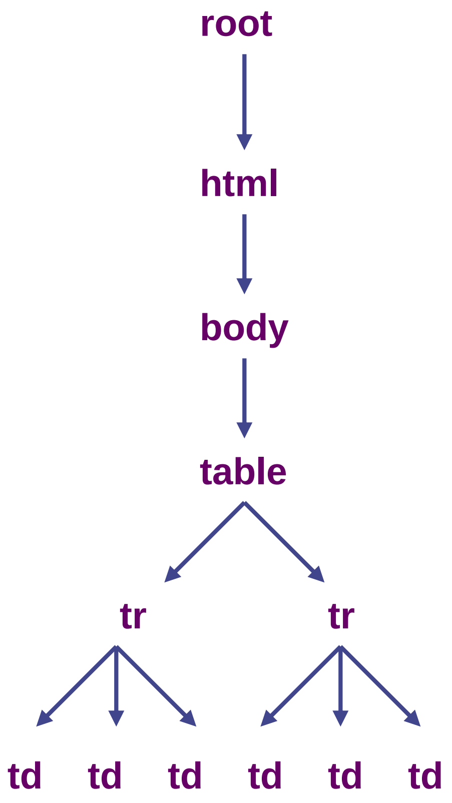
# People @ DBAI

| | | |
|---|---|---|
| Georg Gottlob | gottlob@... | 18420 |
| Christoph Koch | koch@... | 18449 |



# 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).
```



root

html

body

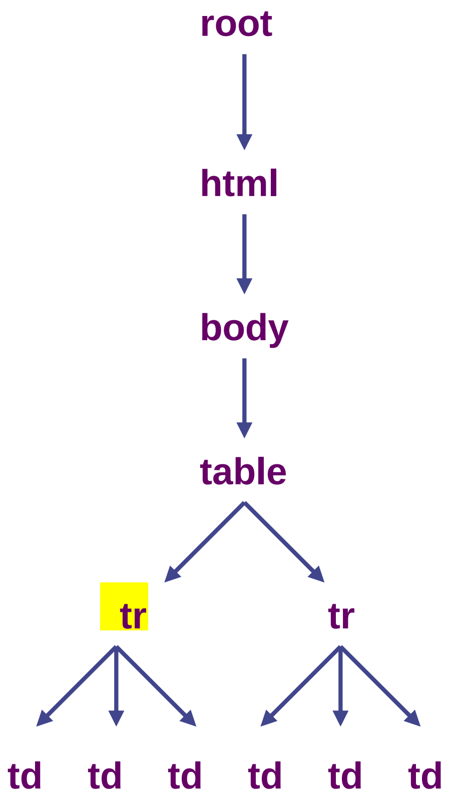
table

tr tr

td td td td td td

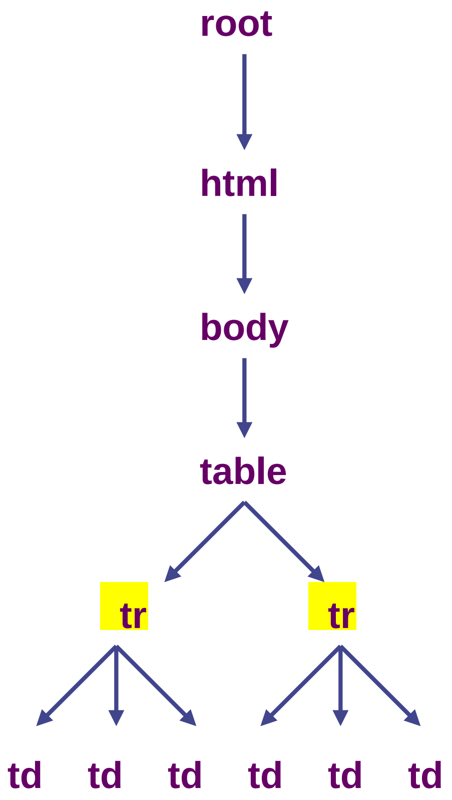
# Monadic Datalog as a Wrapping Language

```
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).
```



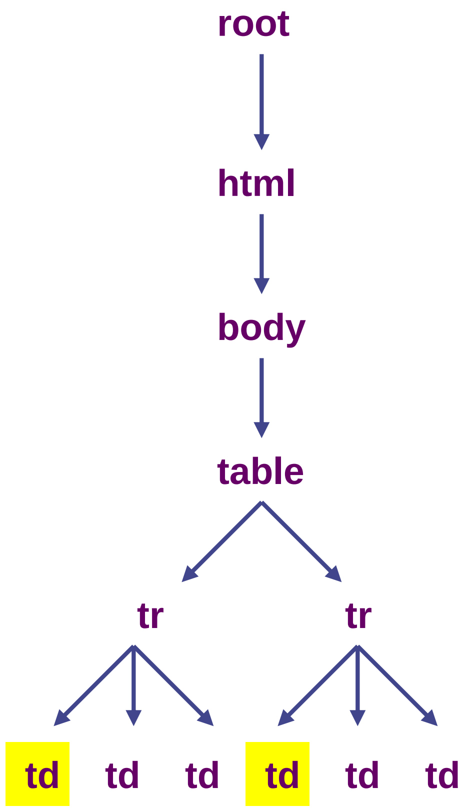
# Monadic Datalog as a Wrapping Language

```
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).
```

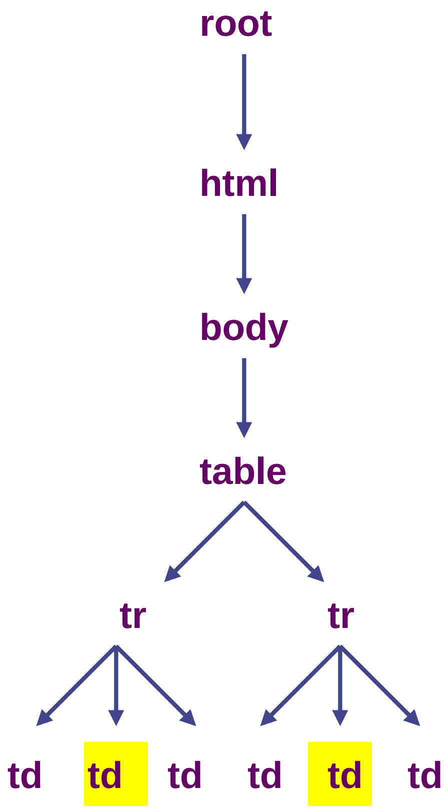
```
email(X) :- name(N), nextsibling(N, X), label[td](X).
```

```
phone(X) :- email(M), nextsibling(M, X), label[td](X).
```



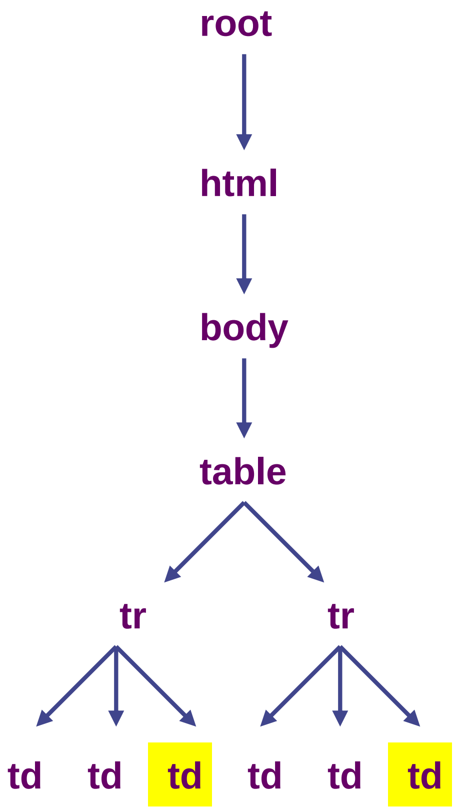
# Monadic Datalog as a Wrapping Language

```
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).
```



# Theoretical Results

Theorem1: Monadic Datalog over trees has combined complexity: O(|data|*|query|)

## Monadic Datalog and the Expressive Power of Languages for Web Information Extraction

GEORG GOTTLOB and CHRISTOPH KOCH
Technische Universität Wien, Austria

Research on information extraction from Web pages (wrapping) has seen much activity recently (particularly systems implementations), but little work has been done on formally studying the expressiveness of the formalisms proposed or on the theoretical foundations of wrapping. In this paper, we first study monadic datalog over trees as a wrapping language. We show that this simple language is equivalent to monadic second order logic (MSO) in its ability to specify wrappers. We believe that MSO has the right expressiveness required for Web information extraction and propose MSO as a yardstick for evaluating and comparing wrappers. Along the way, several other results on the complexity of query evaluation and query containment for monadic datalog over trees are established, and a simple normal form for this language is presented. Using the above results, we subsequently study the kernel fragment Elog$^-$ of the Elog wrapping language used in the Lixto system (a visual wrapper generator). Curiously, Elog$^-$ exactly captures MSO, yet is easier to use. Indeed, programs in this language can be entirely visually specified.

Categories and Subject Descriptors: F.1.1 [**Computation by Abstract Devices**]: Automata; F.4.1 [**Mathematical Logic and Formal Languages**]: Computational Logic; F.4.3 [**Mathematical Logic and Formal Languages**]: Classes defined by grammars or automata; H.2.3 [**Database Management**]: Query languages; I.7.2 [**Document Preparation**]: Markup languages

General Terms: Theory, Languages, Algorithms

Additional Key Words and Phrases: Complexity, Expressiveness, HTML, Information Extraction, Monadic Datalog, MSO, Regular Tree Languages, Web Wrapping

### 1. INTRODUCTION

The Web wrapping problem, i.e., the problem of extracting structured information from HTML documents, is one of high practical importance and has spurred a great

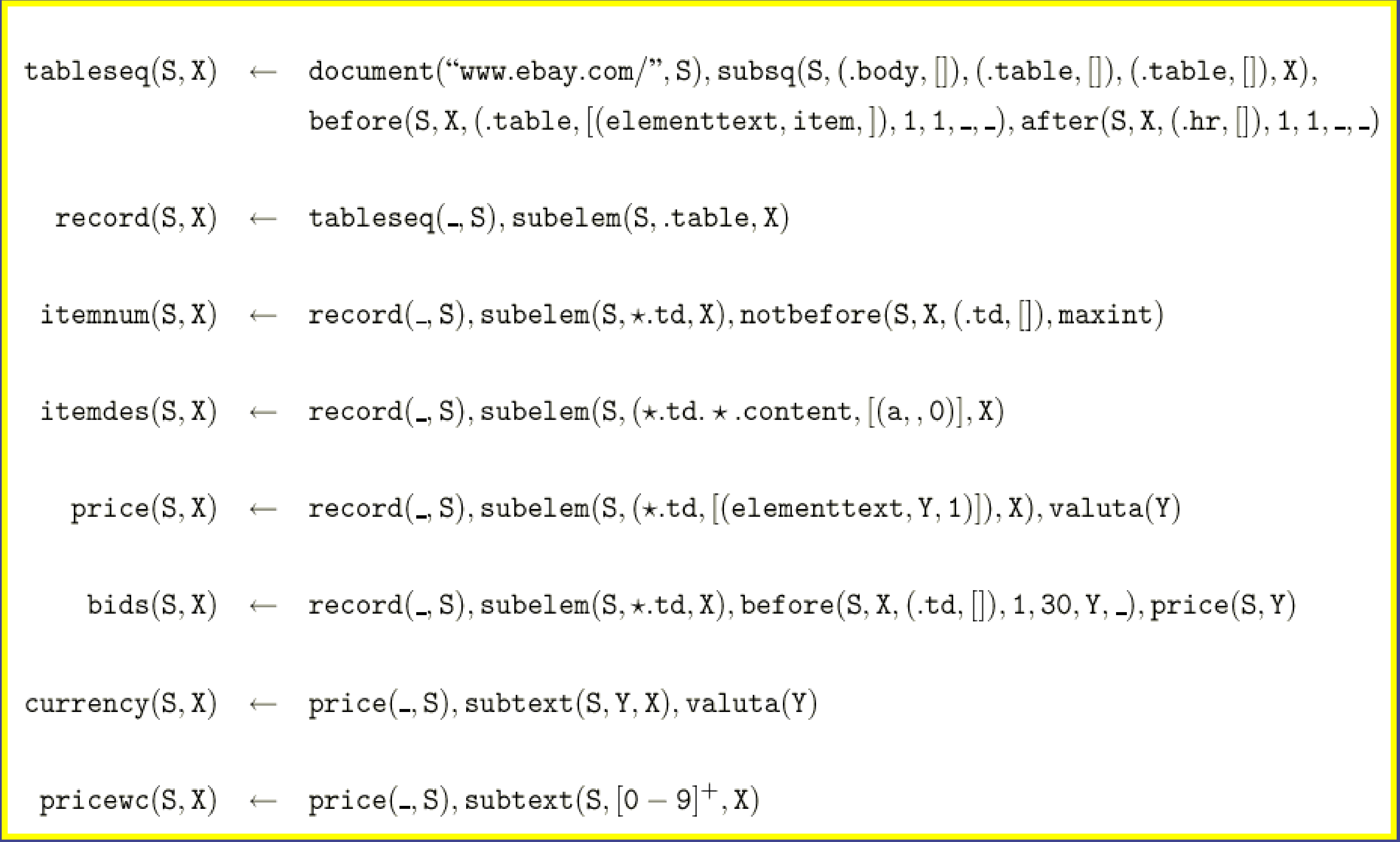
J. ACM 51(1): 74-113 (2004)

Theorem 2 : Over trees, Monadic Datalog = MSO

MSO= Monadic second order logic.

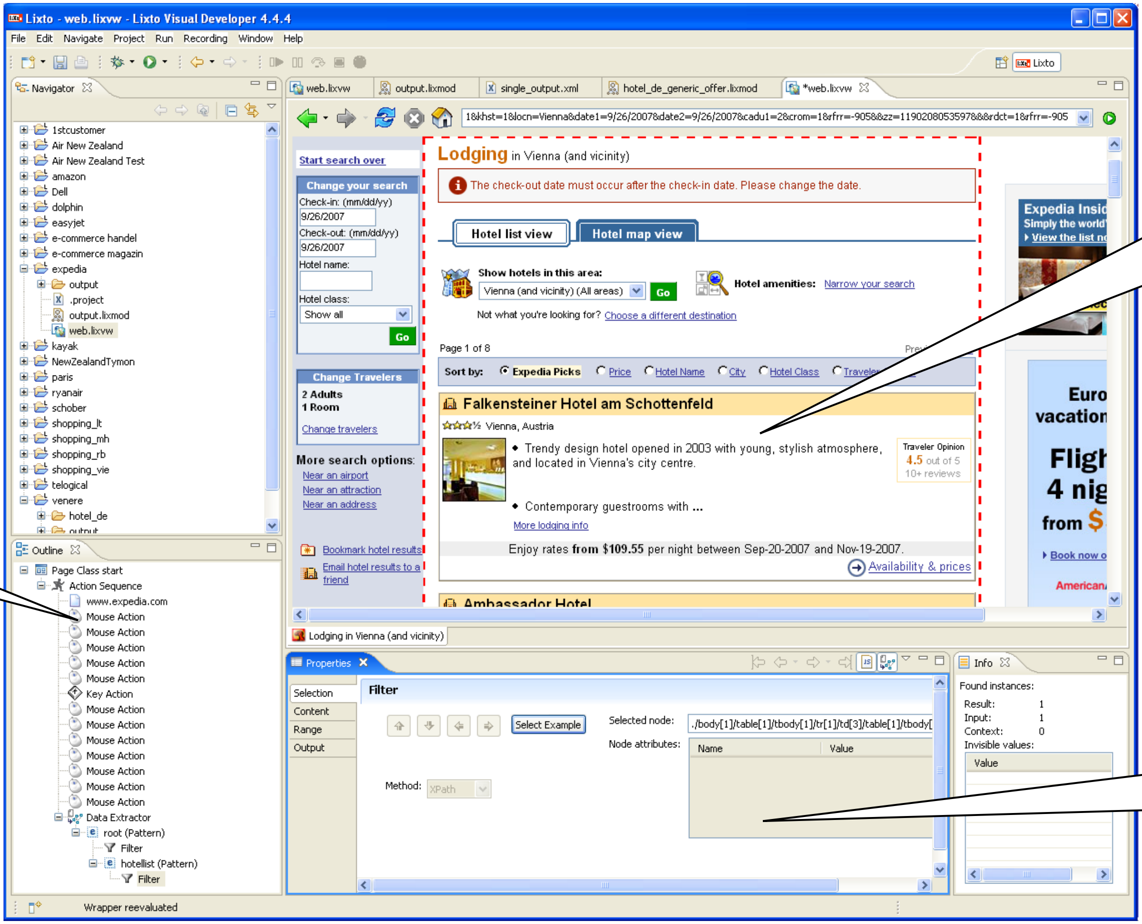
A unary query is definable
in MSO iff it is definable
via a monadic datalog program.

# ELOG Program for eBay pages



```
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)


Lixto - web.lixvw - Lixto Visual Developer 4.4.4
File Edit Navigate Project Run Recording Window Help
Lixto
Navigator
Outline
web.lixvw
output.lixmod
single_output.xml
hotel_de_generic_offer.lixmod
*web.lixvw
Properties
Info

1&khst=1&locn=Vienna&date1=9/26/2007&date2=9/26/2007&cadu1=2&crom=1&rfrr=-905&&zz=1190208053597&&&rdct=1&rfrr=-905

Navigator: 1stcustomer, Air New Zealand, Air New Zealand Test, amazon, Dell, dolphin, easyjet, e-commerce handel, e-commerce magazin, expedia, output, .project, output.lixmod, web.lixvw, kayak, NewZealandTymon, paris, ryanair, schober, shopping_lt, shopping_mh, shopping_rb, shopping_vie, telogical, venere, hotel_de, output

Outline: Page Class start, Action Sequence, www.expedia.com, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Key Action, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Mouse Action, Data Extractor, root (Pattern), Filter, hotellist (Pattern), Filter

Start search over

Change your search
Check-in: (mm/dd/yy) 9/26/2007
Check-out: (mm/dd/yy) 9/26/2007
Hotel name:
Hotel class: Show all
Go

Change Travelers
2 Adults
1 Room
Change travelers

More search options:
Near an airport
Near an attraction
Near an address

Bookmark hotel results
Email hotel results to a friend

Lodging in Vienna (and vicinity)

The check-out date must occur after the check-in date. Please change the date.

Hotel list view | Hotel map view

Show hotels in this area: Vienna (and vicinity) (All areas) Go
Hotel amenities: Narrow your search
Not what you're looking for? Choose a different destination

Page 1 of 8

Sort by: ● Expedia Picks ○ Price ○ Hotel Name ○ City ○ Hotel Class ○ Traveler

Falkensteiner Hotel am Schottenfeld
Vienna, Austria
- Trendy design hotel opened in 2003 with young, stylish atmosphere, and located in Vienna's city centre.
- Contemporary guestrooms with ...

Traveler Opinion 4.5 out of 5 10+ reviews

More lodging info

Enjoy rates from $109.55 per night between Sep-20-2007 and Nov-19-2007.

Availability & prices

Ambassador Hotel

Expedia Insid... Simply the world... View the list n...
Euro vacation
Fligh 4 nig from $
Book now o
American

Lodging in Vienna (and vicinity)

Filter
Selection
Content
Range
Output
Select Example
Selected node: ./body[1]/table[1]/tbody[1]/tr[1]/td[3]/table[1]/tbody[
Node attributes:

| Name | Value |
|---|---|

Method: XPath

Found instances:
Result: 1
Input: 1
Context: 0
Invisible values:

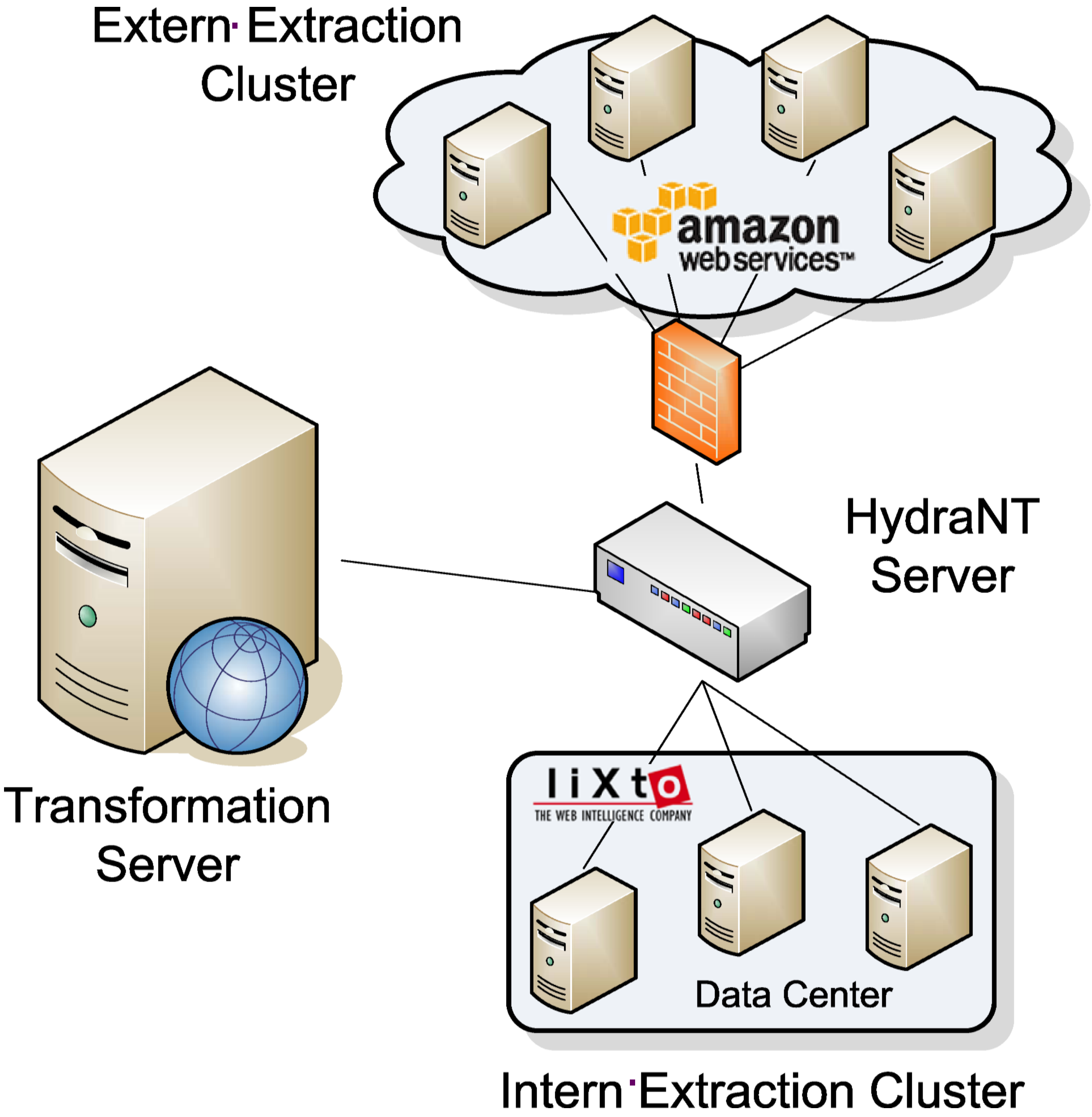
| Value |
|---|

Wrapper reevaluated

Mozilla Web Browser

Navigation Steps

Extraction Configuration



Extern Extraction Cluster

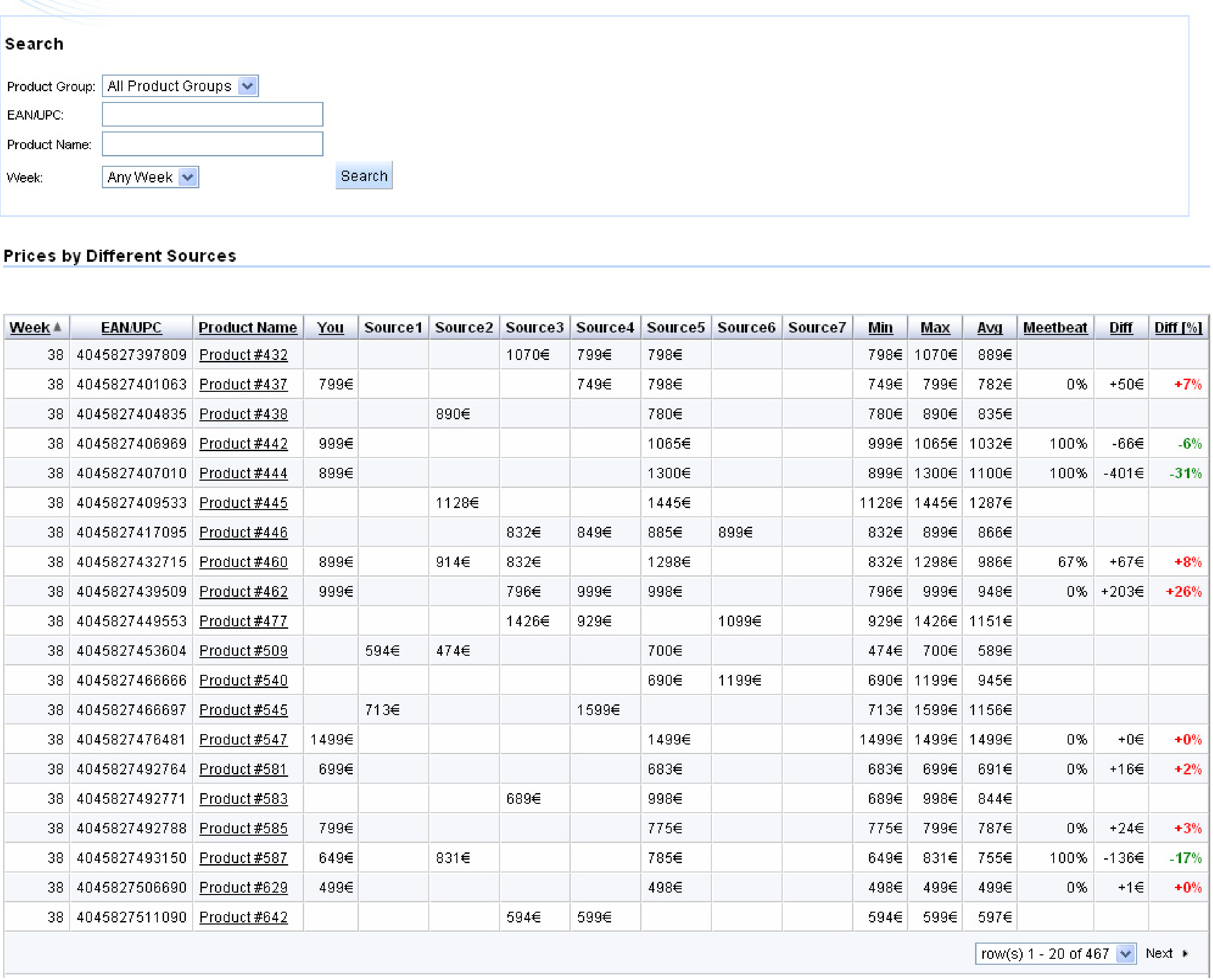
HydraNT Server

Transformation Server

Data Center

Intern Extraction Cluster

SONY
make.believe
IBEROSTAR
HOTELS & RESORTS
FUJITSU
Insight | Technology Solutions
norfolkline.com
WKO
Shopping.com
France
hama
DIE PASSENDE LÖSUNG
ethix™
ZF
Verbund
SAP®
dfv VERLAGSGRU
DEUTSCHER
CONRAD
BUSINESS
PIRELLI
BOOKING.COM
online hotel reservations
MAGNA
T · · Mobile ·
Voss
mondi
business paper
eniro
gta
by Travelport
ThyssenKrupp
TKP
Webasto
Feel the drive
hotel.de®
simply book for less
toerisme
vlaanderen
IBM®
F-IT
INA
l'tur
Last Minute
Best Western
Germany
The travel destination
REPVBBLICA ITALIANA
www.austria.inf



### Search

Product Group: All Product Groups

EAN/UPC:

Product Name:

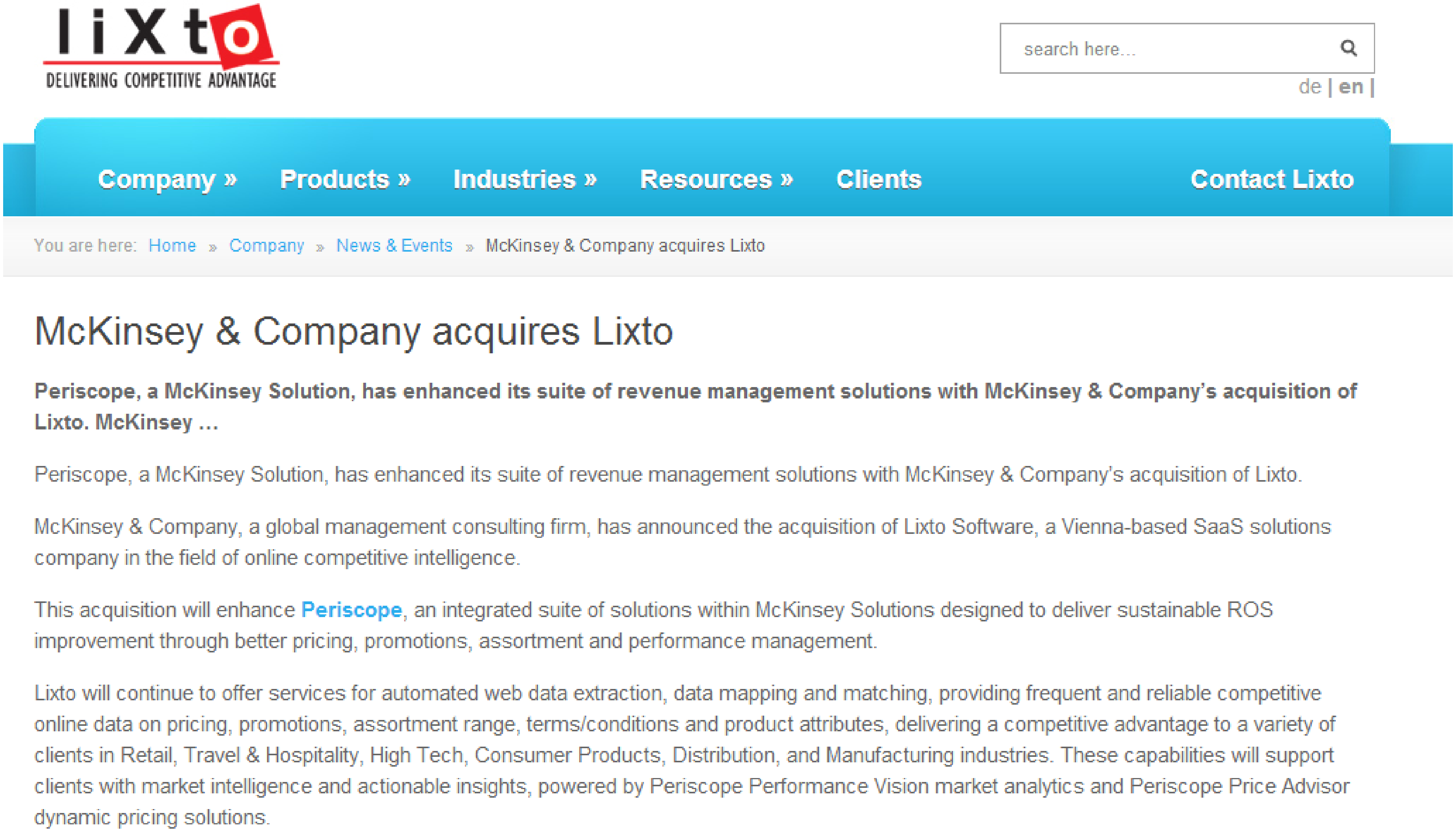
Week: Any Week

### Prices by Different Sources

| Week ▲ | EAN/UPC | Product Name | You | Source1 | Source2 | Source3 | Source4 | Source5 | Source6 | Source7 | Min | Max | Avg | Meetbeat | Diff | Diff [%] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 38 | 4045827397809 | Product #432 | | | | 1070€ | 799€ | 798€ | | | 798€ | 1070€ | 889€ | | | |
| 38 | 4045827401063 | Product #437 | 799€ | | | | 749€ | 798€ | | | 749€ | 799€ | 782€ | 0% | +50€ | +7% |
| 38 | 4045827404835 | Product #438 | | | 890€ | | | 780€ | | | 780€ | 890€ | 835€ | | | |
| 38 | 4045827406969 | Product #442 | 999€ | | | | | 1065€ | | | 999€ | 1065€ | 1032€ | 100% | -66€ | -6% |
| 38 | 4045827407010 | Product #444 | 899€ | | | | | 1300€ | | | 899€ | 1300€ | 1100€ | 100% | -401€ | -31% |
| 38 | 4045827409533 | Product #445 | | | 1128€ | | | 1445€ | | | 1128€ | 1445€ | 1287€ | | | |
| 38 | 4045827417095 | Product #446 | | | | 832€ | 849€ | 885€ | 899€ | | 832€ | 899€ | 866€ | | | |
| 38 | 4045827432715 | Product #460 | 899€ | | 914€ | 832€ | | 1298€ | | | 832€ | 1298€ | 986€ | 67% | +67€ | +8% |
| 38 | 4045827439509 | Product #462 | 999€ | | | 796€ | 999€ | 998€ | | | 796€ | 999€ | 948€ | 0% | +203€ | +26% |
| 38 | 4045827449553 | Product #477 | | | | 1426€ | 929€ | | 1099€ | | 929€ | 1426€ | 1151€ | | | |
| 38 | 4045827453604 | Product #509 | | 594€ | 474€ | | | 700€ | | | 474€ | 700€ | 589€ | | | |
| 38 | 4045827466666 | Product #540 | | | | | | 690€ | 1199€ | | 690€ | 1199€ | 945€ | | | |
| 38 | 4045827466697 | Product #545 | | 713€ | | | 1599€ | | | | 713€ | 1599€ | 1156€ | | | |
| 38 | 4045827476481 | Product #547 | 1499€ | | | | | 1499€ | | | 1499€ | 1499€ | 1499€ | 0% | +0€ | +0% |
| 38 | 4045827492764 | Product #581 | 699€ | | | | | 683€ | | | 683€ | 699€ | 691€ | 0% | +16€ | +2% |
| 38 | 4045827492771 | Product #583 | | | | 689€ | | 998€ | | | 689€ | 998€ | 844€ | | | |
| 38 | 4045827492788 | Product #585 | 799€ | | | | | 775€ | | | 775€ | 799€ | 787€ | 0% | +24€ | +3% |
| 38 | 4045827493150 | Product #587 | 649€ | | 831€ | | | 785€ | | | 649€ | 831€ | 755€ | 100% | -136€ | -17% |
| 38 | 4045827506690 | Product #629 | 499€ | | | | | 498€ | | | 498€ | 499€ | 499€ | 0% | +1€ | +0% |
| 38 | 4045827511090 | Product #642 | | | | 594€ | 599€ | | | | 594€ | 599€ | 597€ | | | |

row(s) 1 - 20 of 467 Next ▸

# *August 2013:*



You are here: Home » Company » News & Events » McKinsey & Company acquires Lixto

## 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.

# 3. From Lixto over DIADEM to Wrapidity

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?

lixto

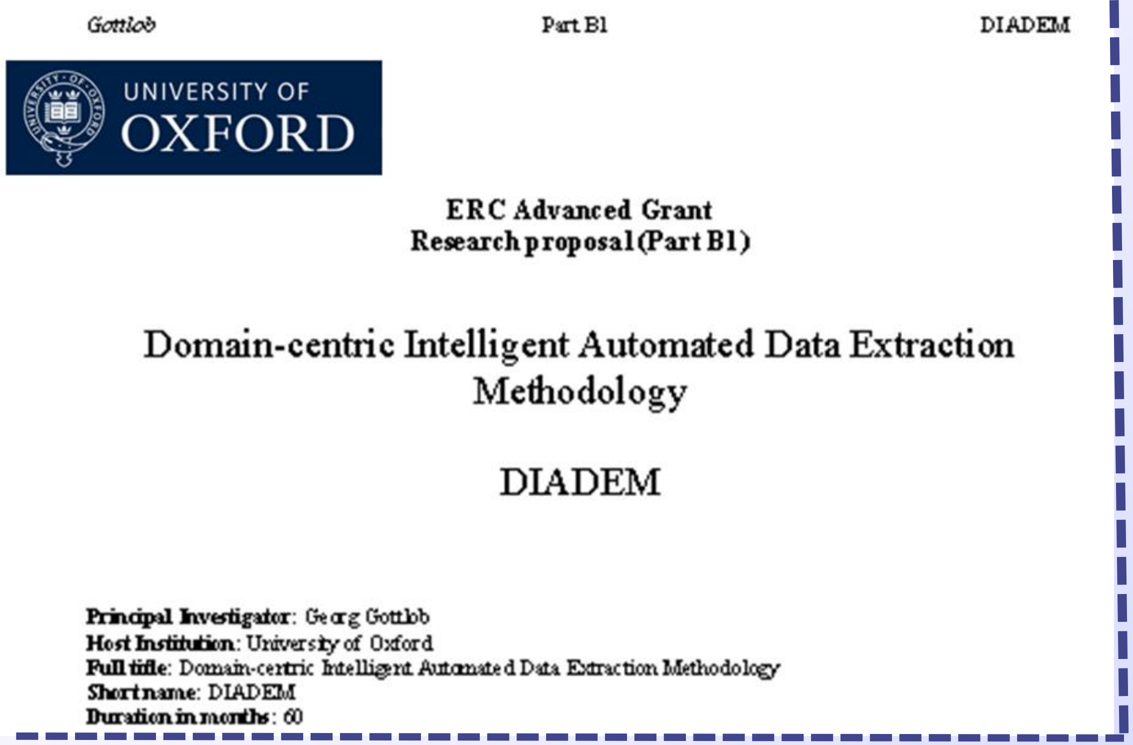
THE WEB INTELLIGENCE COMPANY

# 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



Gottlob Part B1 DIADEM

UNIVERSITY OF OXFORD

**ERC Advanced Grant**
**Research proposal (Part B1)**

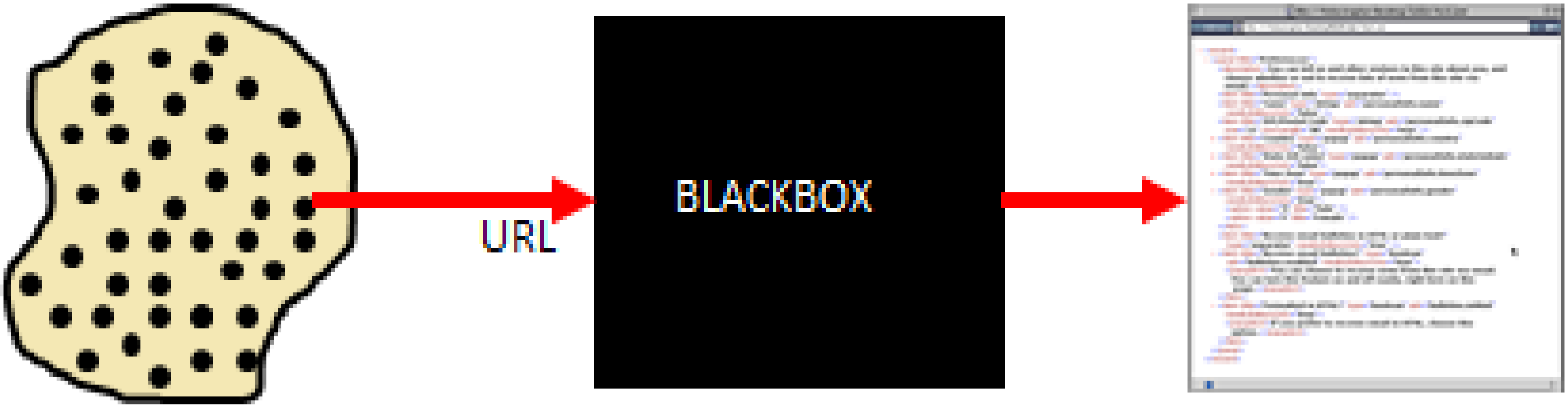
Domain-centric Intelligent Automated Data Extraction Methodology

DIADEM

**Principal Investigator**: Georg Gottlob
**Host Institution**: University of Oxford
**Full title**: Domain-centric Intelligent Automated Data Extraction Methodology
**Short name**: DIADEM
**Duration in months**: 60

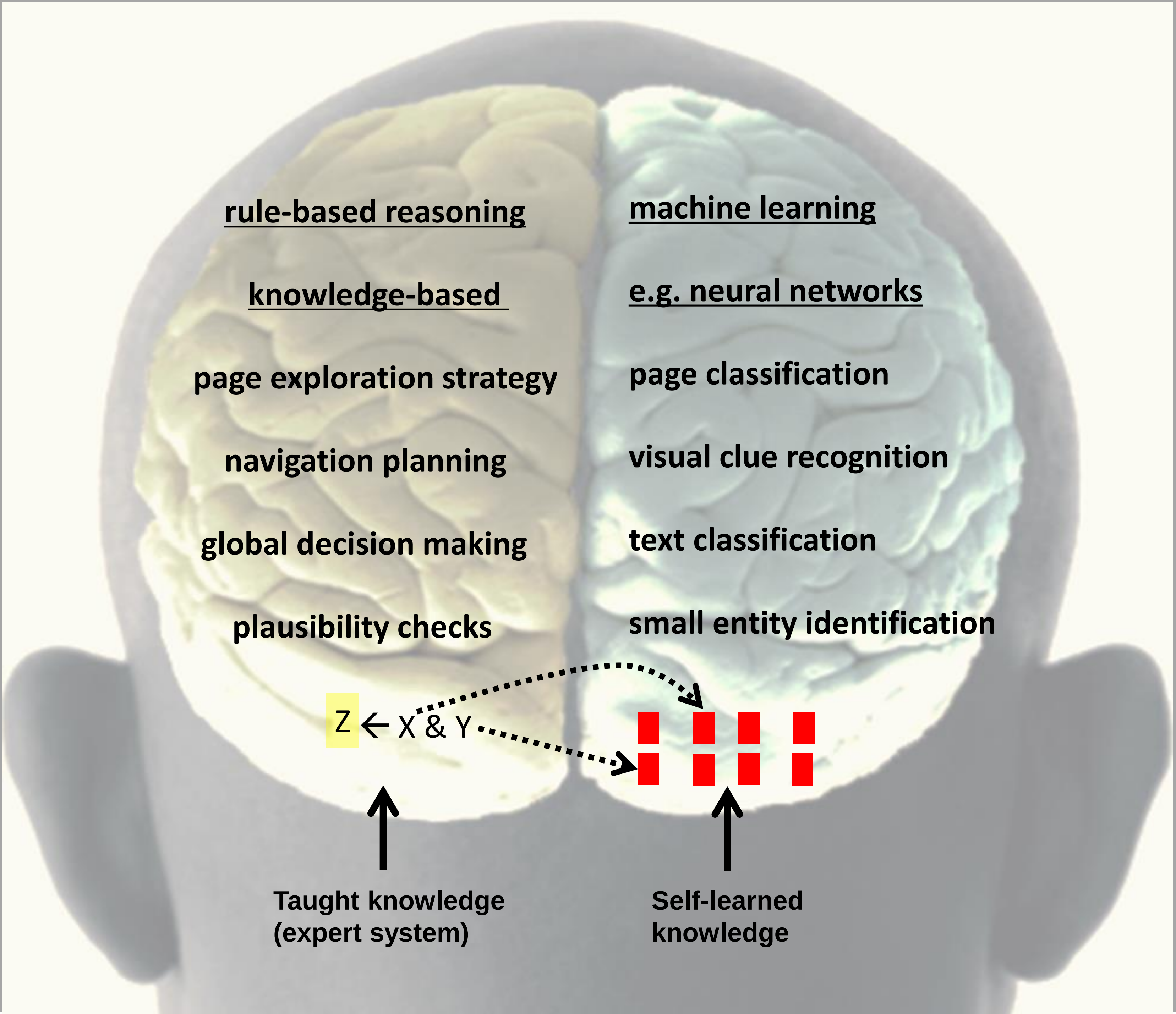
# DIADEM

## a €2.5M project started in April '10



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 Datalog 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:**

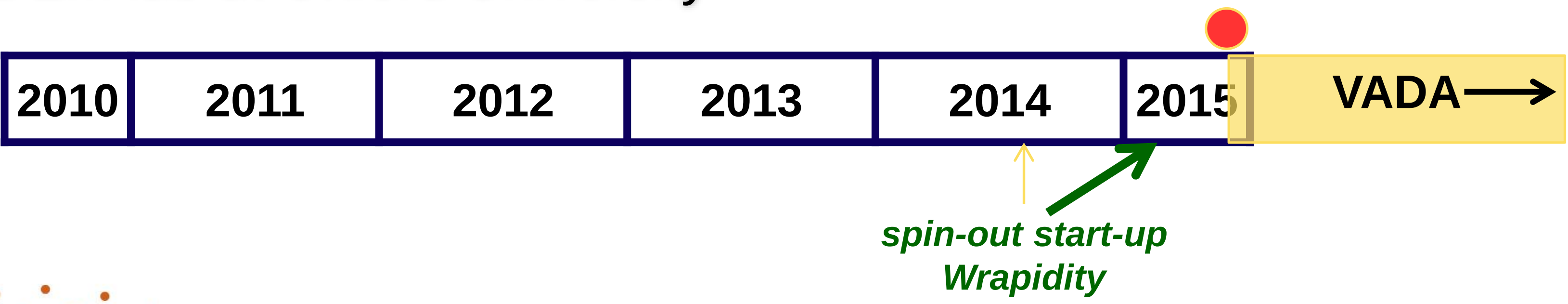
  - before filling a field try to leave it empty

  - rules for handling next-page links

- **Other types of rules**

# DIADEM PROJECT

- DIADEM lab at Oxford University



erc

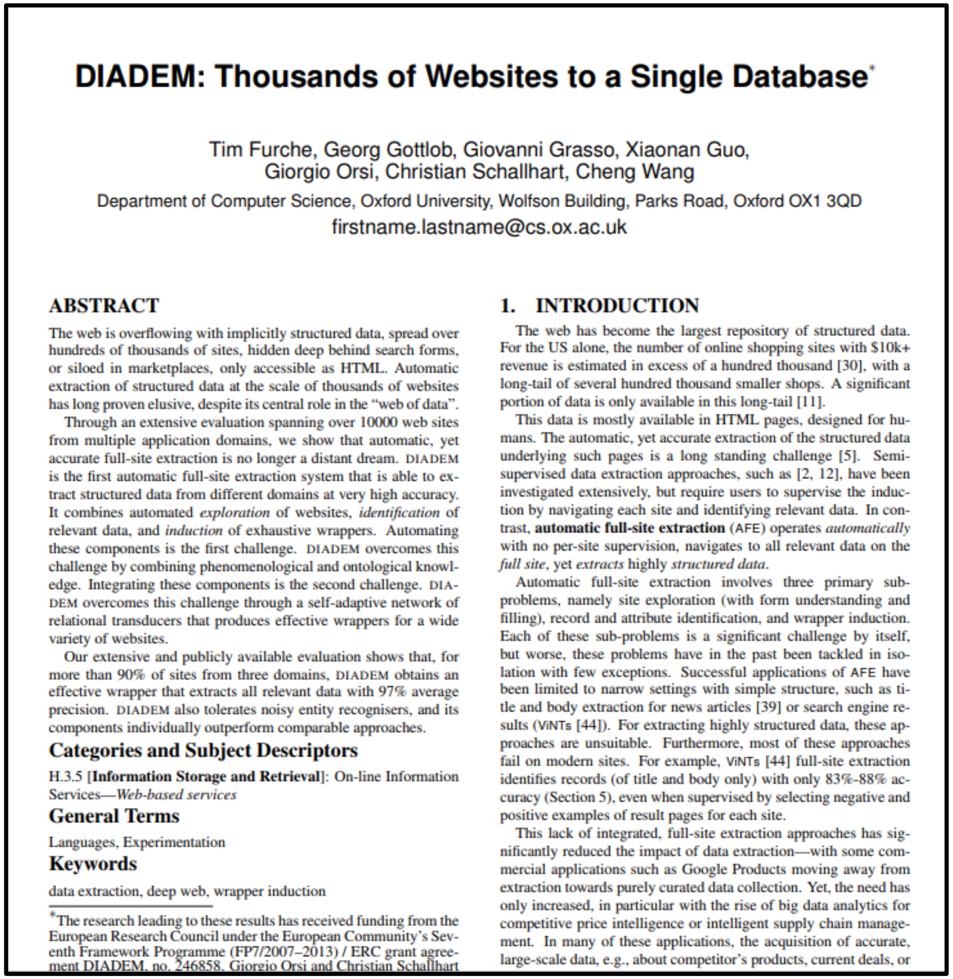
EPSRC

Google
YAHOO!
amazon.com

OXFORD
MARTIN
SCHOOL

Funding so far > $7M

# Research Papers (examples)



## DIADEM: Thousands of Websites to a Single Database*

Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo,
Giorgio Orsi, Christian Schallhart, Cheng Wang
Department of Computer Science, Oxford University, Wolfson Building, Parks Road, Oxford OX1 3QD
firstname.lastname@cs.ox.ac.uk

### ABSTRACT

The web is overflowing with implicitly structured data, spread over hundreds of thousands of sites, hidden deep behind search forms, or siloed in marketplaces, only accessible as HTML. Automatic extraction of structured data at the scale of thousands of websites has long proven elusive, despite its central role in the "web of data".

Through an extensive evaluation spanning over 10000 web sites from multiple application domains, we show that automatic, yet accurate full-site extraction is no longer a distant dream. DIADEM is the first automatic full-site extraction system that is able to extract structured data from different domains at very high accuracy. It combines automated *exploration* of websites, *identification* of relevant data, and *induction* of exhaustive wrappers. Automating these components is the first challenge. DIADEM overcomes this challenge by combining phenomenological and ontological knowledge. Integrating these components is the second challenge. DIADEM overcomes this challenge through a self-adaptive network of relational transducers that produces effective wrappers for a wide variety of websites.

Our extensive and publicly available evaluation shows that, for more than 90% of sites from three domains, DIADEM obtains an effective wrapper that extracts all relevant data with 97% average precision. DIADEM also tolerates noisy entity recognisers, and its components individually outperform comparable approaches.

### Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-based services*

### General Terms

Languages, Experimentation

### Keywords

data extraction, deep web, wrapper induction

### 1. INTRODUCTION

The web has become the largest repository of structured data. For the US alone, the number of online shopping sites with $10k+ revenue is estimated in excess of a hundred thousand [30], with a long-tail of several hundred thousand smaller shops. A significant portion of data is only available in this long-tail [11].

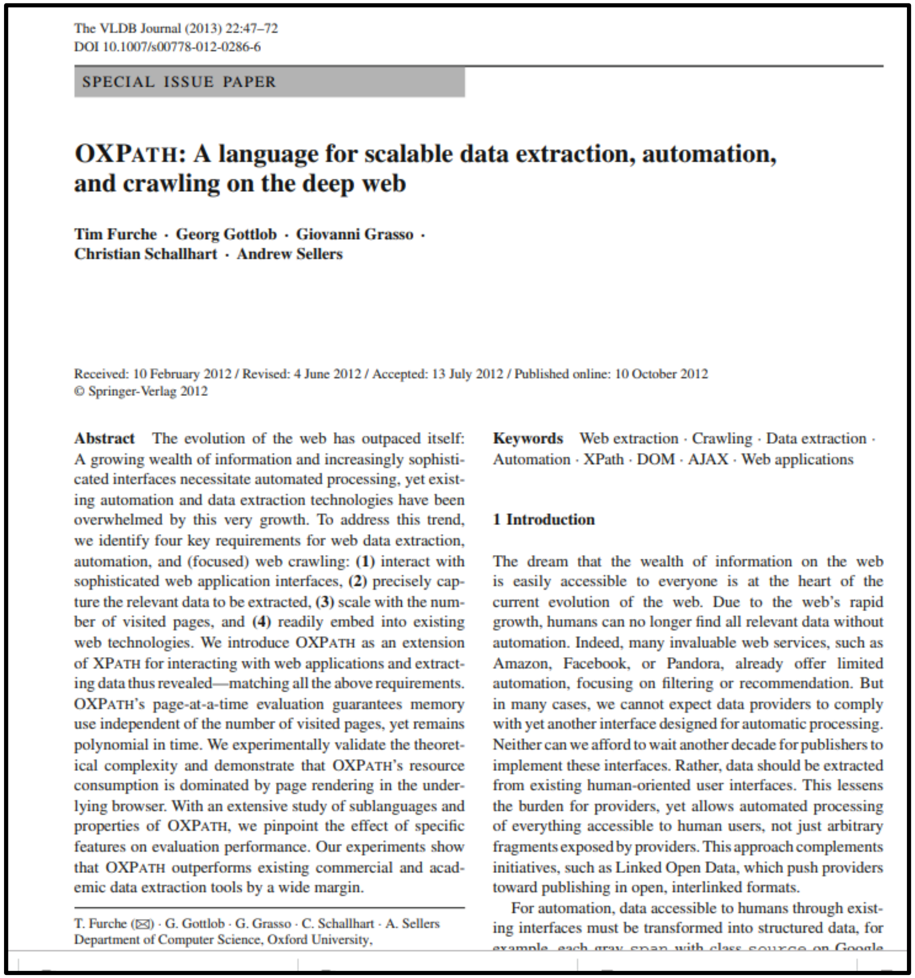
This data is mostly available in HTML pages, designed for humans. The automatic, yet accurate extraction of the structured data underlying such pages is a long standing challenge [5]. Semi-supervised data extraction approaches, such as [2, 12], have been investigated extensively, but require users to supervise the induction by navigating each site and identifying relevant data. In contrast, **automatic full-site extraction** (AFE) operates *automatically* with no per-site supervision, navigates to all relevant data on the *full site*, yet *extracts* highly *structured data*.

Automatic full-site extraction involves three primary subproblems, namely site exploration (with form understanding and filling), record and attribute identification, and wrapper induction. Each of these sub-problems is a significant challenge by itself, but worse, these problems have in the past been tackled in isolation with few exceptions. Successful applications of AFE have been limited to narrow settings with simple structure, such as title and body extraction for news articles [39] or search engine results (ViNTs [44]). For extracting highly structured data, these approaches are unsuitable. Furthermore, most of these approaches fail on modern sites. For example, ViNTs [44] full-site extraction identifies records (of title and body only) with only 83%-88% accuracy (Section 5), even when supervised by selecting negative and positive examples of result pages for each site.

This lack of integrated, full-site extraction approaches has significantly reduced the impact of data extraction—with some commercial applications such as Google Products moving away from extraction towards purely curated data collection. Yet, the need has only increased, in particular with the rise of big data analytics for competitive price intelligence or intelligent supply chain management. In many of these applications, the acquisition of accurate, large-scale data, e.g., about competitor's products, current deals, or

*The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement DIADEM, no. 246858. Giorgio Orsi and Christian Schallhart

VLDB 2014



The VLDB Journal (2013) 22:47–72
DOI 10.1007/s00778-012-0286-6

SPECIAL ISSUE PAPER

## OXPATH: A language for scalable data extraction, automation, and crawling on the deep web

**Tim Furche · Georg Gottlob · Giovanni Grasso ·
Christian Schallhart · Andrew Sellers**

Received: 10 February 2012 / Revised: 4 June 2012 / Accepted: 13 July 2012 / Published online: 10 October 2012

**Abstract** The evolution of the web has outpaced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth. To address this trend, we identify four key requirements for web data extraction, automation, and (focused) web crawling: **(1)** interact with sophisticated web application interfaces, **(2)** precisely capture the relevant data to be extracted, **(3)** scale with the number of visited pages, and **(4)** readily embed into existing web technologies. We introduce OXPATH as an extension of XPATH for interacting with web applications and extracting data thus revealed—matching all the above requirements. OXPATH's page-at-a-time evaluation guarantees memory use independent of the number of visited pages, yet remains polynomial in time. We experimentally validate the theoretical complexity and demonstrate that OXPATH's resource consumption is dominated by page rendering in the underlying browser. With an extensive study of sublanguages and properties of OXPATH, we pinpoint the effect of specific features on evaluation performance. Our experiments show that OXPATH outperforms existing commercial and academic data extraction tools by a wide margin.

T. Furche (✉) · G. Gottlob · G. Grasso · C. Schallhart · A. Sellers
Department of Computer Science, Oxford University,

**Keywords** Web extraction · Crawling · Data extraction · Automation · XPath · DOM · AJAX · Web applications

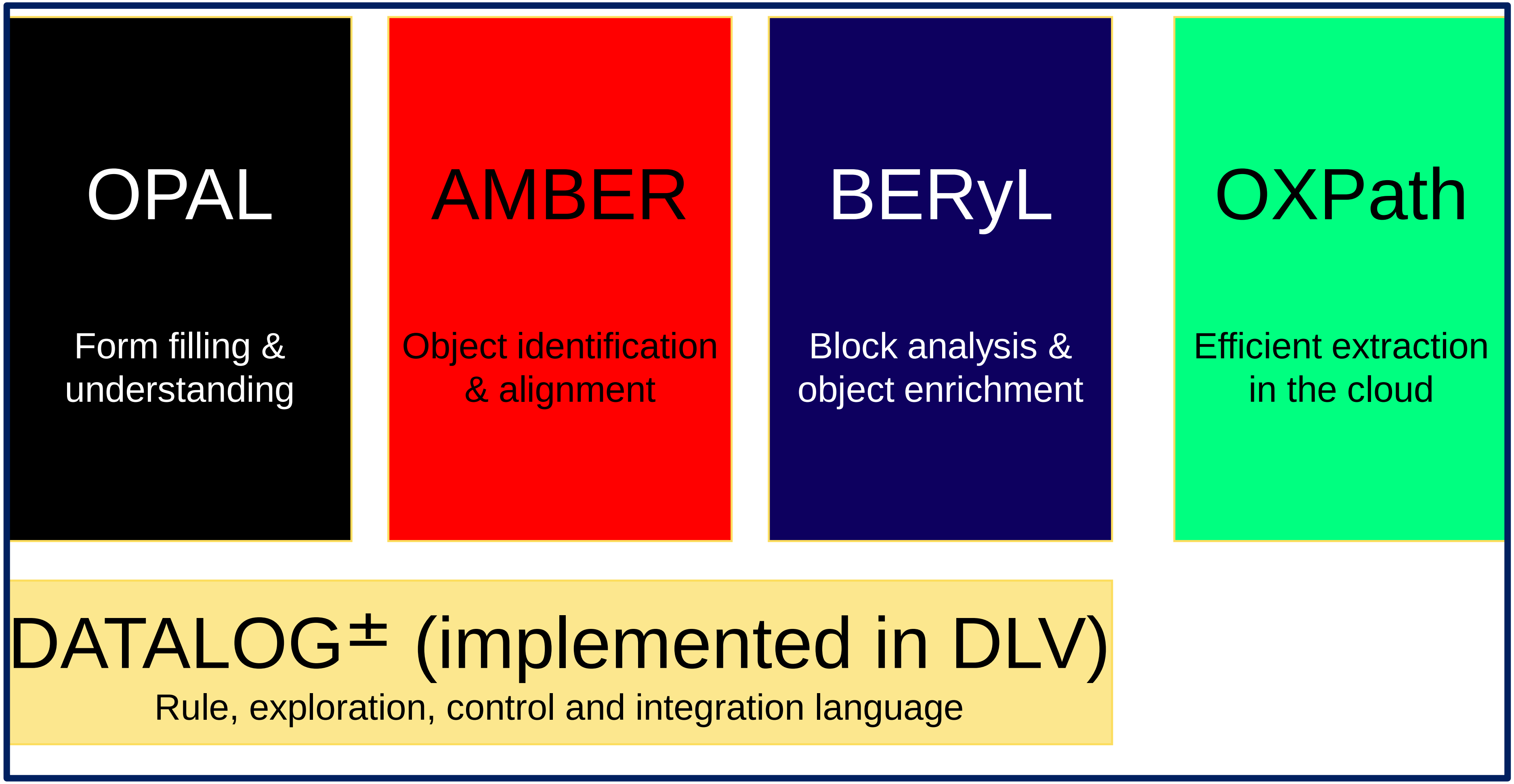
### 1 Introduction

The dream that the wealth of information on the web is easily accessible to everyone is at the heart of the current evolution of the web. Due to the web's rapid growth, humans can no longer find all relevant data without automation. Indeed, many invaluable web services, such as Amazon, Facebook, or Pandora, already offer limited automation, focusing on filtering or recommendation. But in many cases, we cannot expect data providers to comply with yet another interface designed for automatic processing. Neither can we afford to wait another decade for publishers to implement these interfaces. Rather, data should be extracted from existing human-oriented user interfaces. This lessens the burden for providers, yet allows automated processing of everything accessible to human users, not just arbitrary fragments exposed by providers. This approach complements initiatives, such as Linked Open Data, which push providers toward publishing in open, interlinked formats.

For automation, data accessible to humans through existing interfaces must be transformed into structured data, for example, each gray span with class source on Google

VLDB Journal 2013

# DIADEM Architecture



OPAL

Form filling & understanding

AMBER

Object identification & alignment

BERyL

Block analysis & object enrichment

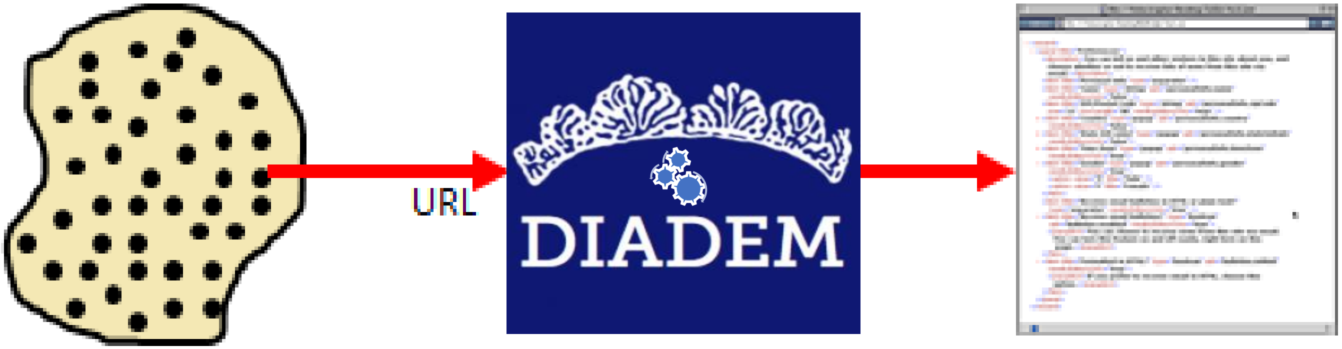
OXPath

Efficient extraction in the cloud

DATALOG$^{\pm}$ (implemented in DLV)

Rule, exploration, control and integration language

New knowledge-based technology combining formalised knowledge with machine learning



**Application domain with thousands of websites**

**Application-relevant highly structured  data**

Properties

Used cars

Forms with ranges & dependencies

Combines ML with rule-based Reasoning.

In production: Extracts from  200K websites in various application domains.

Successful on > 95%  of test sites.

# Evaluation on 10k+ Sites

**10,493**

**Sites** from real-estate and used-car

**92%**

**Effective wrappers** for more than 92% of sites on average

**100**

**Domain-dependent** concepts and relations

**45**

**Node** Amazon EC2 cluster running 2.1 days

**97%**

**Precision** of extracted primary attributes

**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
- Companies

# Commercial Impact

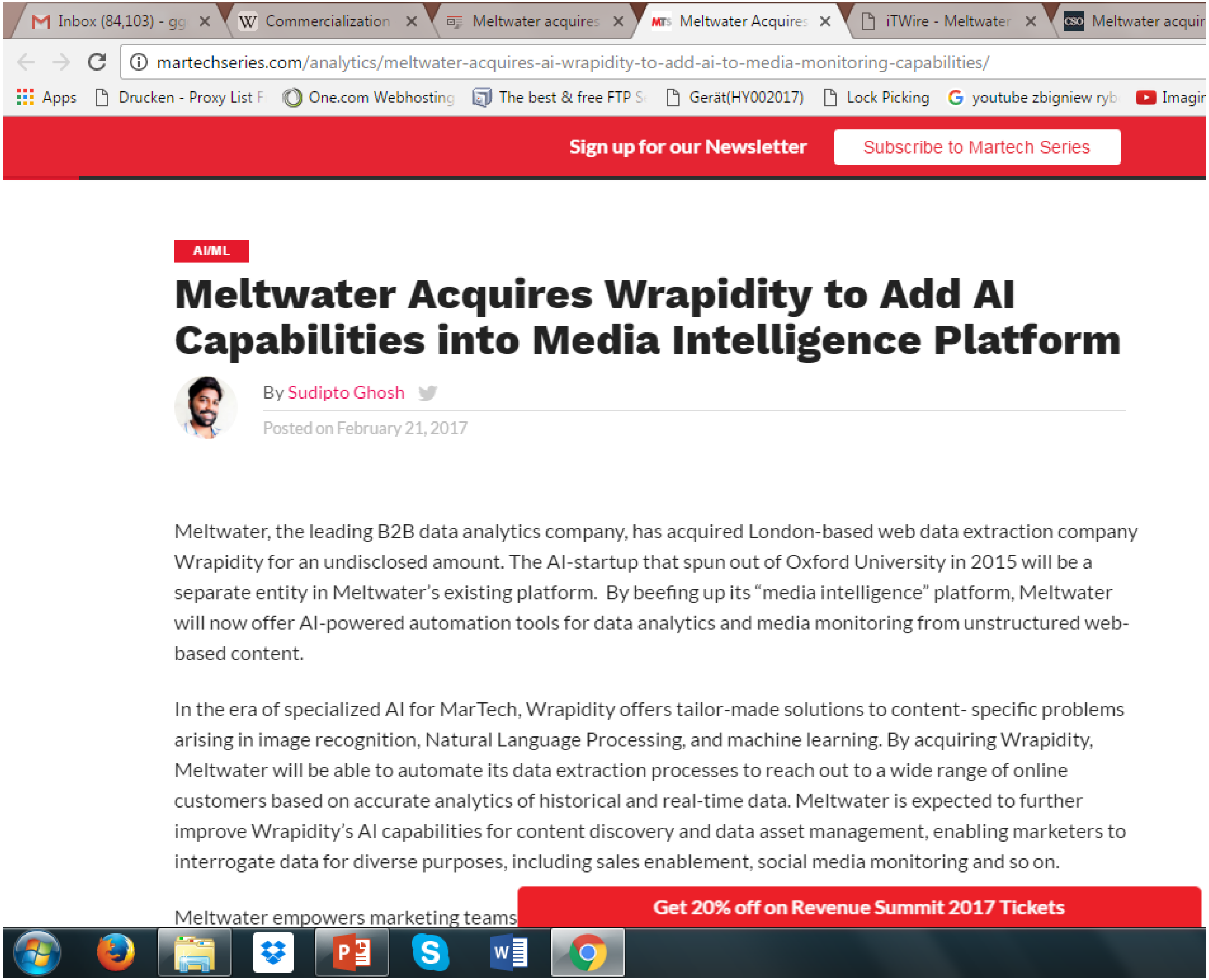
DIADEM | domain-centric intelligent automated data extraction methodology → Wrapidity

ERC Advanced Grant DIADEM
+ ERC Proof of Concept Grant
EXTRALYTICS

Founded February 2015
operating initially in Oxford
now in London

2 possibilities:

- Build up company with large client portfolio
- Sell technology, IP & software to strategic partner



AI/ML

# 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.

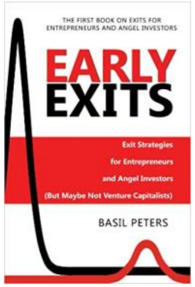
In the era of specialized AI for MarTech, Wrapidity offers tailor-made solutions to content- specific problems arising in image recognition, Natural Language Processing, and machine learning. By acquiring Wrapidity, Meltwater will be able to automate its data extraction processes to reach out to a wide 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.

Meltwater empowers marketing teams

# Remarks to Prospective Academic Founders

**Question:** Big exit (8+ years) vs moderate exit (2-3 years) ?

Big exit: *expected value* lower: Risks (see Anki) + VCs
There are exceptions such as DeepMind.
If you aim at a big exit, leave academia.



THE FIRST BOOK ON EXITS FOR ENTREPRENEURS AND ANGEL INVESTORS

EARLY EXITS

Exit Strategies for Entrepreneurs and Angel Investors (But Maybe Not Venture Capitalists)

BASIL PETERS

**Management:** There are 1001 things to do, so take a full-time CEO.

**Customers:** Selling your new technology provides satisfaction.
But: Many customers → many problems.
A single failure can lead to threats and endless discussions.
Automotive suppliers are particularly harsh, being pressured by OEMs.

**Engineers vs Salespeople**: Salespeople promise what Engineers cannot deliver.
You are in the middle…

**Lead promises are broken**. 50% of the times “we will buy / want to buy your product”
is broken. Main reasons:

(i) not a ``decider” (AT Biz jargon: “nicht satisfaktionsfähig”)
(ii) sudden change of management (happens all the time)

**VCs** : - They can afford to the risk that only 1/10 of their companies succeed.
- They do not want moderate exits and can block them.
- They can force you to sell in some cases.
- Good VCs help you.

**Business Angels**: Are normally the better solution when aiming at moderate exits.

**Universities**: They are usually the IP holders and want to get equity.
Oxford: 20-30%, even though
Austria??

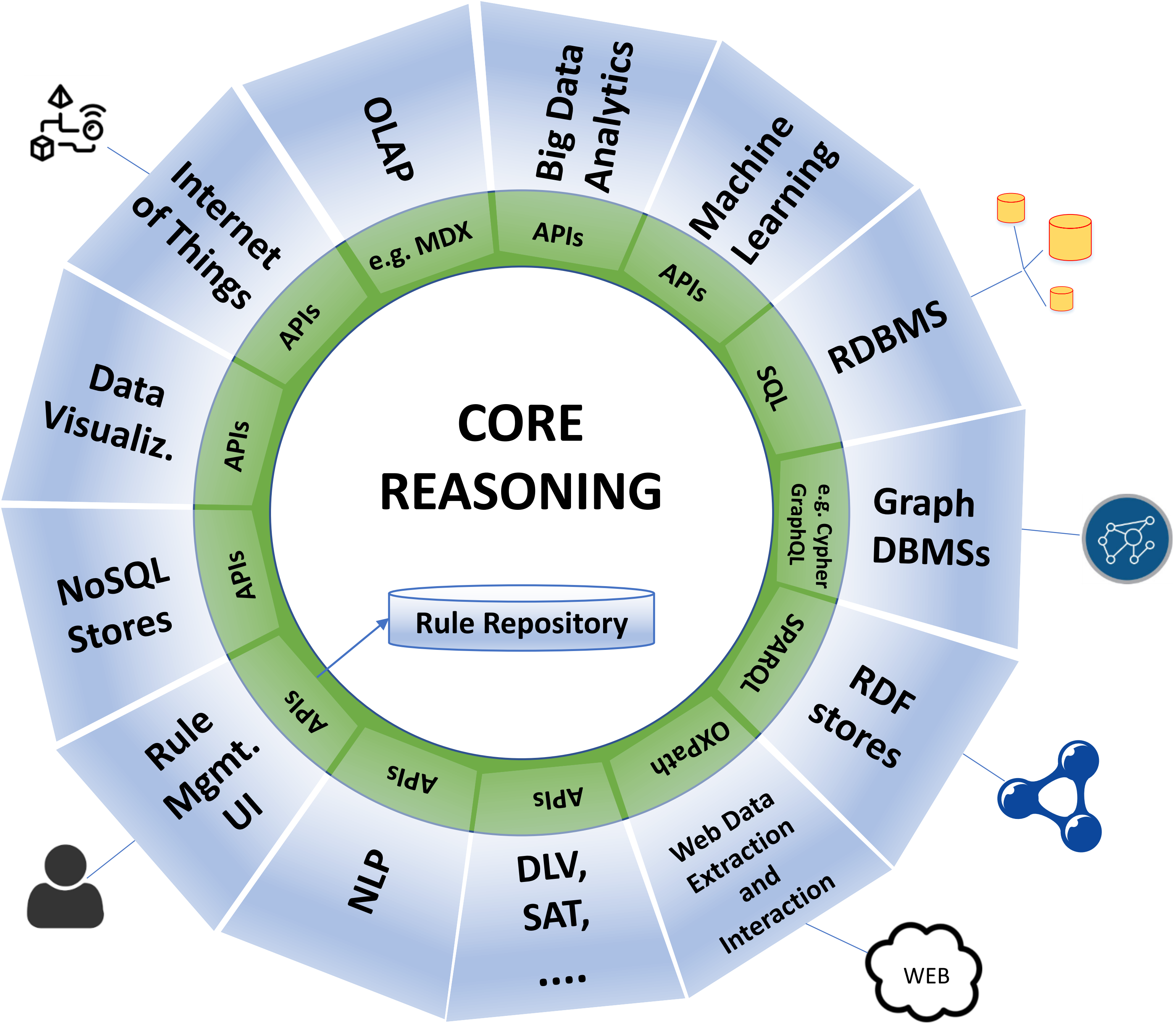
**Advantages**: Many very nice research problems arise from customer needs.

Diverse activities let you “live longer”.

You prove that your research is useful. – better credibility, also for publications.

Additional income.

Selling shares is tax efficient. In AT: 27.5% CGT vs 50% income tax.



CORE REASONING

Rule Repository

OLAP

e.g. MDX

Big Data Analytics

APIs

Machine Learning

APIs

RDBMS

SQL

Graph DBMSs

e.g. Cypher GraphQL

RDF stores

SPARQL

Web Data Extraction and Interaction

OXPath

WEB

DLV, SAT, ....

APIs

NLP

APIs

Rule Mgmt. UI

APIs

NoSQL Stores

APIs

Data Visualiz.

APIs

Internet of Things

APIs

# Enterprise Knowledge Graphs

**Facebook Knowledge Graph:** Social graph with people, places and things + information from Wikipedia

**Amazon Knowledge Graph:** Started as product categorization ontology

**Wolfram KB:** World facts + mathematics

**Factual:** Businesses & places

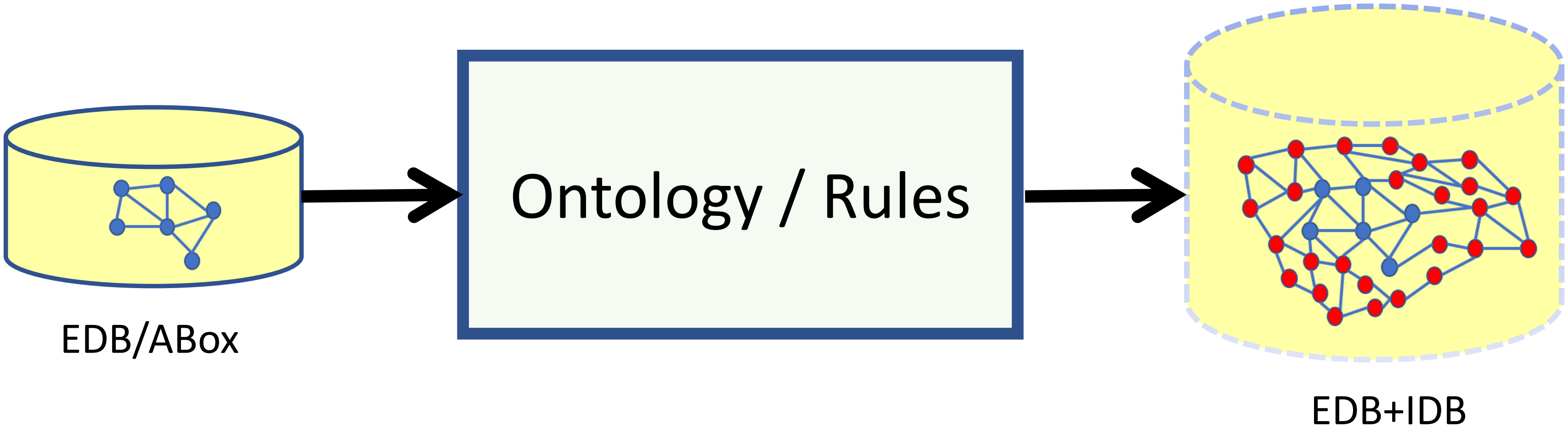
**Recruit Institute of Technology** (R.I.T): People, skills, recruiting

**Central Banks:** Company register – ownership graph

**Credit Rating Agencies …**

Thousands of medium to large size companies now want their own corporate knowledge graph. This not just for semantic indexing and search, but for advanced reasoning tasks on top of machine learning.

# Reasoning in Knowledge Graphs



EDB/ABox

Ontology / Rules

EDB+IDB

Many still think that DLs or graph databases suffice. However:

Reasoning tasks are required that cannot be expressed by description logics, and cannot be reasonably managed by relational DBMS, nor by graph DBMS.

# Example: Wikidata Marriage Intervals

[Krötzsch DL 2017]

Wikidata contains the statement :

**Taylor was married to Burton starting from 1964 and ending 1974**

This can be represented in relational DB or Datalog-notation by :

```
married(taylor,burton,1964,1974)
```

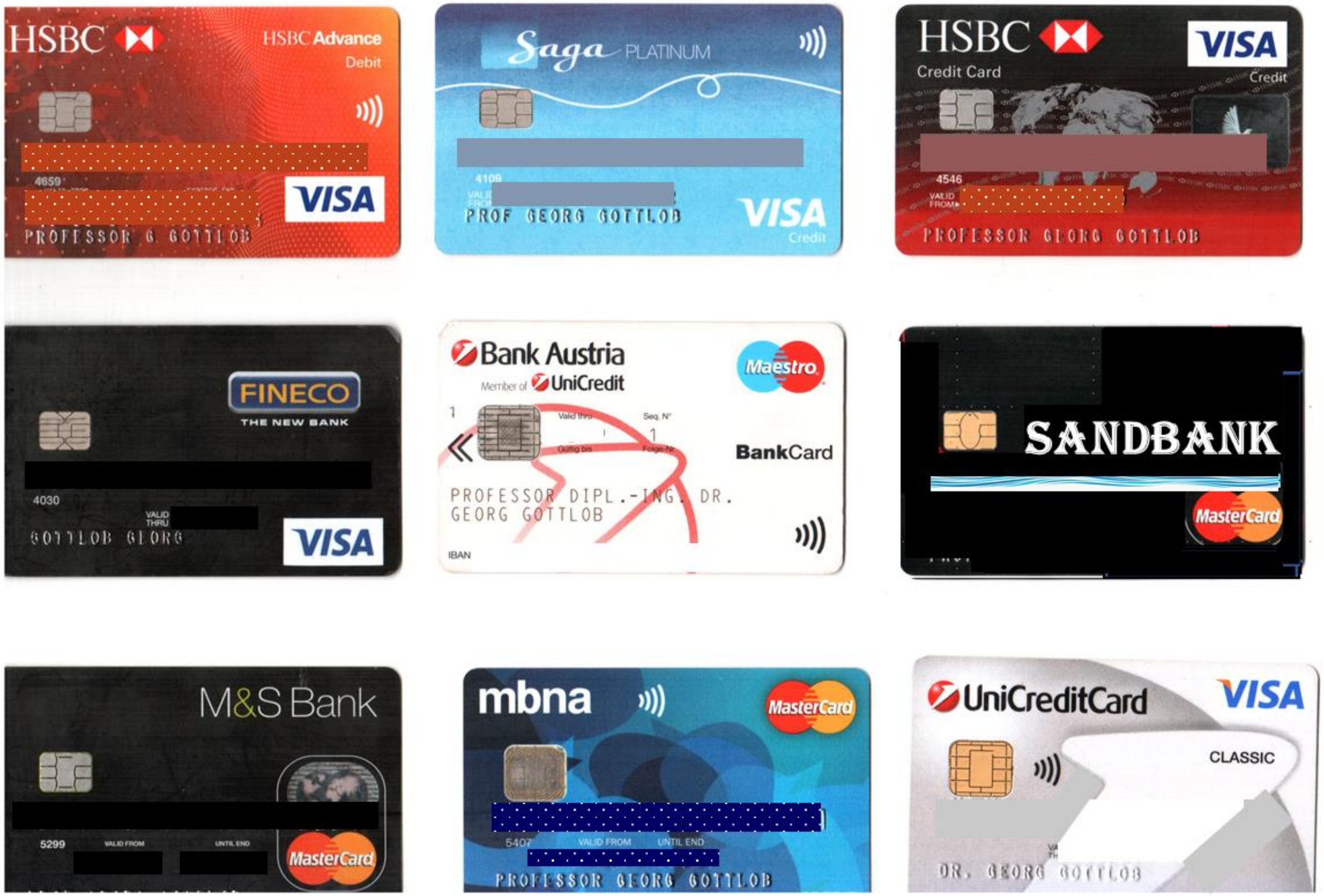
Symmetry rule for marriage intervals in Datalog:

```
∀ u,v,x,y. married(u,v,x,y) → married(v,u,x,y)
```

**This cannot be expressed in DLs!**

Note: In what follows, we will often omit universal quantifiers.

# Example: My Creditworthiness



# Example: My Creditworthiness

**up to £10,000**

**£8,500**

**£12,000**

**up to EUR 10,000**

**up to EUR 20,000**

**£500**

**£ 8,000**

**£ 12,500**

**EUR 14,000**

# My Explanation

A machine-learning program has "reasonably" learned:

> ***People who live in a joint household with someone who does not pay their bills are likely to fail repaying their own debts.***

This ethically questionable rule was applied to
incomplete and wrong data:

- Before I bought the house there was a tenant who indeed did not pay his bills (*tons of unpaid bills & overdue notices in my mailbox*).
- The tenant had moved out before I moved in, but the Credit Rating Agency did not know, and simply assumed he still lived there.

ML should be complemented and, where necessary, overruled by domain-specific “expert rules” that express domain knowledge:

- **A new house-owner is most likely unrelated to a previous tenant.**
- **If a house is bought by somebody who did not live there previously, and now lives there, then the previous occupiers have most likely moved out. [→Verify!]**
- **If someone has closed their bank account without opening a new one then it is likely that the person has moved out of the country.**

...

Automatically accessing outside sources such as the Land Register and/or Social Networks may help.

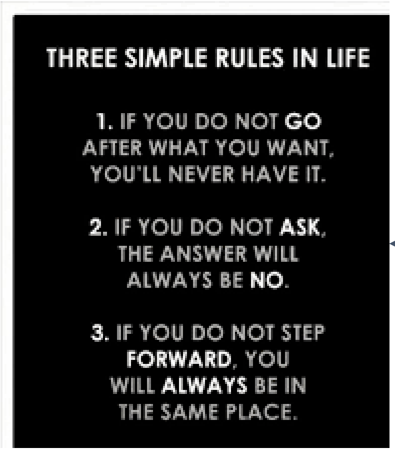
→ Data extraction from external sources is a requirement for KGMS.

# Knowledge Graph Management Systems

KGMS combine the power of rule-based reasoning with machine learning over Big Data:

**KGMS = KBMS + Big Data + Analytics**

**Misusing the lateralization thesis for illustration**

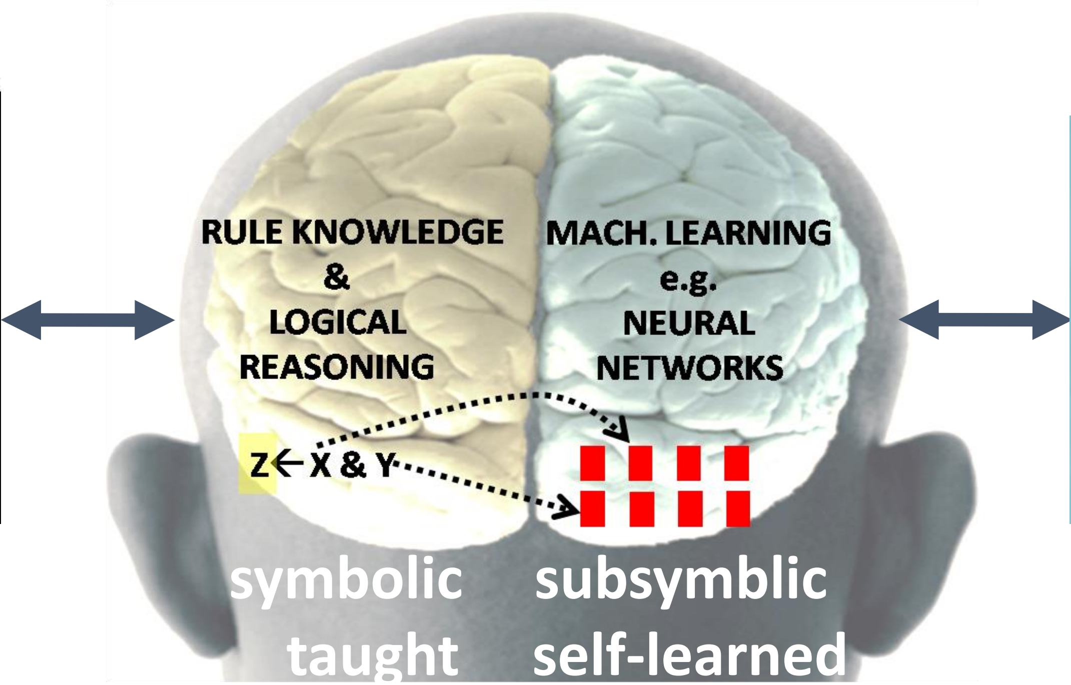


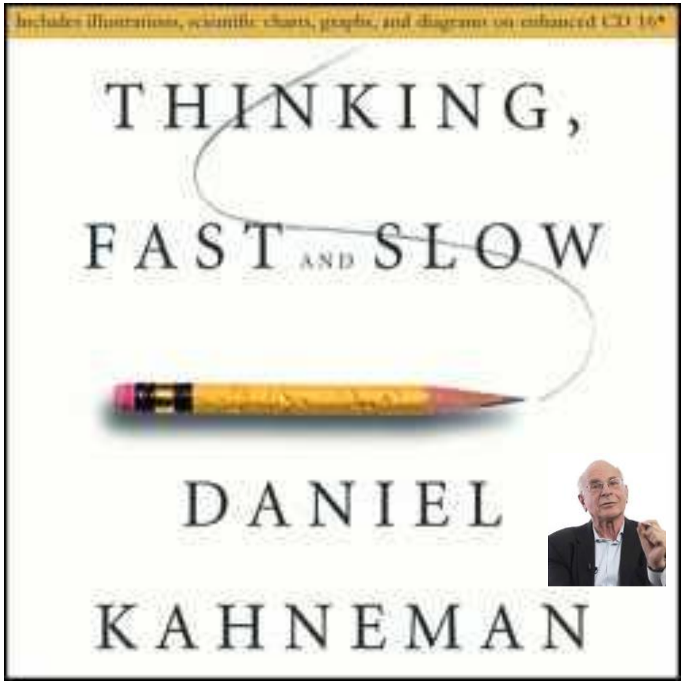
THREE SIMPLE RULES IN LIFE

1. IF YOU DO NOT GO AFTER WHAT YOU WANT, YOU'LL NEVER HAVE IT.

2. IF YOU DO NOT ASK, THE ANSWER WILL ALWAYS BE NO.

3. IF YOU DO NOT STEP FORWARD, YOU WILL ALWAYS BE IN THE SAME PLACE.



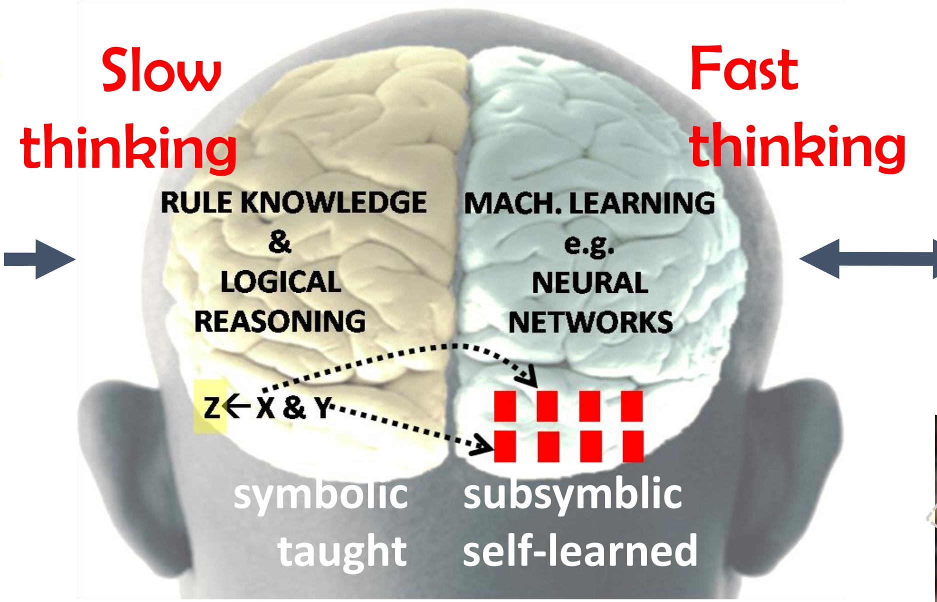


Includes illustrations, scientific charts, graphs, and diagrams on enhanced CD 16*

THINKING, FAST AND SLOW

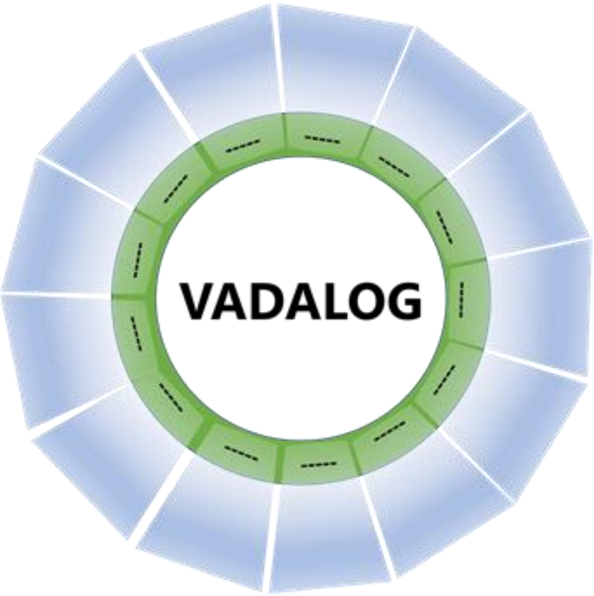
DANIEL KAHNEMAN

Grandma: “Fly agarics are poisonous mushrooms. If you eat a poisonous mushroom, you may die”.


Slow thinking
RULE KNOWLEDGE & LOGICAL REASONING
Z←X & Y
symbolic
taught
Fast thinking
MACH. LEARNING e.g. NEURAL NETWORKS
subsymblic
self-learned

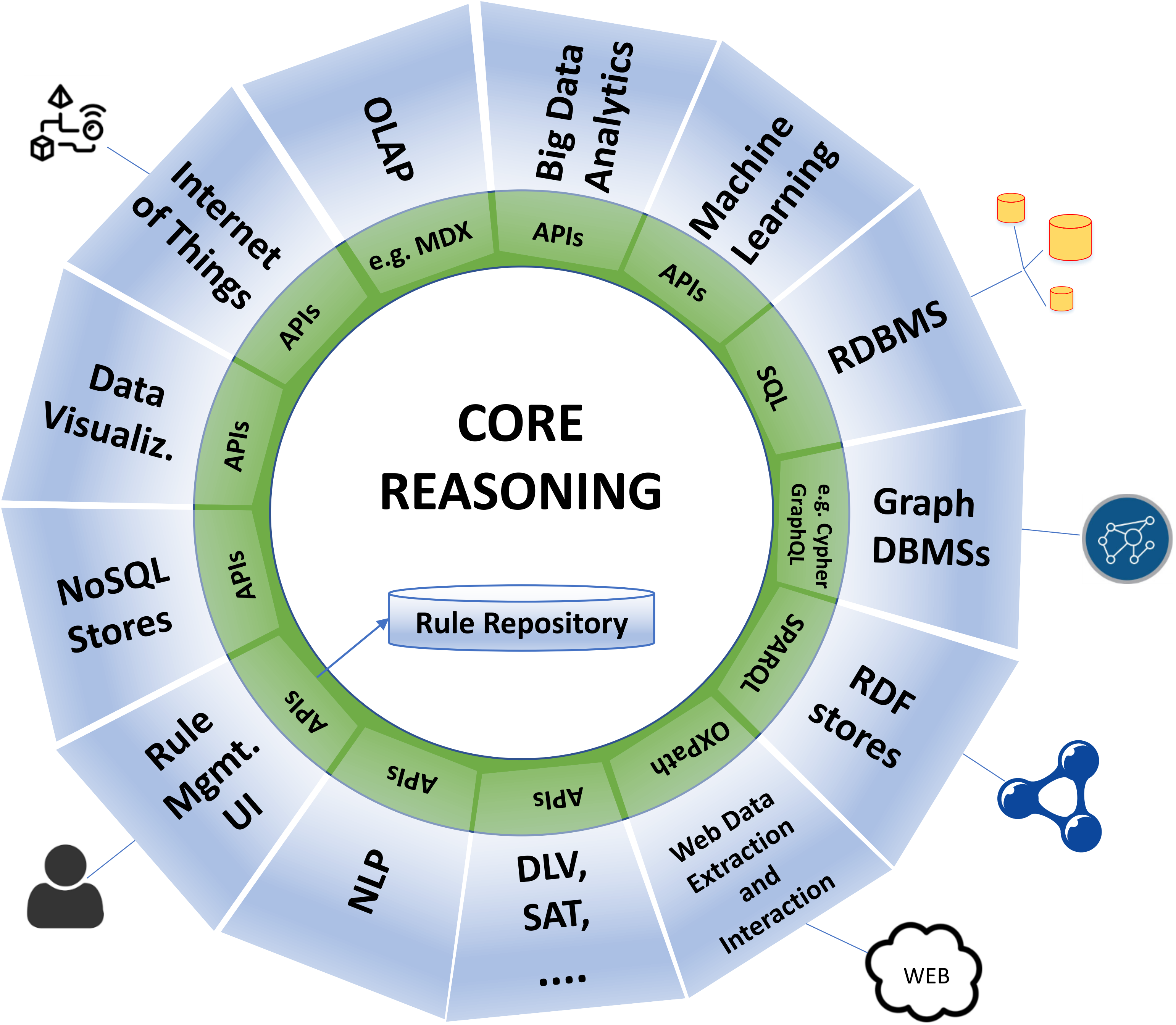
Yikes, a fly agaric!

# Vadalog KGMS Being Built at Oxford

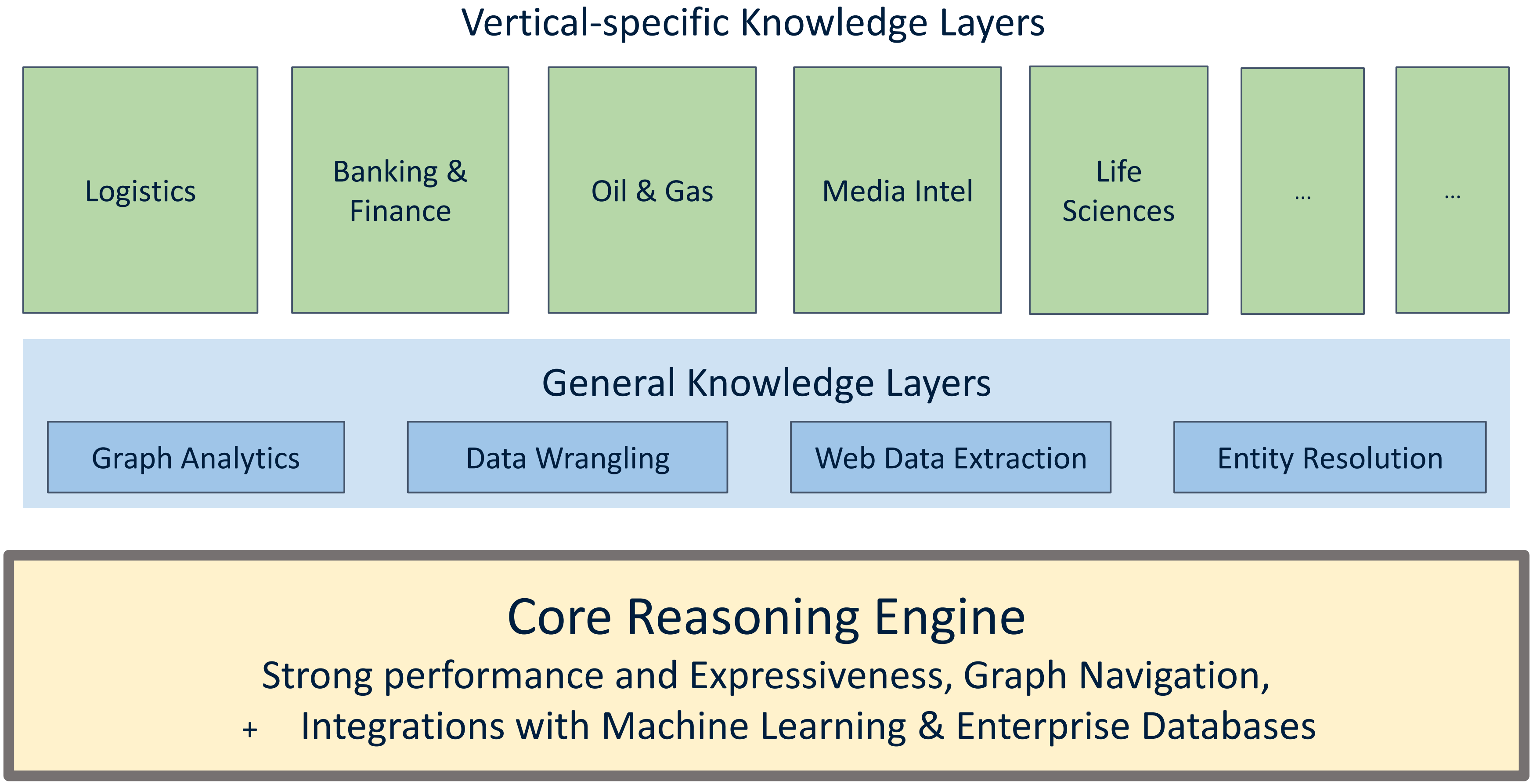


Current Team Members

- VADA = **V**alue-**A**dded **DA**ta
- General architecture
- The core reasoning language “Warded Datalog” and its extensions
- Connectivity: Some plug-ins


CORE
REASONING
Rule Repository
OLAP
e.g. MDX
Big Data Analytics
APIs
Machine Learning
APIs
RDBMS
SQL
Graph DBMSs
e.g. Cypher GraphQL
RDF stores
SPARQL
Web Data Extraction and Interaction
OXPath
WEB
DLV, SAT, ....
APIs
NLP
APIs
Rule Mgmt. UI
APIs
NoSQL Stores
APIs
Data Visualiz.
APIs
Internet of Things
APIs

# Knowledge Layers



Vertical-specific Knowledge Layers

Logistics

Banking & Finance

Oil & Gas

Media Intel

Life Sciences

…

…

General Knowledge Layers

Graph Analytics

Data Wrangling

Web Data Extraction

Entity Resolution

Core Reasoning Engine

Strong performance and Expressiveness, Graph Navigation,

+ Integrations with Machine Learning & Enterprise Databases

DeepReason.ai