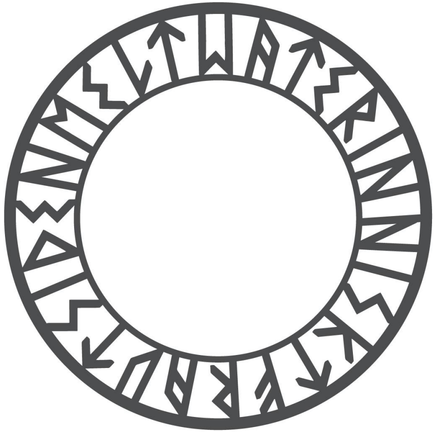
Wrapping Millions of Documents Per Day

and How that's Just the Beginning



Background: About me

OXFORD-MAN INSTITUTE

OXFORD MARTIN SCHOOL

Lecturer in Databases and Co-investigator VADA

Fellow of Oxford Martin School & OMI

Co-founder and CTO

Meltwater Outside Insight

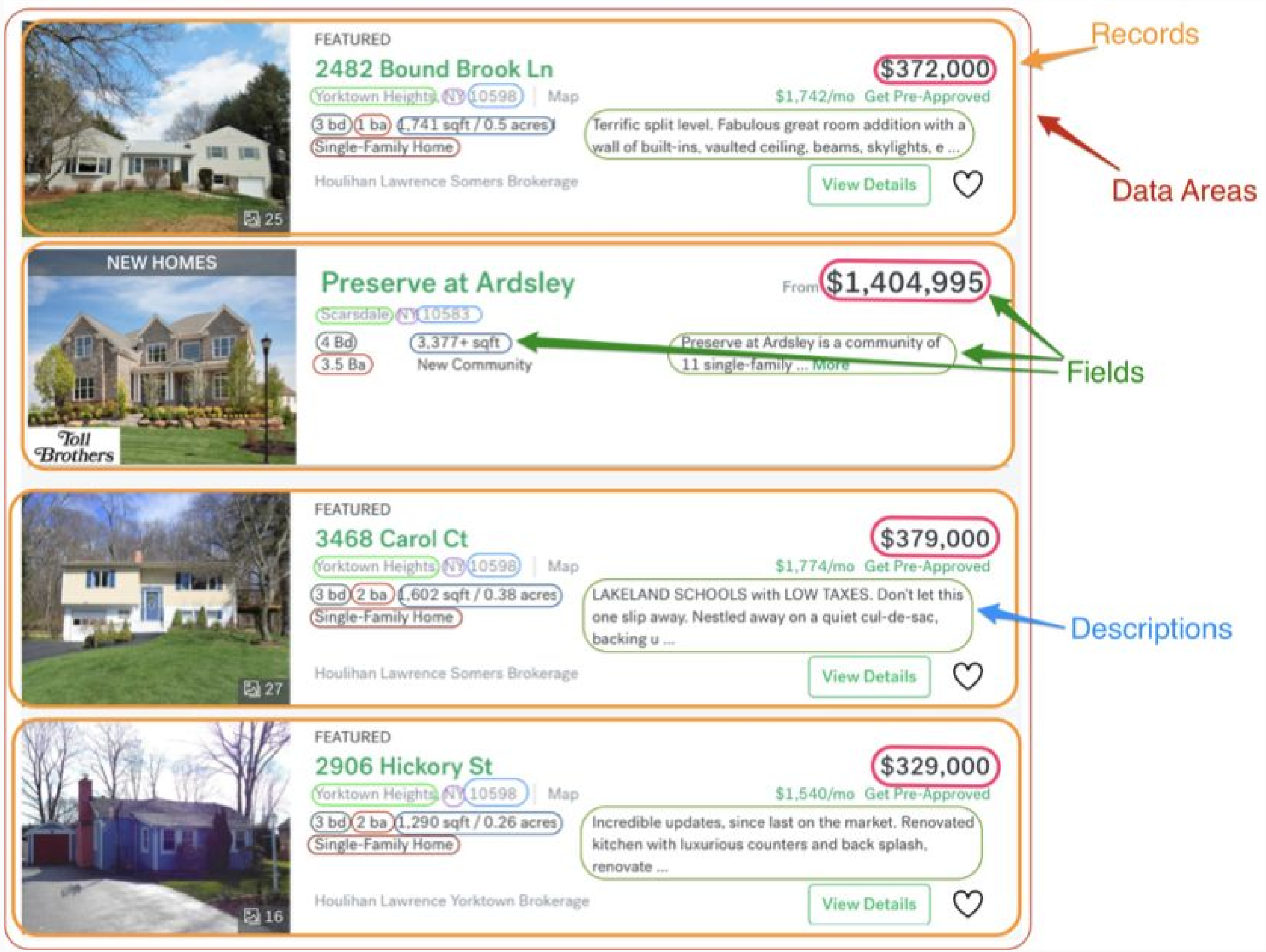
Senior Research Manager

Meltwater

Mission

SHACK15

Mission



Records

Data Areas

Fields

Descriptions

FEATURED

2482 Bound Brook Ln

$372,000

Yorktown Heights, NY 10598 | Map

$1,742/mo Get Pre-Approved

3 bd 1 ba 1,741 sqft / 0.5 acres

Single-Family Home

Terrific split level. Fabulous great room addition with a wall of built-ins, vaulted ceiling, beams, skylights, e ...

Houlihan Lawrence Somers Brokerage

View Details

NEW HOMES

Preserve at Ardsley

From $1,404,995

Scarsdale, NY 10583

4 Bd 3,377+ sqft

3.5 Ba New Community

Preserve at Ardsley is a community of 11 single-family ... More

FEATURED

3468 Carol Ct

$379,000

Yorktown Heights, NY 10598 | Map

$1,774/mo Get Pre-Approved

3 bd 2 ba 1,602 sqft / 0.38 acres

Single-Family Home

LAKELAND SCHOOLS with LOW TAXES. Don't let this one slip away. Nestled away on a quiet cul-de-sac, backing u ...

Houlihan Lawrence Somers Brokerage

View Details

FEATURED

2906 Hickory St

$329,000

Yorktown Heights, NY 10598 | Map

$1,540/mo Get Pre-Approved

3 bd 2 ba 1,290 sqft / 0.26 acres

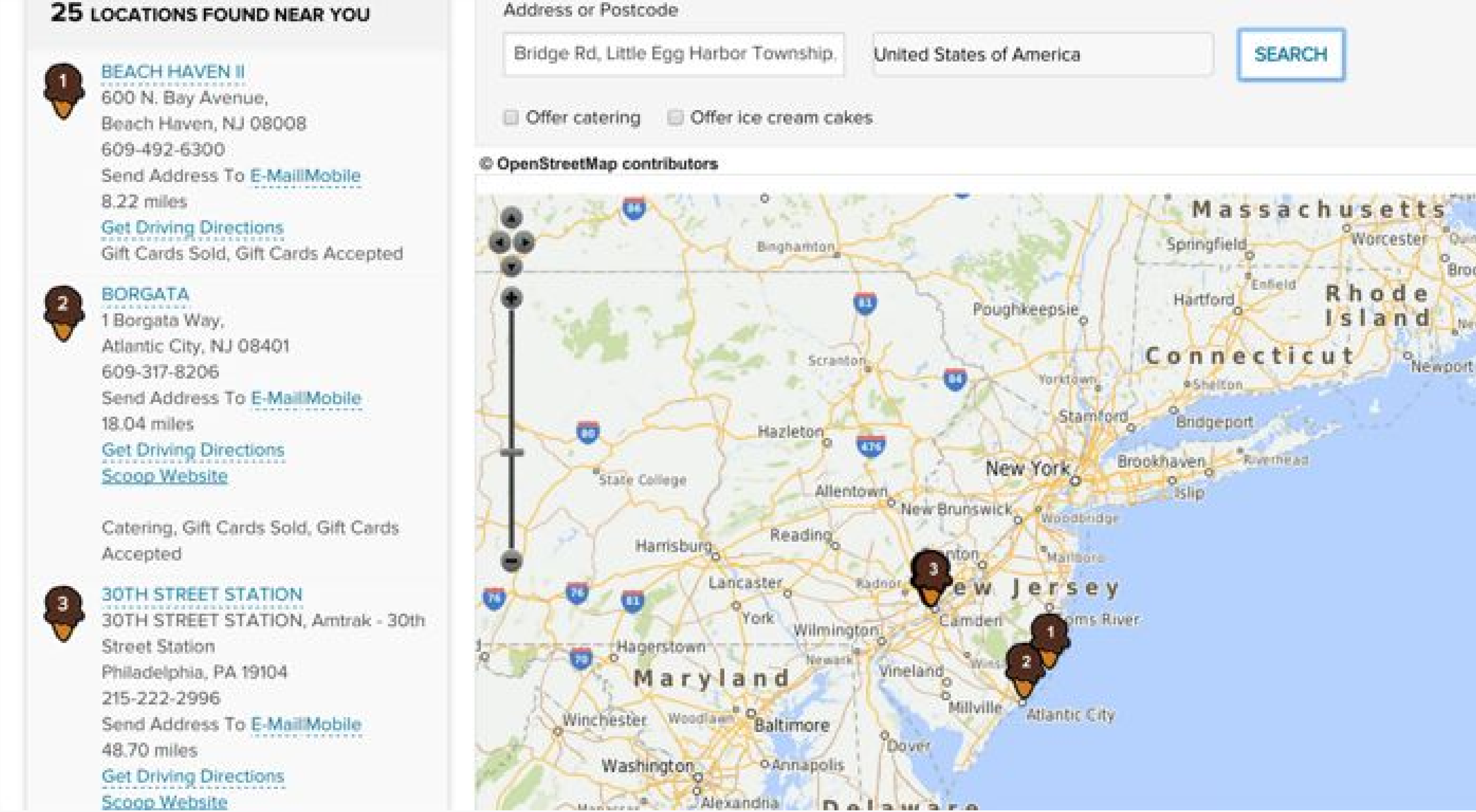
Single-Family Home

Incredible updates, since last on the market. Renovated kitchen with luxurious counters and back splash, renovate ...

Houlihan Lawrence Yorktown Brokerage

View Details

Mission



25 LOCATIONS FOUND NEAR YOU

1. BEACH HAVEN II
600 N. Bay Avenue,
Beach Haven, NJ 08008
609-492-6300
Send Address To E-MailMobile
8.22 miles
Get Driving Directions
Gift Cards Sold, Gift Cards Accepted

2. BORGATA
1 Borgata Way,
Atlantic City, NJ 08401
609-317-8206
Send Address To E-MailMobile
18.04 miles
Get Driving Directions
Scoop Website

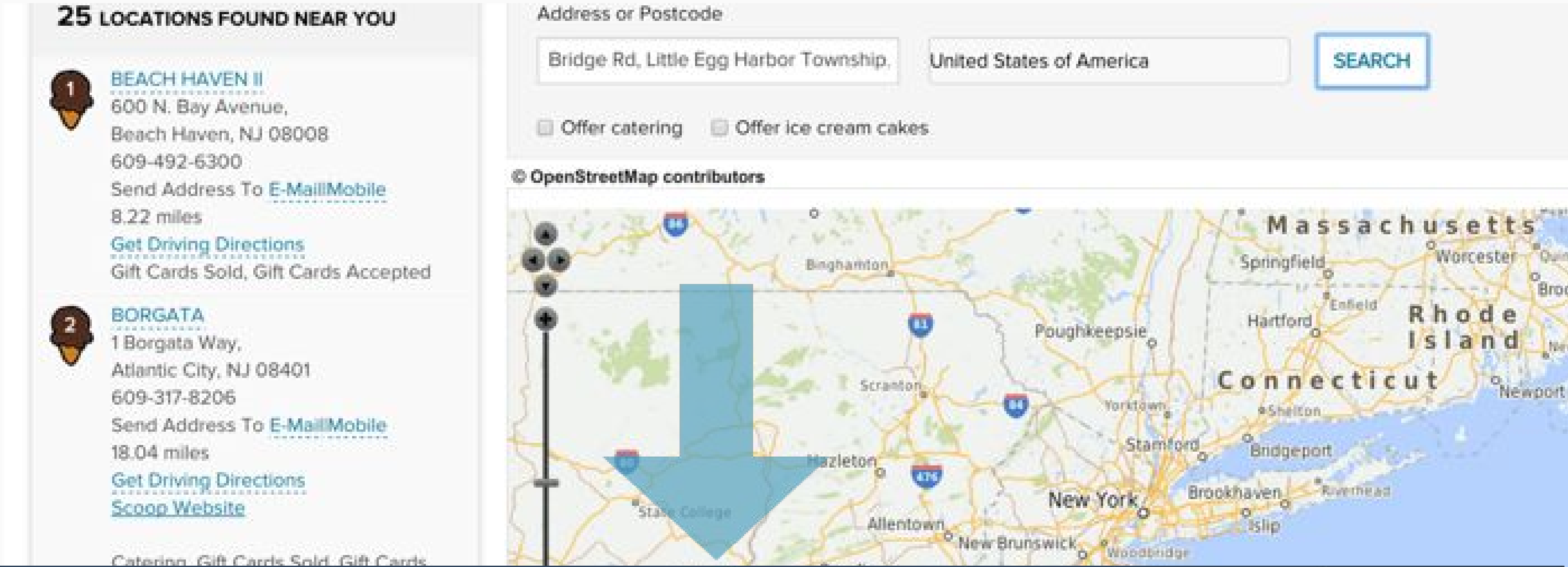
Catering, Gift Cards Sold, Gift Cards Accepted

3. 30TH STREET STATION
30TH STREET STATION, Amtrak - 30th Street Station
Philadelphia, PA 19104
215-222-2996
Send Address To E-MailMobile
48.70 miles
Get Driving Directions
Scoop Website

Address or Postcode: Bridge Rd, Little Egg Harbor Township. | United States of America | SEARCH

□ Offer catering □ Offer ice cream cakes

© OpenStreetMap contributors



25 LOCATIONS FOUND NEAR YOU

1 BEACH HAVEN II
600 N. Bay Avenue,
Beach Haven, NJ 08008
609-492-6300
Send Address To E-MailMobile
8.22 miles
Get Driving Directions
Gift Cards Sold, Gift Cards Accepted

2 BORGATA
1 Borgata Way,
Atlantic City, NJ 08401
609-317-8206
Send Address To E-MailMobile
18.04 miles
Get Driving Directions
Scoop Website

Catering, Gift Cards Sold, Gift Cards

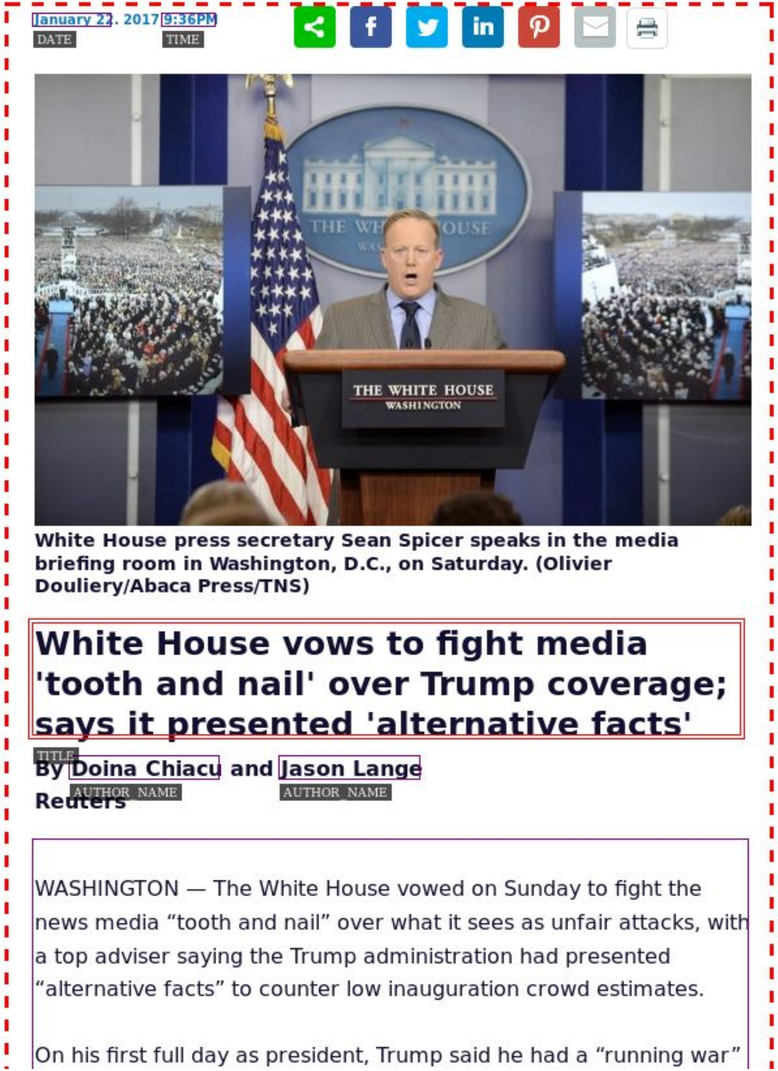
Address or Postcode

Bridge Rd, Little Egg Harbor Township, | United States of America | SEARCH

□ Offer catering □ Offer ice cream cakes

© OpenStreetMap contributors

NAME	STREET ADDRESS	LOCALITY	STATE	POSTCODE	PHONE
BEACH HAVEN II	600 N. Bay Avenue	Beach Haven	NJ	08008	609-492-6300
BORGATA	1 Borgata Way	Atlantic city	NJ	08401	609-317-8206
30TH STREET STATION	30TH STREET STATION	Philadelphia	PA	19104	215-222-2996



January 22, 2017 9:36PM

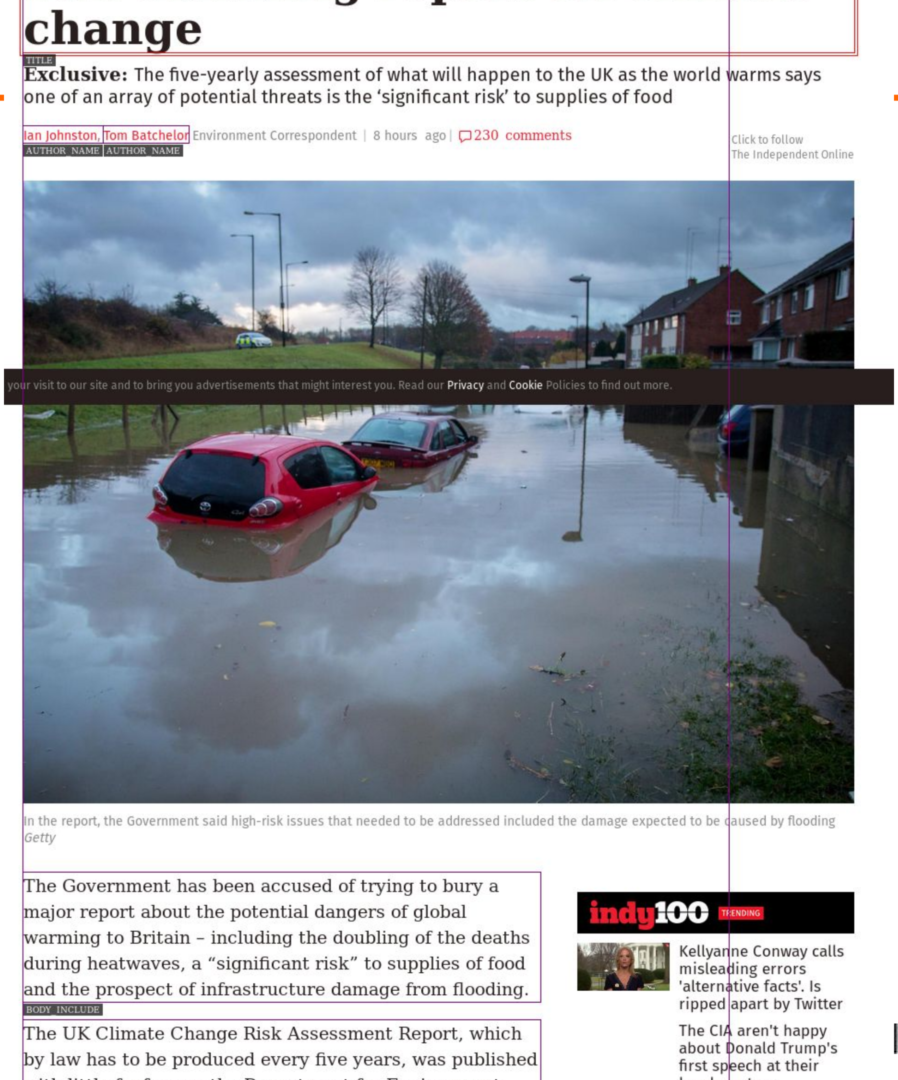
White House press secretary Sean Spicer speaks in the media briefing room in Washington, D.C., on Saturday. (Olivier Douliery/Abaca Press/TNS)

White House vows to fight media 'tooth and nail' over Trump coverage; says it presented 'alternative facts'

By Doina Chiacu and Jason Lange
Reuters

WASHINGTON — The White House vowed on Sunday to fight the news media "tooth and nail" over what it sees as unfair attacks, with a top adviser saying the Trump administration had presented "alternative facts" to counter low inauguration crowd estimates.

On his first full day as president, Trump said he had a "running war"



change

Exclusive: The five-yearly assessment of what will happen to the UK as the world warms says one of an array of potential threats is the 'significant risk' to supplies of food

Ian Johnston, Tom Batchelor Environment Correspondent | 8 hours ago | 230 comments

Click to follow
The Independent Online

your visit to our site and to bring you advertisements that might interest you. Read our Privacy and Cookie Policies to find out more.

In the report, the Government said high-risk issues that needed to be addressed included the damage expected to be caused by flooding *Getty*

The Government has been accused of trying to bury a major report about the potential dangers of global warming to Britain – including the doubling of the deaths during heatwaves, a "significant risk" to supplies of food and the prospect of infrastructure damage from flooding.

The UK Climate Change Risk Assessment Report, which by law has to be produced every five years, was published

Kellyanne Conway calls misleading errors 'alternative facts'. Is ripped apart by Twitter

The CIA aren't happy about Donald Trump's first speech at their

Demo

https://youtu.be/j_0IZdNJ-aw

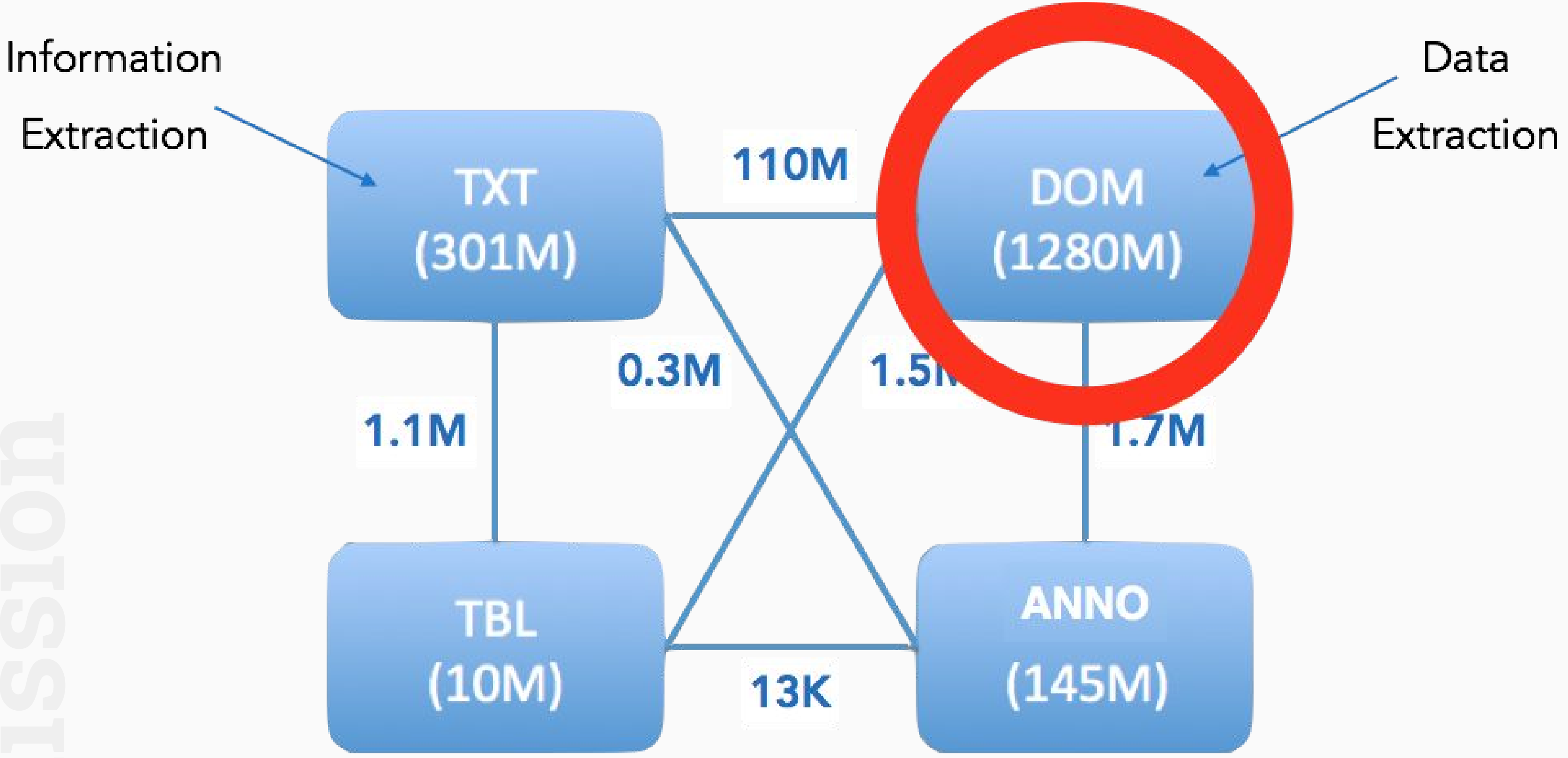
Mission

World class crawling platform

to largely automate

outside content collection

Crawling Space & Volume



Information Extraction

Data Extraction

TXT (301M)

DOM (1280M)

TBL (10M)

ANNO (145M)

110M

0.3M

1.5M

1.1M

1.7M

13K

Source: Xin Luna Dong (Google) - PVLDB '14

Crawling Coverage

For many kinds of information

one has to extract from

thousands of sites

in order to build a

comprehensive database

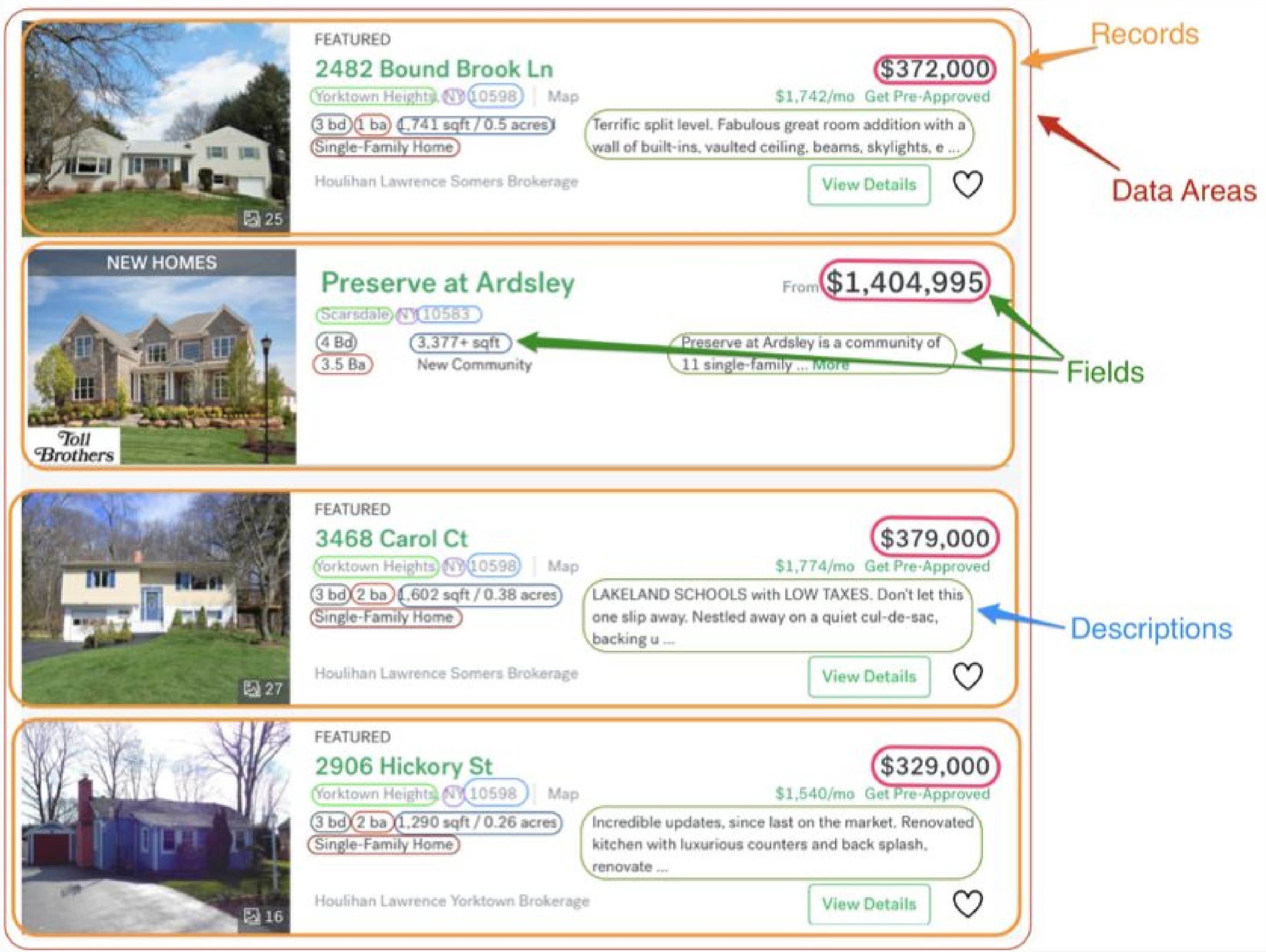
Source: Nilesh Dalvi (Yahoo!) et al. – VLDB 2012

Vertical 1 & 2:

Real Estate & Used Cars, UK

Results

Mission



FEATURED

2482 Bound Brook Ln

Yorktown Heights, NY 10598 | Map

3 bd 1 ba 1,741 sqft / 0.5 acres

Single-Family Home

Houlihan Lawrence Somers Brokerage

$372,000

$1,742/mo Get Pre-Approved

Terrific split level. Fabulous great room addition with a wall of built-ins, vaulted ceiling, beams, skylights, e ...

View Details

NEW HOMES

Toll Brothers

Preserve at Ardsley

Scarsdale, NY 10583

4 Bd 3,377+ sqft

3.5 Ba New Community

From $1,404,995

Preserve at Ardsley is a community of 11 single-family ... More

FEATURED

3468 Carol Ct

Yorktown Heights, NY 10598 | Map

3 bd 2 ba 1,602 sqft / 0.38 acres

Single-Family Home

Houlihan Lawrence Somers Brokerage

$379,000

$1,774/mo Get Pre-Approved

LAKELAND SCHOOLS with LOW TAXES. Don't let this one slip away. Nestled away on a quiet cul-de-sac, backing u ...

View Details

FEATURED

2906 Hickory St

Yorktown Heights, NY 10598 | Map

3 bd 2 ba 1,290 sqft / 0.26 acres

Single-Family Home

Houlihan Lawrence Yorktown Brokerage

$329,000

$1,540/mo Get Pre-Approved

Incredible updates, since last on the market. Renovated kitchen with luxurious counters and back splash, renovate ...

View Details

Records

Data Areas

Fields

Descriptions

Results

10,493

Sites from real-estate and used-car

92%

Effective wrappers (where we get all the data)

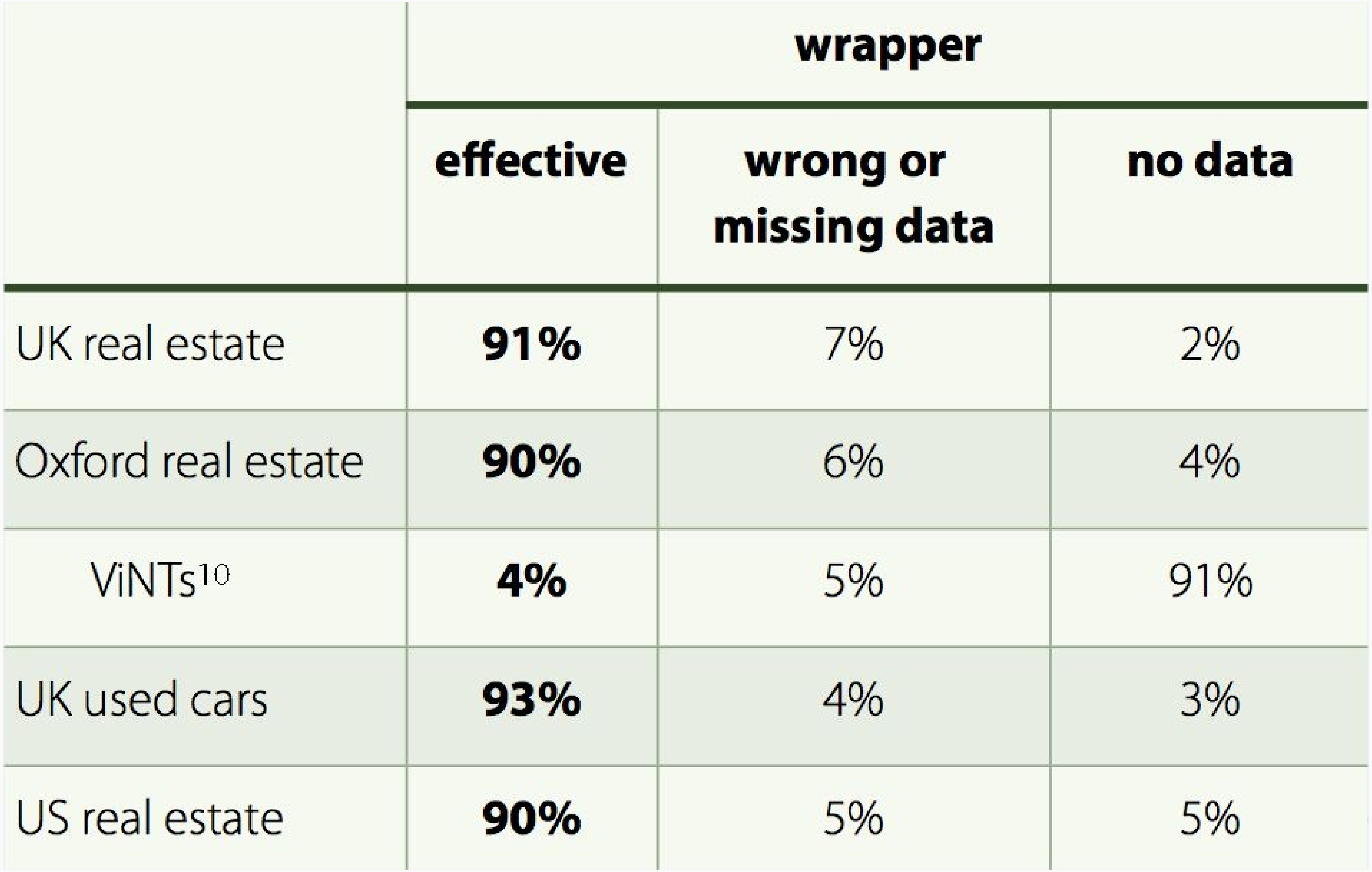
96%

Precision of extracted primary attributes

20

Days (one expert) to adjust system to a new domain

Results



	wrapper		
	effective	**wrong or missing data**	**no data**
UK real estate	**91%**	7%	2%
Oxford real estate	**90%**	6%	4%
ViNTs[10]	**4%**	5%	91%
UK used cars	**93%**	4%	3%
US real estate	**90%**	5%	5%

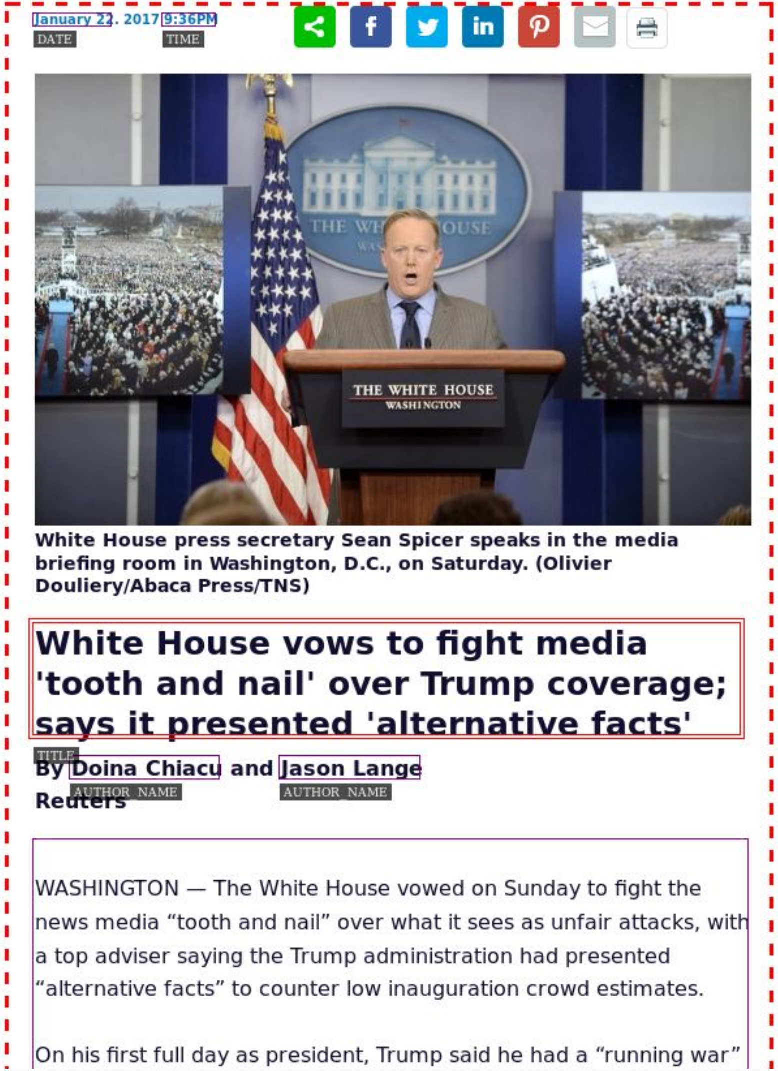
```
doc('http://www.wwagency.com/')//label[@for='sale_type_id']/following-sibling::select/{0 /}
       //form/div[@class='formbtn-ctn'][last()]/button[@class='formbtn']/{click /}
   /.:<data_area>[?.//div[@class='pagenumlinks'][1]//span/text():<number_results=.>]
   /(//div[contains(@class,'proplist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/{nextclick /})*
      //div[contains(@class,'proplist_wrap')]:<record>[? .:<origin_url=current-url()>]
      [? .//span[@class='prop_price']/text():<price=normalize-space(.)> ]
      [? .//span[.='Type:']/following-sibling::strong/text():<property_type=normalize-space(.)> ]
      [? .//div[@class='prop_statuses']//text():<property_status=normalize-space(.)> ]
      [? .//span[.='Bathrooms:']/following-sibling::strong/text():<bathroom_number=normalize-space(.)> ]
      [? .//span[.='Bedrooms:']/following-sibling::strong/text():<bedroom_number=normalize-space(.)> ]
      [? .//strong[@class='orange']/preceding-sibling::text():<location_raw=string(.)> ]
      [? .//strong[@class='orange']/text():<postcode=normalize-space(.)> ]
      [? .//strong/preceding-sibling::strong/text():<street_address=normalize-space(.)> ]
      [? .//@src:<image=normalize-space(.)> ]
      [? .//div[@class='prop_statuses']/following-sibling::a/@href:<url=normalize-space(.)> ]
      [? .//div[@class='prop_maininfo']:<description=normalize-space(.)> ]
```

Results

Vertical 3:

News Articles

Results



January 22. 2017 9:36PM

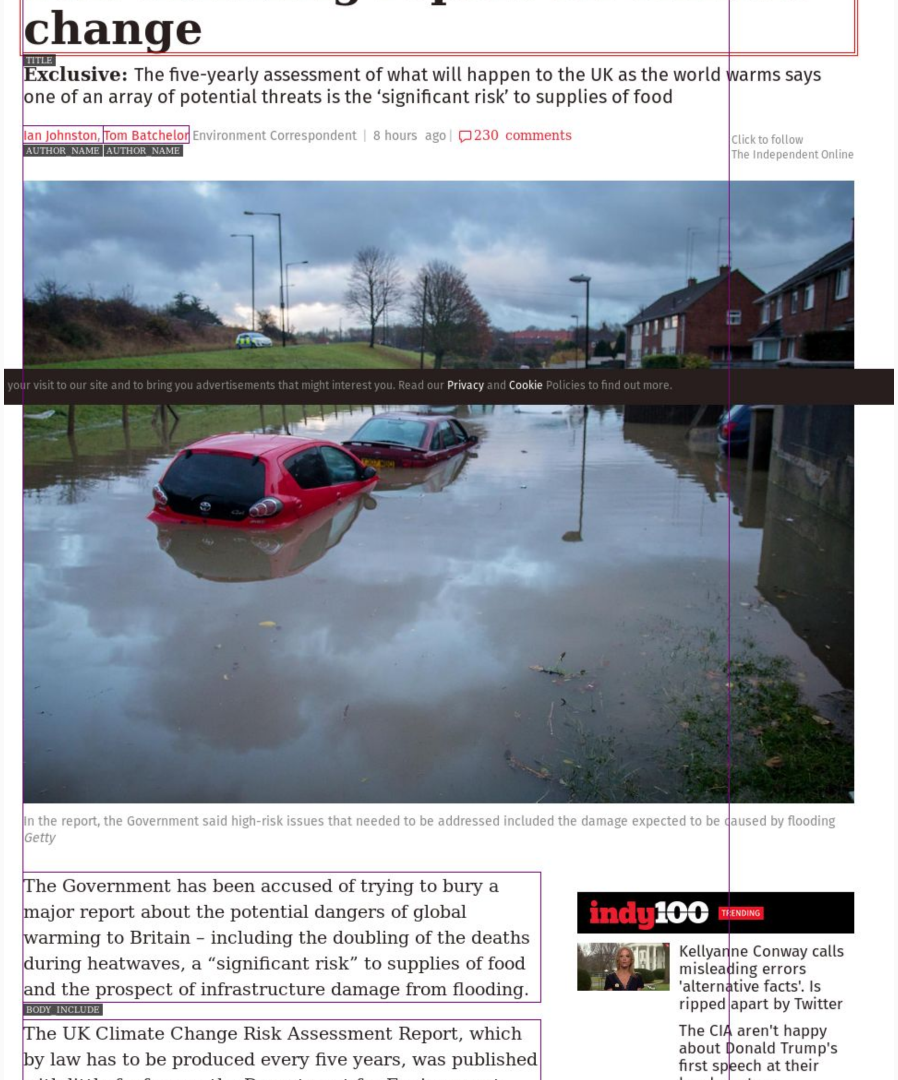
White House press secretary Sean Spicer speaks in the media briefing room in Washington, D.C., on Saturday. (Olivier Douliery/Abaca Press/TNS)

White House vows to fight media 'tooth and nail' over Trump coverage; says it presented 'alternative facts'

By Doina Chiacu and Jason Lange
Reuters

WASHINGTON — The White House vowed on Sunday to fight the news media "tooth and nail" over what it sees as unfair attacks, with a top adviser saying the Trump administration had presented "alternative facts" to counter low inauguration crowd estimates.

On his first full day as president, Trump said he had a "running war"



change

Exclusive: The five-yearly assessment of what will happen to the UK as the world warms says one of an array of potential threats is the 'significant risk' to supplies of food

Ian Johnston, Tom Batchelor Environment Correspondent | 8 hours ago | 230 comments

Click to follow
The Independent Online

your visit to our site and to bring you advertisements that might interest you. Read our Privacy and Cookie Policies to find out more.

In the report, the Government said high-risk issues that needed to be addressed included the damage expected to be caused by flooding *Getty*

The Government has been accused of trying to bury a major report about the potential dangers of global warming to Britain – including the doubling of the deaths during heatwaves, a "significant risk" to supplies of food and the prospect of infrastructure damage from flooding.

The UK Climate Change Risk Assessment Report, which by law has to be produced every five years, was published

indy100 TRENDING

Kellyanne Conway calls misleading errors 'alternative facts'. Is ripped apart by Twitter

The CIA aren't happy about Donald Trump's first speech at their

Results

~40,000

US/UK news sources from Meltwater media intelligence

85%

Effective wrappers (where we get all the data)

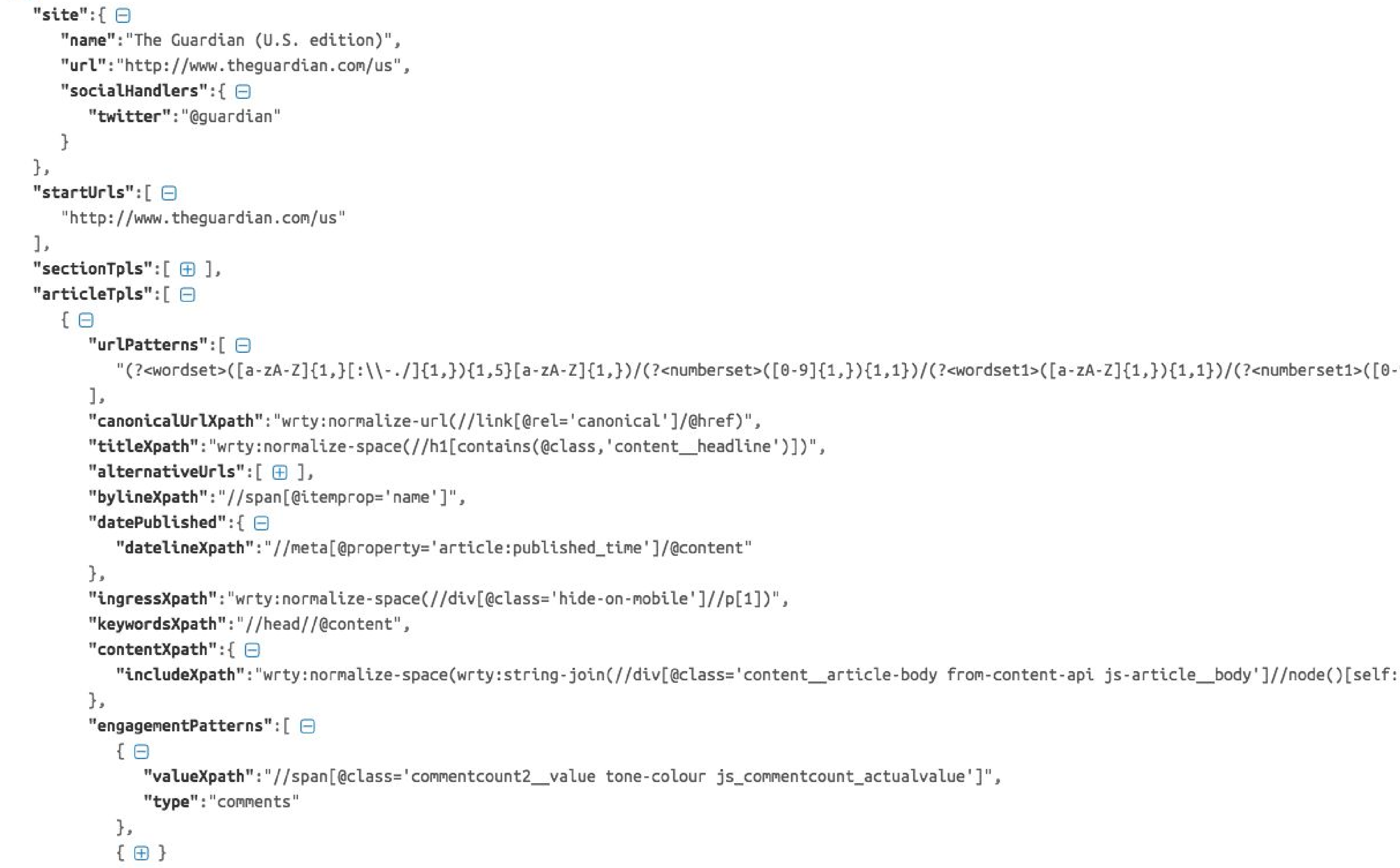
89%

Precision of extracted ADICT+ attributes

90

Days to adjust system to this vertical (why? 30min refresh)

Results



```
"site":{ ⊟
   "name":"The Guardian (U.S. edition)",
   "url":"http://www.theguardian.com/us",
   "socialHandlers":{ ⊟
      "twitter":"@guardian"
   }
},
"startUrls":[ ⊟
   "http://www.theguardian.com/us"
],
"sectionTpls":[ ⊞ ],
"articleTpls":[ ⊟
   { ⊟
      "urlPatterns":[ ⊟
         "(?<wordset>([a-zA-Z]{1,}[:\\-./]{1,}){1,5}[a-zA-Z]{1,})/(?<numberset>([0-9]{1,}){1,1})/(?<wordset1>([a-zA-Z]{1,}){1,1})/(?<numberset1>([0-
      ],
      "canonicalUrlXpath":"wrty:normalize-url(//link[@rel='canonical']/@href)",
      "titleXpath":"wrty:normalize-space(//h1[contains(@class,'content__headline')])",
      "alternativeUrls":[ ⊞ ],
      "bylineXpath":"//span[@itemprop='name']",
      "datePublished":{ ⊟
         "datelineXpath":"//meta[@property='article:published_time']/@content"
      },
      "ingressXpath":"wrty:normalize-space(//div[@class='hide-on-mobile']//p[1])",
      "keywordsXpath":"//head//@content",
      "contentXpath":{ ⊟
         "includeXpath":"wrty:normalize-space(wrty:string-join(//div[@class='content__article-body from-content-api js-article__body']//node()[self:
      },
      "engagementPatterns":[ ⊟
         { ⊟
            "valueXpath":"//span[@class='commentcount2__value tone-colour js_commentcount_actualvalue']",
            "type":"comments"
         },
         { ⊞ }
```

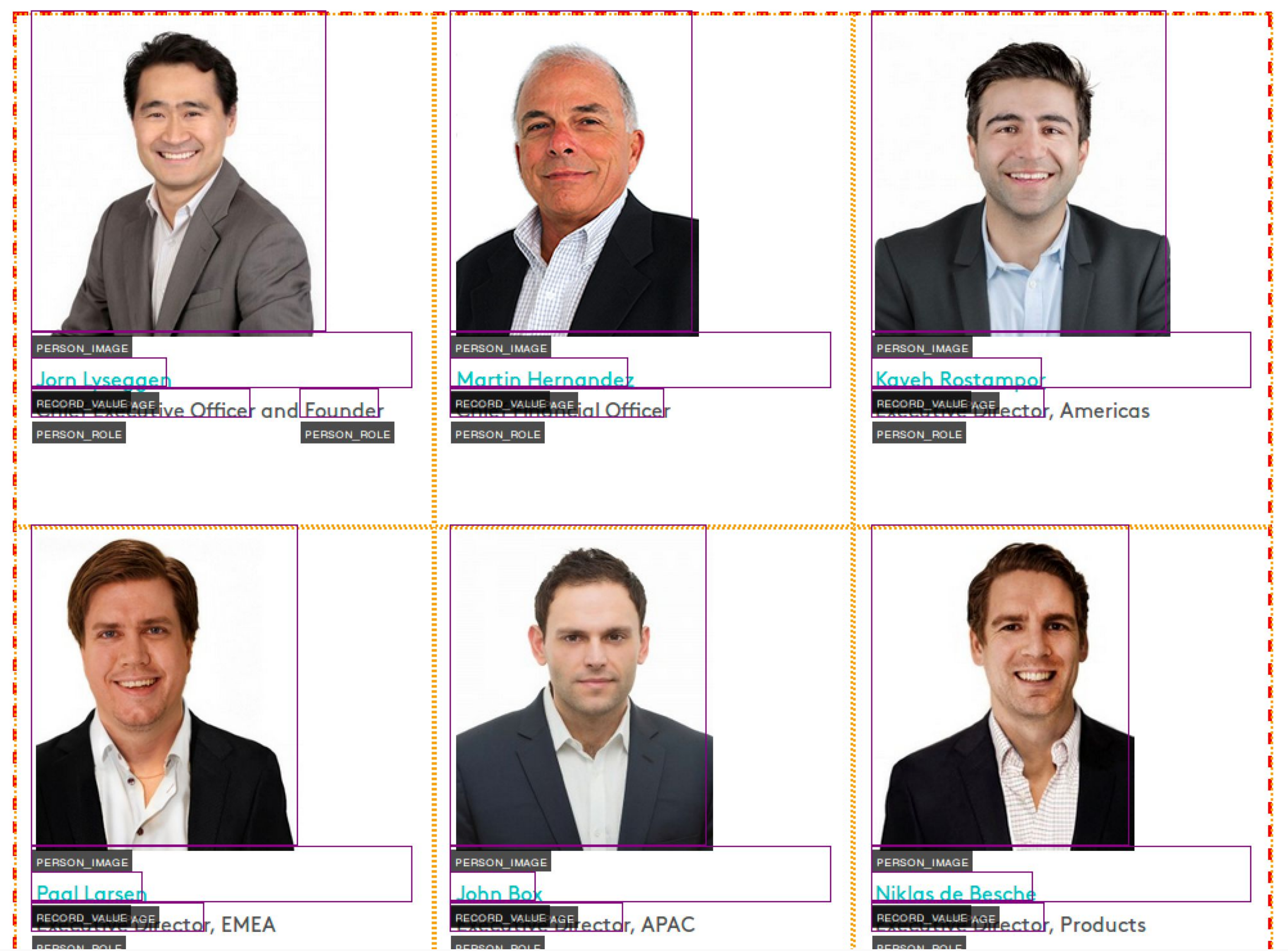
Why different wrapper format?

- **OXPath**: perfect for interactive, search engine style websites
- However: in **media intelligence** – freshness of data is critical
    - 30min maximum between publishing and indexing time
    - (almost) every article has an indexable, unique URL
    - large variety of different article templates
- **Decompose** OXPath wrapper into **single page segments**
    - memorise set of section pages encountered in a run
    - recrawl stored section pages in next run
        - to find new article (& section) pages

Vertical 4:

Company Extractors

Results



PERSON_IMAGE

Jorn Lyseggen

RECORD_VALUE AGE

Chief Executive Officer and Founder

PERSON_ROLE PERSON_ROLE

PERSON_IMAGE

Martin Hernandez

RECORD_VALUE AGE

Chief Financial Officer

PERSON_ROLE

PERSON_IMAGE

Kaveh Rostampor

RECORD_VALUE AGE

Executive Director, Americas

PERSON_ROLE

PERSON_IMAGE

Paal Larsen

RECORD_VALUE AGE

Executive Director, EMEA

PERSON_ROLE

PERSON_IMAGE

John Box

RECORD_VALUE AGE

Executive Director, APAC

PERSON_ROLE

PERSON_IMAGE

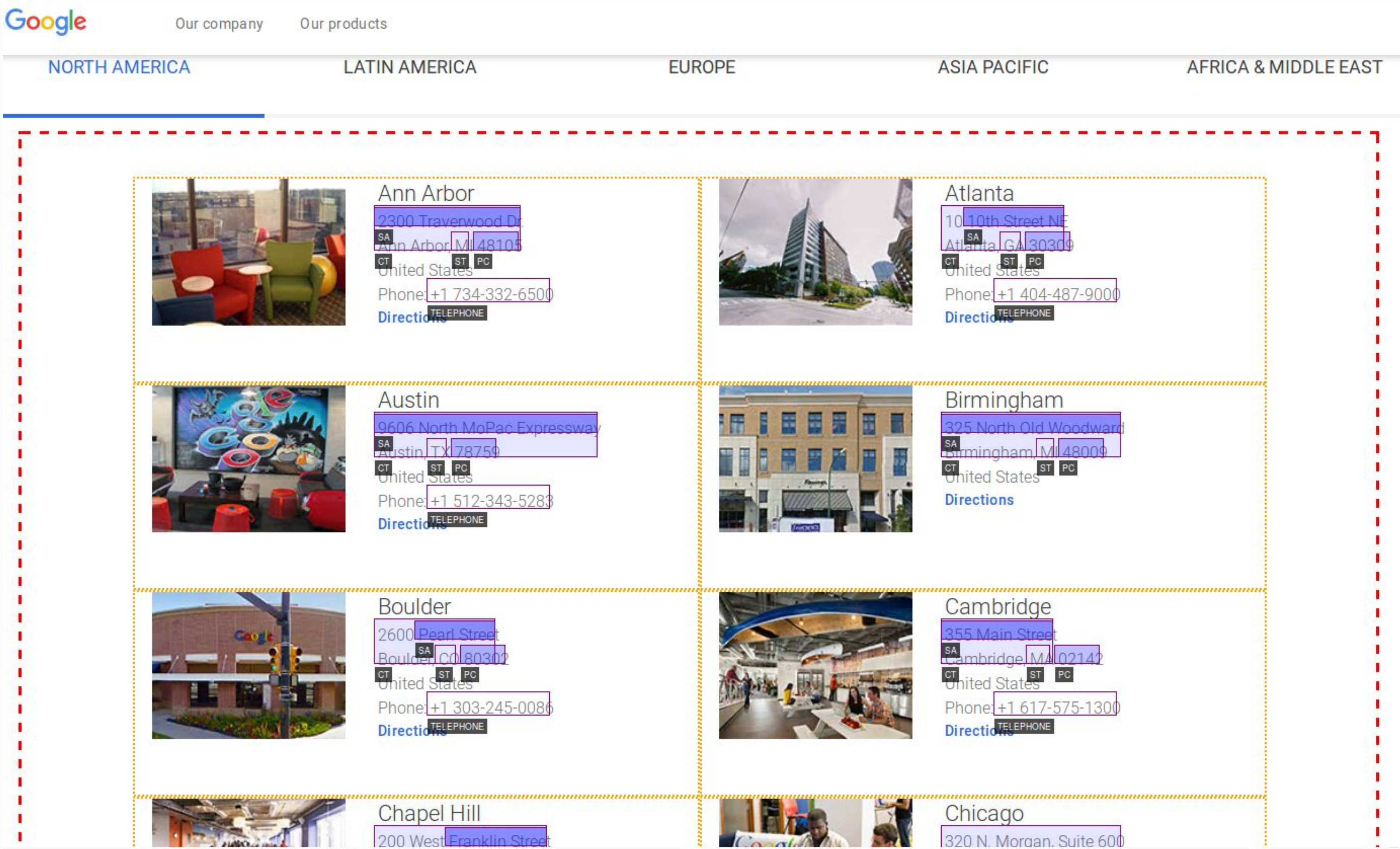
Niklas de Besche

RECORD_VALUE AGE

Executive Director, Products

PERSON_ROLE

Results



Google Our company Our products

NORTH AMERICA LATIN AMERICA EUROPE ASIA PACIFIC AFRICA & MIDDLE EAST

Ann Arbor
2300 Traverwood Dr.
Ann Arbor, MI 48105
United States
Phone: +1 734-332-6500
Directions

Atlanta
10 10th Street NE
Atlanta, GA 30309
United States
Phone: +1 404-487-9000
Directions

Austin
9606 North MoPac Expressway
Austin, TX 78759
United States
Phone: +1 512-343-5283
Directions

Birmingham
325 North Old Woodward
Birmingham, MI 48009
United States
Directions

Boulder
2600 Pearl Street
Boulder, CO 80302
United States
Phone: +1 303-245-0086
Directions

Cambridge
355 Main Street
Cambridge, MA 02142
United States
Phone: +1 617-575-1300
Directions

Chapel Hill
200 West Franklin Street

Chicago
320 N. Morgan, Suite 600

Company Extraction: **Goals**

- Given **only** a company website
    - Extract as much relevant information from structured sources
        - executive team, locations, subsidiaries, …
    - Identify unstructured sources
        - press releases, financial reports, …
- **Scale** to millions of companies in multiple languages

Company Extraction: **Results**

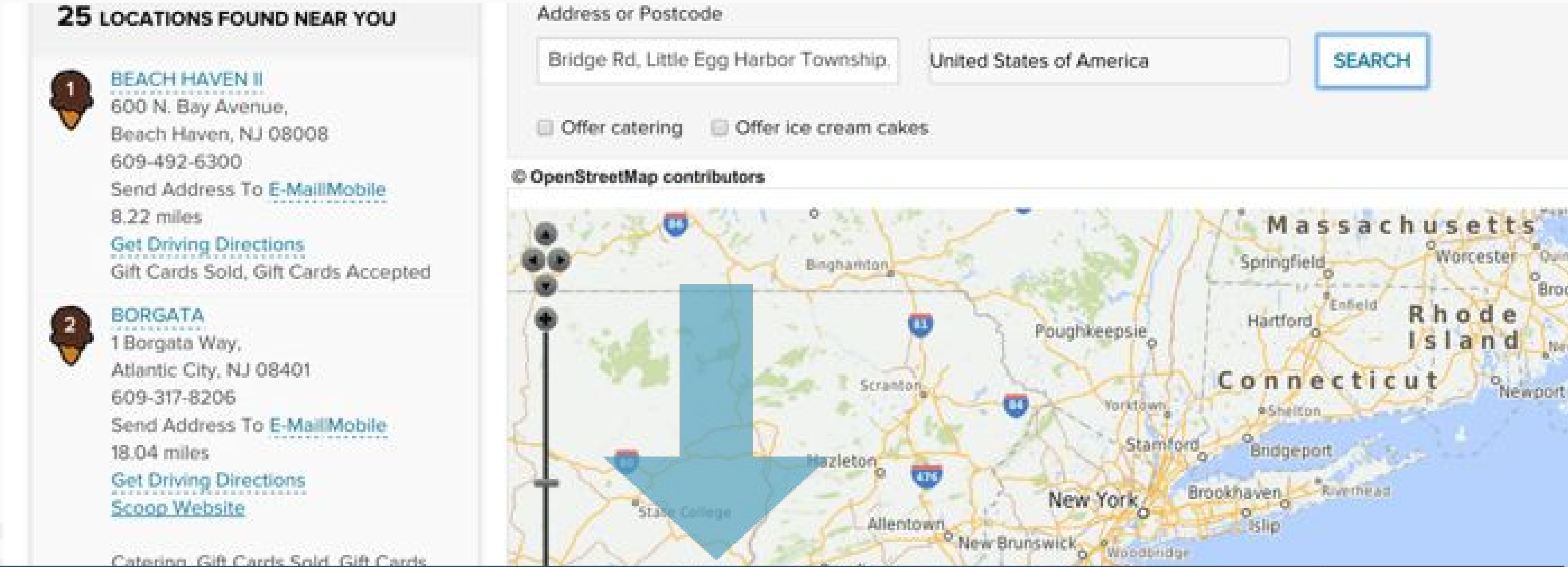
	% of companies			
	CEO & execs	description	location	logo
recall	85%	74%	82%	89%
present	58%	79%	69%	74%

Results

Restaurant locations

Don’t believe us? You aren’t the first – major US technology company

- **Need:** US restaurant locations (including chains) for check-ins
- **Problem:** existing location databases incomplete and full of errors
- **Want:** Get that data from the “authoritative” source, i.e.,
  - the restaurant (chain) websites
- They evaluated state-of-the-art – most solutions to crude
  - Settled on scrapy, but: 2 months for top 20 US chains
  - Very worried about maintenance



25 LOCATIONS FOUND NEAR YOU

1 BEACH HAVEN II
600 N. Bay Avenue,
Beach Haven, NJ 08008
609-492-6300
Send Address To E-MailMobile
8.22 miles
Get Driving Directions
Gift Cards Sold, Gift Cards Accepted

2 BORGATA
1 Borgata Way,
Atlantic City, NJ 08401
609-317-8206
Send Address To E-MailMobile
18.04 miles
Get Driving Directions
Scoop Website

Catering, Gift Cards Sold, Gift Cards

Address or Postcode

Bridge Rd, Little Egg Harbor Township, | United States of America | SEARCH

□ Offer catering □ Offer ice cream cakes

© OpenStreetMap contributors

NAME	STREET ADDRESS	LOCALITY	STATE	POSTCODE	PHONE
BEACH HAVEN II	600 N. Bay Avenue	Beach Haven	NJ	08008	609-492-6300
BORGATA	1 Borgata Way	Atlantic city	NJ	08401	609-317-8206
30TH STREET STATION	30TH STREET STATION	Philadelphia	PA	19104	215-222-2996

Restaurant location: **Results**

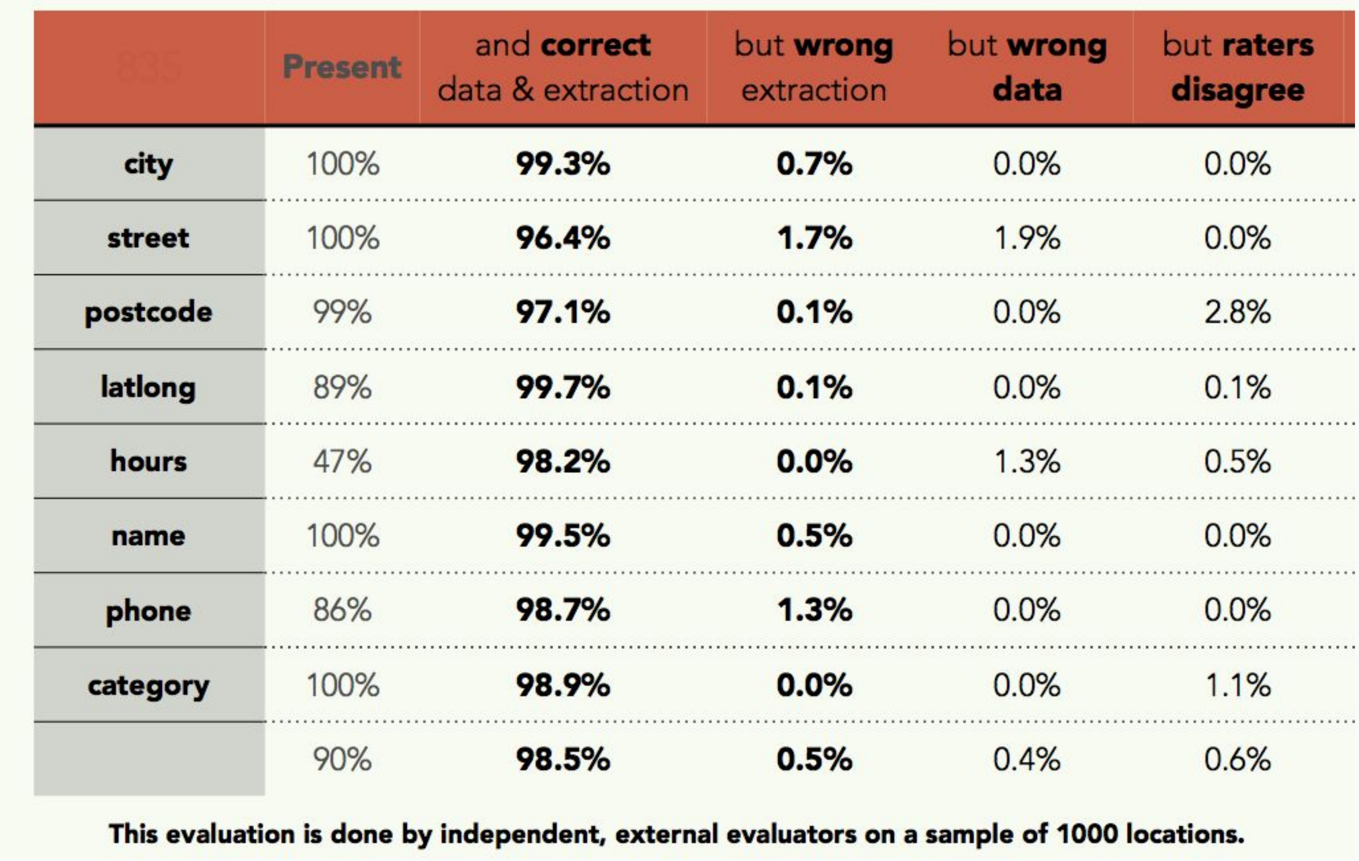
- After 1 month applying Wrapidity technology:
  - over 300 US chains, over 100k websites
  - more than 3M locations in total
  - fully automated maintenance for those sources

coverage **85%**

precision **95%**

- But: *they still didn’t believe*
  - hired Accenture to assess quality of the data
  - result: **over 97% precision**

Restaurant location: **Independent Evaluation**



835	**Present**	and **correct** data & extraction	but **wrong** extraction	but **wrong** **data**	but **raters** **disagree**
city	100%	**99.3%**	**0.7%**	0.0%	0.0%
street	100%	**96.4%**	**1.7%**	1.9%	0.0%
postcode	99%	**97.1%**	**0.1%**	0.0%	2.8%
latlong	89%	**99.7%**	**0.1%**	0.0%	0.1%
hours	47%	**98.2%**	**0.0%**	1.3%	0.5%
name	100%	**99.5%**	**0.5%**	0.0%	0.0%
phone	86%	**98.7%**	**1.3%**	0.0%	0.0%
category	100%	**98.9%**	**0.0%**	0.0%	1.1%
	90%	**98.5%**	**0.5%**	0.4%	0.6%

This evaluation is done by independent, external evaluators on a sample of 1000 locations.

Results

Summary

NEWS

~50k sources, 30 min recrawl interval

87%

90%

over 300 US chains, over 100k websites;
more than 3M locations

FORK, KNIFE & GLASS
DINNER CLUB

COMPANY NAME
Company Slogan

over 30 attributes, where present; 1M+ company's site crawl ongoing

91%

Approach

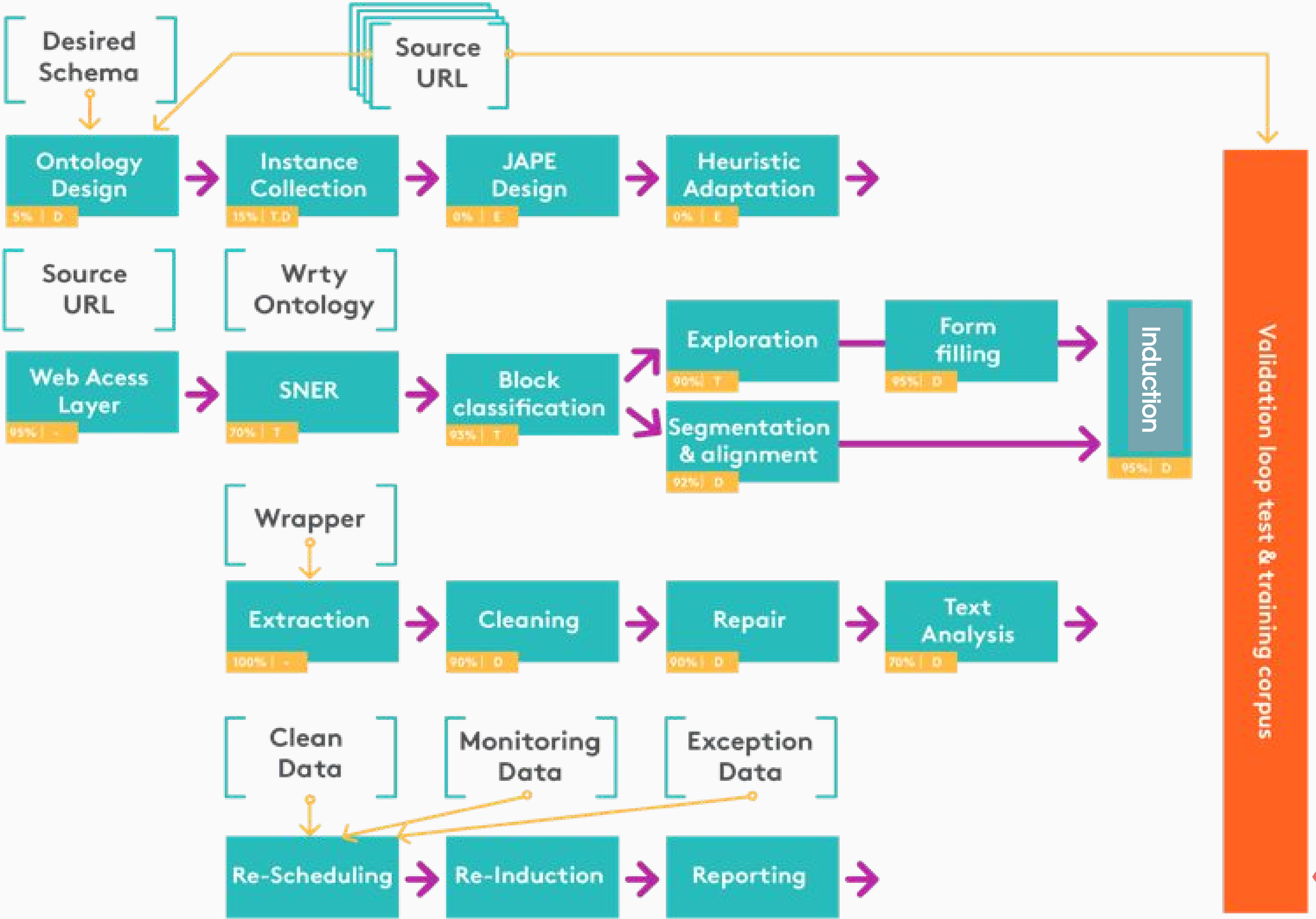
Web-Scale Wrapper Induction

- We need to **scale to the web**
  - minimize supervision per source
- But: we can afford **prior knowledge**
  - about entities and attributes
  - mostly in form of known **labels** & **instances** and “**appearance**”
    - expressed as Gazetteers or rules for local, textual information
    - higher-level rules or classifiers for complex structures

Web-Scale Wrapper Induction

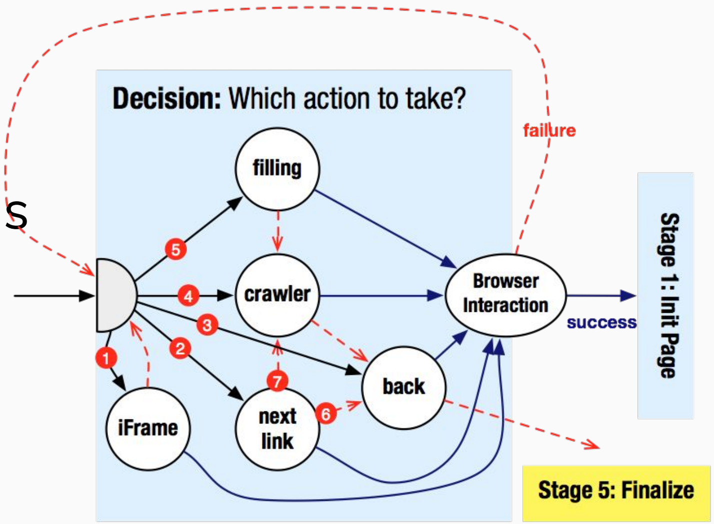
- **Problem:** application of prior knowledge is costly & noisy
  - wrapper induction to generalise to other pages of site
  - “template” hypothesis
- **Solution:** Generate “wrapper” program from examples
  - then apply to all pages of a site
  - when to apply which extractor
- **Full site extraction** needs to also deal with
  - Interactivity such as pagination & form filling (deep web)
  - Detecting complex structures such as lists, tables, …

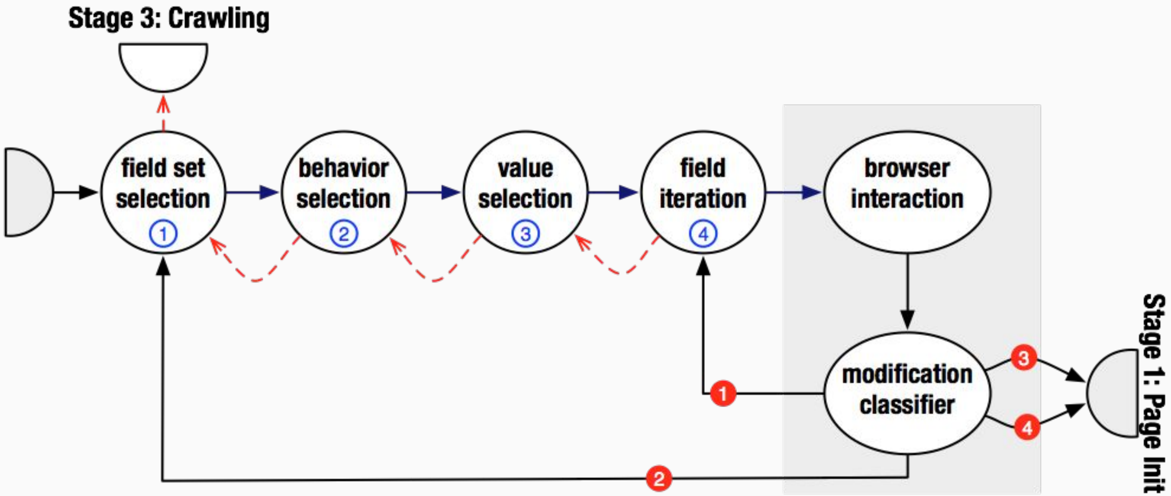
Approach


Desired Schema
Source URL
Ontology Design
5% | D
Instance Collection
15% | T,D
JAPE Design
0% | E
Heuristic Adaptation
0% | E
Source URL
Wrty Ontology
Web Acess Layer
95% | -
SNER
70% | T
Block classification
93% | T
Exploration
90% | T
Form filling
95% | D
Segmentation & alignment
92% | D
Induction
95% | D
Wrapper
Extraction
100% | -
Cleaning
90% | D
Repair
90% | D
Text Analysis
70% | D
Clean Data
Monitoring Data
Exception Data
Re-Scheduling
Re-Induction
Reporting
Validation loop test & training corpus

Exploration: Self-Adaptive

- Self-adaptive, dynamic exploration plans
  - planers expressed as guarded FSTs
  - with Datalog rules as guards
  - 1000's of unique exploration plans

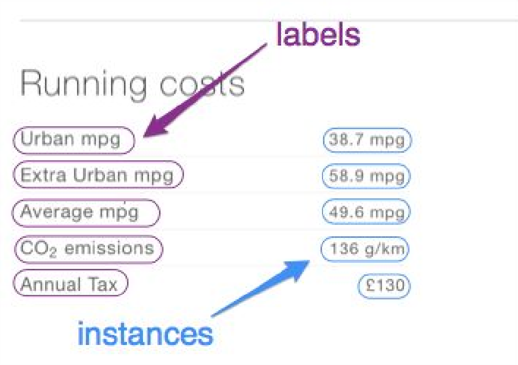




NER for DOMs: Labels, structure, ...

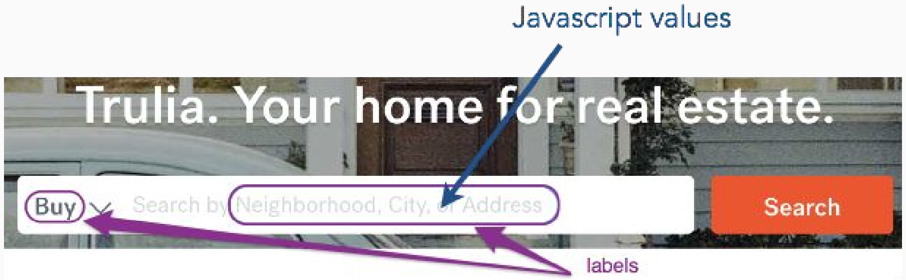
Labels and instances, visible and invisible (HTML structure, Javascript values)

Approach

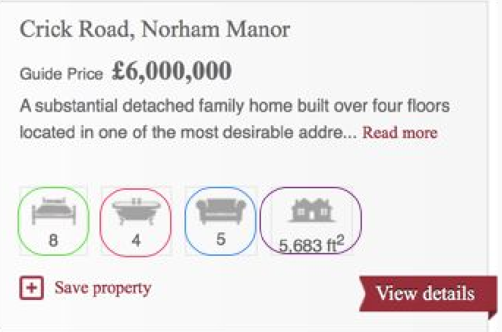


Running costs

Urban mpg	38.7 mpg
Extra Urban mpg	58.9 mpg
Average mpg	49.6 mpg
CO_2 emissions	136 g/km
Annual Tax	£130



Javascript values
Trulia. Your home for real estate.
Buy
Search by Neighborhood, City, or Address
Search
labels



Crick Road, Norham Manor

Guide Price £6,000,000

A substantial detached family home built over four floors located in one of the most desirable addre... Read more

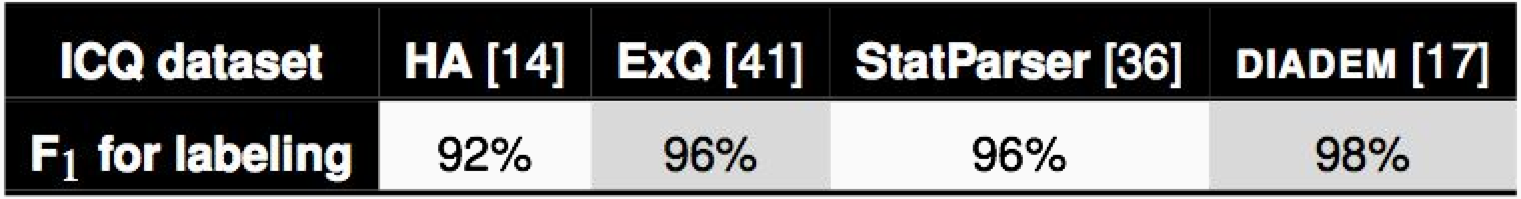
```
<div class="icon first">
  <img src=".../bdes.jpg" alt="Bedrooms" title="Bedrooms">
  <br>8
</div>
<div class="icon">
  <img src=".../bath.jpg" alt="Bathrooms" title="Bathrooms">
  <br>4
</div>
```

Form understanding

WORLD WIDE WEB 2012 LYON - FRANCE

Approach

- Sub-problems: form **labeling**, form **segmentation**, **classification**
- **Combines** structural, textual, visual, and semantic clues
  - structural = structure of the DOM, e.g., distance
  - visual = rendering of the form, e.g., for alignment
  - textual = detectors for a vertical's types (e.g., "LHR")
  - semantic = class, id, … with semantic labels (e.g. "finput_dest")
- **Polynomial time** labeling, grouping, and classification algorithm



ICQ dataset	HA [14]	ExQ [41]	StatParser [36]	DIADEM [17]
F_1 for labeling	92%	96%	96%	98%

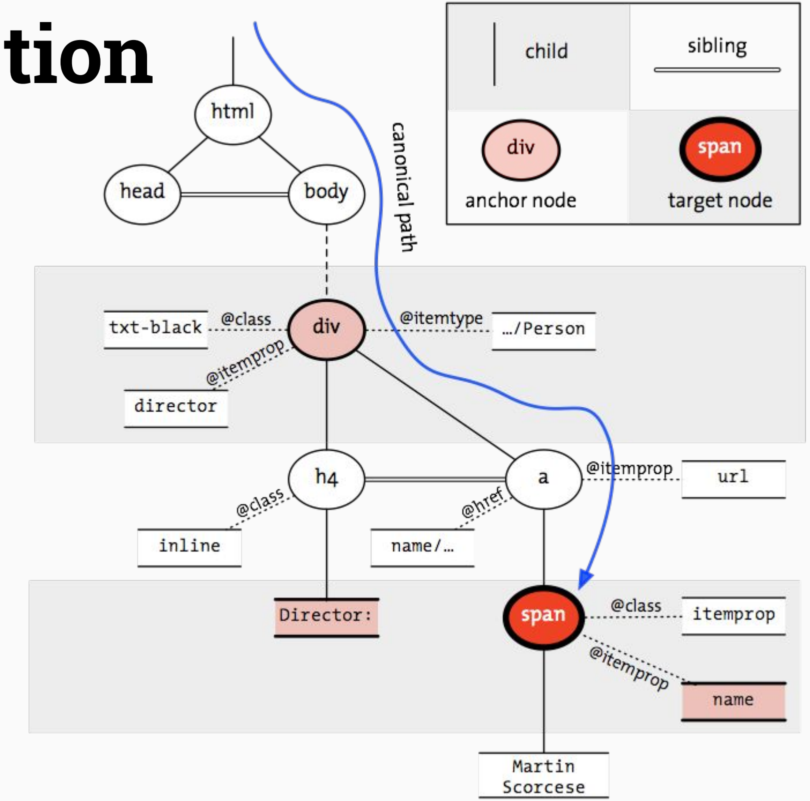
Pick a Path: Wrapper induction

- **Pick** robust, “semantic” paths
  - less affected by changes
  - over time and within a template
- Suitable as “foundation” for
  - **template discovery**
- E.g.: Select the director
  - Firebug (“canonical” XPath)

```
/html[1]/body[1]/ ... /div[4]/a[1]/span[1]
```

  - Ours:

```
//div[starts-with(.,”Director:")]//span[(@class="itemprop")
```



Wrapper Repair

ICDE'16

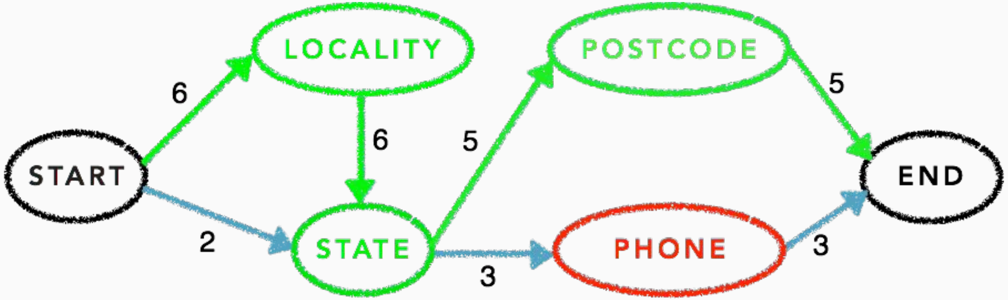
Postcode	Phone	Locality	State	Street Address
San Diego, CA 92101	619-234-1802	San Diego, CA 92101	<NULL>	471 Horton Plaza, near Westland park
Boise, ID 83702	208-342-1992	Boise, ID 83702	<NULL>	103 North 10th Street
Portland, OR 97209	503-796-3033	Portland, OR 97209	<NULL>	301 NW 10th Avenue, near the Fish Market

Postcode	Phone	Locality	State	Street Address
92101	619-234-1802	San Diego	CA	471 Horton Plaza
83702	208-342-1992	Boise	ID	103 North 10th Street
97209	503-796-3033	Portland	OR	301 NW 10th Avenue

San Diego, CA 92101

Boise, ID 83702

Portland, OR 97209



🟩 Locality

🟦 State

🟥 Postcode

🟧 Phone

Wrapper Repair

- **Joint repair** of wrapper and output data (relation)
- Problem related to table segmentation problem
    - generally NP-complete
    - but we show that it's polynomial under atomic misplacement
- **Atomic misplacement:** attribute value is either
    - entirely misplaced, or
    - its fragments are in adjacent fields

Results

1 **ROSeAnn** (VLDB'14)
Entity extraction from DOMs

2 **OPAL** (WWW'12, VLDBJ'13)
Form understanding & filling

3 **AMBER** (ICWE'11)
Record identification for lists

4 **OXPath** (VLDB'11, VLDBJ'13)
Extraction language

5 **Robust XPaths** (SIGMOD'16)
Change-resistant wrappers

6 **Oxtractor** (Coling'16)
Attribute extraction

7 **WaDaR** (ICDE'16)
Joint **wrapper & relation repair**

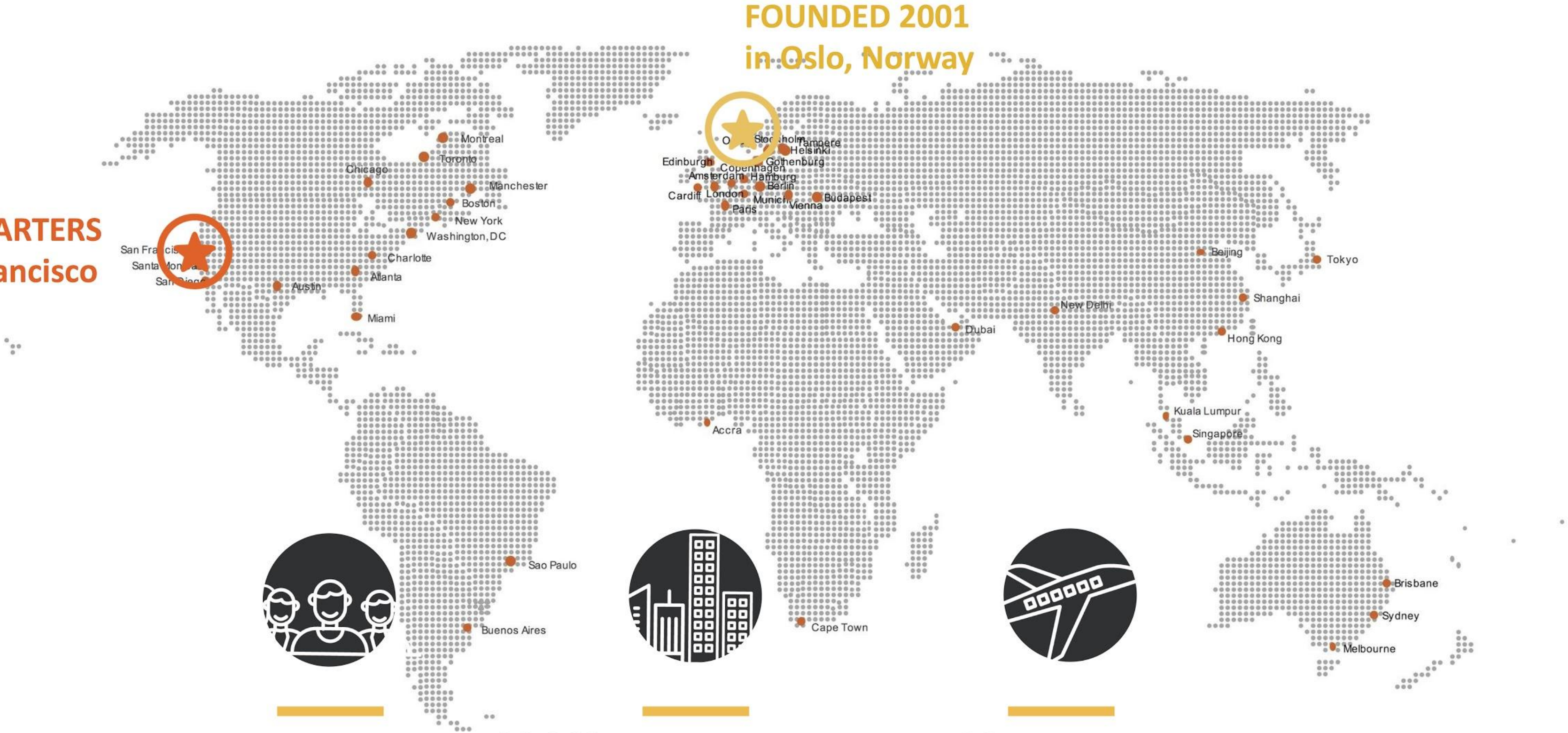
8 **VADA** (EDBT'16)
Wrangling of extracted data (in progress)

6 **DIADEM** (VLDB'14)
World-first accurate, automatic **full-site extraction system**

citi
HSBC
HSBC

citi
HSBC
HSBC

Meltwater: Who are we?



FOUNDED 2001
in Oslo, Norway

HEADQUARTERS
in San Francisco

1500 employees
worldwide

26,000 Business customers
in **108** countries

60 offices across
27 Countries

Meltwater: **Media Intelligence**

Sources: Editorial, Social, Broadcasts

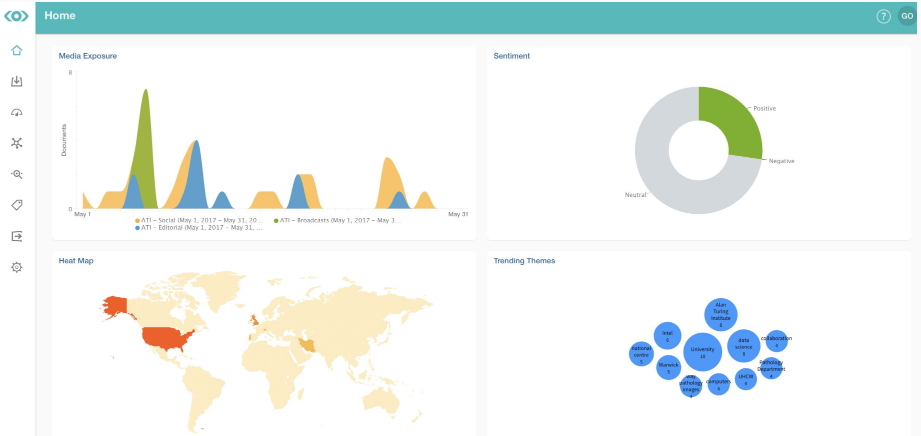
media exposure

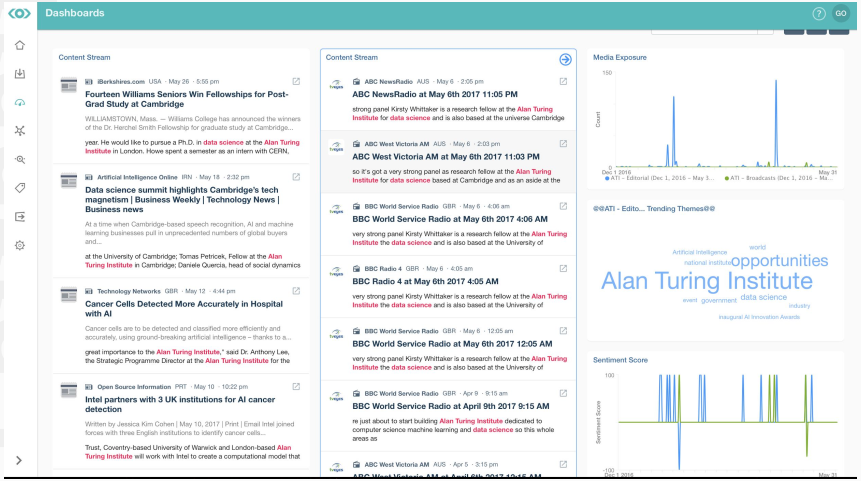
trends

influencers

sentiment analysis

More than 300k different types of user queries





Content Stream

iBerkshires.com USA · May 26 · 5:55 pm

Fourteen Williams Seniors Win Fellowships for Post-Grad Study at Cambridge

WILLIAMSTOWN, Mass. — Williams College has announced the winners of the Dr. Herchel Smith Fellowship for graduate study at Cambridge...

year. He would like to pursue a Ph.D. in data science at the Alan Turing Institute in London. Howe spent a semester as an intern with CERN,

Artificial Intelligence Online IRN · May 18 · 2:32 pm

Data science summit highlights Cambridge's tech magnetism | Business Weekly | Technology News | Business news

At a time when Cambridge-based speech recognition, AI and machine learning businesses pull in unprecedented numbers of global buyers and...

at the University of Cambridge; Tomas Petricek, Fellow at the Alan Turing Institute in Cambridge; Daniele Quercia, head of social dynamics

Technology Networks GBR · May 12 · 4:44 pm

Cancer Cells Detected More Accurately in Hospital with AI

Cancer cells are to be detected and classified more efficiently and accurately, using ground-breaking artificial intelligence – thanks to a...

great importance to the Alan Turing Institute," said Dr. Anthony Lee, the Strategic Programme Director at the Alan Turing Institute for the

Open Source Information PRT · May 10 · 10:22 pm

Intel partners with 3 UK institutions for AI cancer detection

Written by Jessica Kim Cohen | May 10, 2017 | Print | Email Intel joined forces with three English institutions to identify cancer cells...

Trust, Coventry-based University of Warwick and London-based Alan Turing Institute will work with Intel to create a computational model that

Content Stream

ABC NewsRadio AUS · May 6 · 2:05 pm

ABC NewsRadio at May 6th 2017 11:05 PM

strong panel Kirsty Whittaker is a research fellow at the Alan Turing Institute for data science and is also based at the universe Cambridge

ABC West Victoria AM AUS · May 6 · 2:03 pm

ABC West Victoria AM at May 6th 2017 11:03 PM

so it's got a very strong panel as research fellow at the Alan Turing Institute for data science based at Cambridge and as an aside at the

BBC World Service Radio GBR · May 6 · 4:06 am

BBC World Service Radio at May 6th 2017 4:06 AM

very strong panel Kirsty Whittaker is a research fellow at the Alan Turing Institute the data science and is also based at the University of

BBC Radio 4 GBR · May 6 · 4:05 am

BBC Radio 4 at May 6th 2017 4:05 AM

very strong panel Kirsty Whittaker is a research fellow at the Alan Turing Institute the data science and is also based at the University of

BBC World Service Radio GBR · May 6 · 12:05 am

BBC World Service Radio at May 6th 2017 12:05 AM

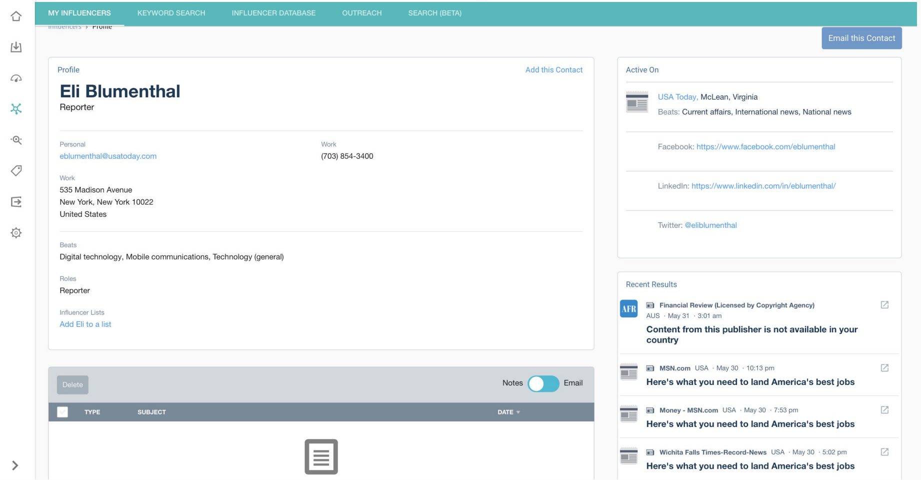
very strong panel Kirsty Whittaker is a research fellow at the Alan Turing Institute the data science and is also based at the University of

BBC World Service Radio GBR · Apr 9 · 9:15 am

BBC World Service Radio at April 9th 2017 9:15 AM

re just about to start building Alan Turing Institute dedicated to computer science machine learning and data science so this whole areas as

ABC West Victoria AM AUS · Apr 5 · 3:15 pm



Profile

Eli Blumenthal

Reporter

Personal
eblumenthal@usatoday.com

Work
(703) 854-3400

Work
535 Madison Avenue
New York, New York 10022
United States

Beats
Digital technology, Mobile communications, Technology (general)

Roles
Reporter

Influencer Lists
Add Eli to a list

Active On

USA Today, McLean, Virginia
Beats: Current affairs, International news, National news

Facebook: https://www.facebook.com/eblumenthal

LinkedIn: https://www.linkedin.com/in/eblumenthal/

Twitter: @eliblumenthal

Recent Results

Financial Review (Licensed by Copyright Agency) AUS · May 31 · 3:01 am

Content from this publisher is not available in your country

MSN.com USA · May 30 · 10:13 pm

Here's what you need to land America's best jobs

Money - MSN.com USA · May 30 · 7:53 pm

Here's what you need to land America's best jobs

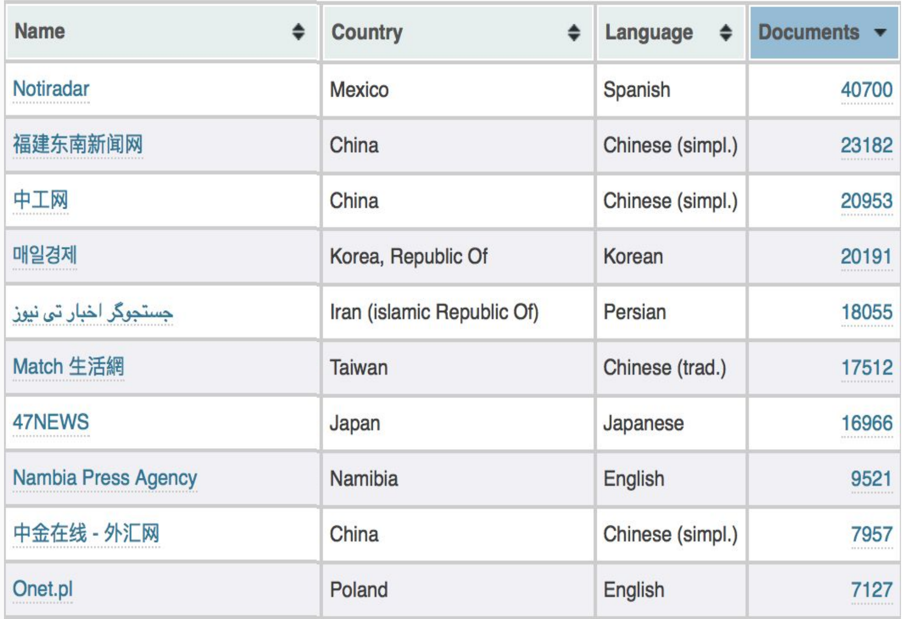
Wichita Falls Times-Record-News USA · May 30 · 5:02 pm

Here's what you need to land America's best jobs

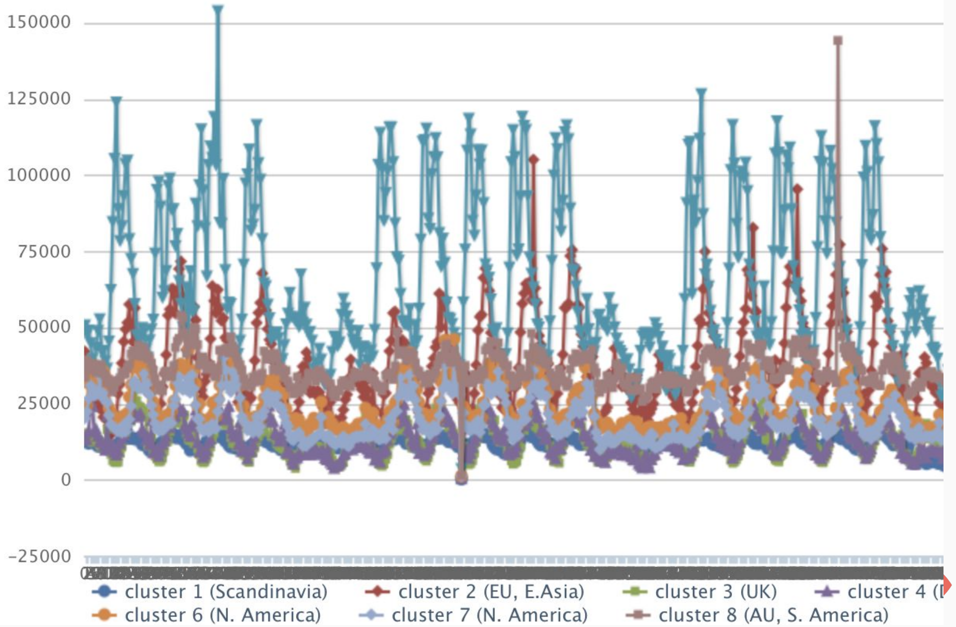
Meltwater: **In Numbers**

~**200B indexed documents**

- Crawlers fetch ~**3.3M** articles/day from 190k editorial sources
    - re-crawled every 30 minutes
- With the social fire hoses we go up to 30M docs/ day.



Name	Country	Language	Documents
Notiradar	Mexico	Spanish	40700
福建东南新闻网	China	Chinese (simpl.)	23182
中工网	China	Chinese (simpl.)	20953
매일경제	Korea, Republic Of	Korean	20191
جستجوگر اخبار تی نیوز	Iran (islamic Republic Of)	Persian	18055
Match 生活網	Taiwan	Chinese (trad.)	17512
47NEWS	Japan	Japanese	16966
Nambia Press Agency	Namibia	English	9521
中金在线 - 外汇网	China	Chinese (simpl.)	7957
Onet.pl	Poland	English	7127



Meltwater: **Existing Technology Stack**

Ingestion:

- Social media hoses (partnerships)
- Editorial/News (partnerships + web crawling)
- Broadcasts (views on the above)

Enrichments (15 languages):

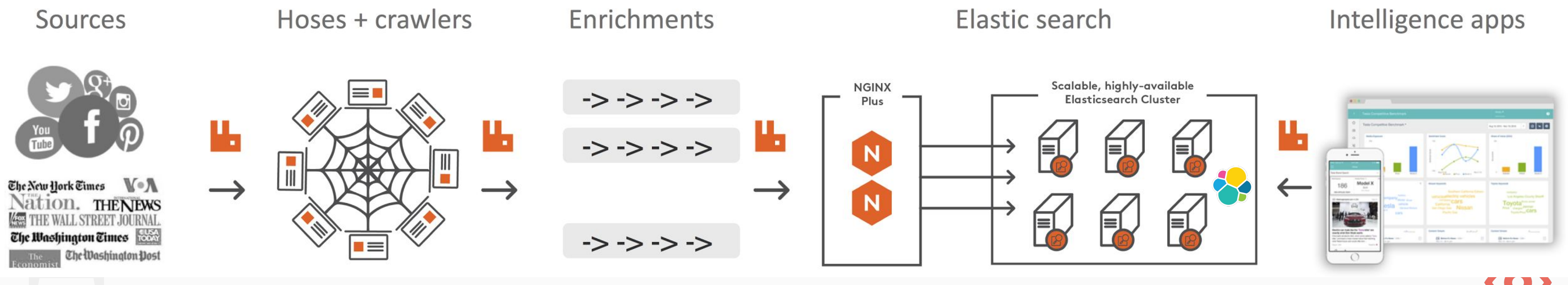
- Text categorization (topic, *language*)
- NERD (*person*, *location*, *organization*, …)
- NED (https://en.wikipedia.org/wiki/Tim_Cook)
- Sentiment Analysis

Storage and search:

- Elastic search
- Rabbit MQ (distributed queues)
- AWS

Media Intelligence applications (Custom)

- Boolean queries (*keywords / entities*)
- Counters
- Aggregates
- Drill downs / pivoting



Vision: **Insight** Building on **Outside** Data

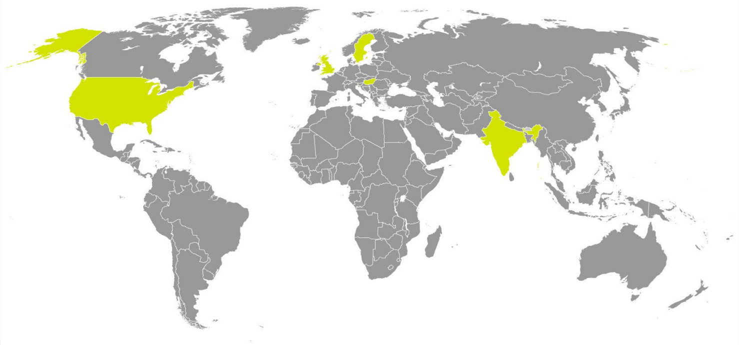
Build a world class AI platform for a new software category

Outside Insight

FAIRHAIR

Fairhair: People & Community

5 Data Science **Research offices**



6 **Data Science Hubs** (co-working spaces)

- London
- San Francisco
- Singapore
- Sydney
- Berlin
- New York

SHACK15
The Data Science Hub

University **collaborations**

Stanford University

Carnegie Mellon University

ASU Arizona State University

Università della Calabria

mest

Meltwater Entrepreneurial School of Technology

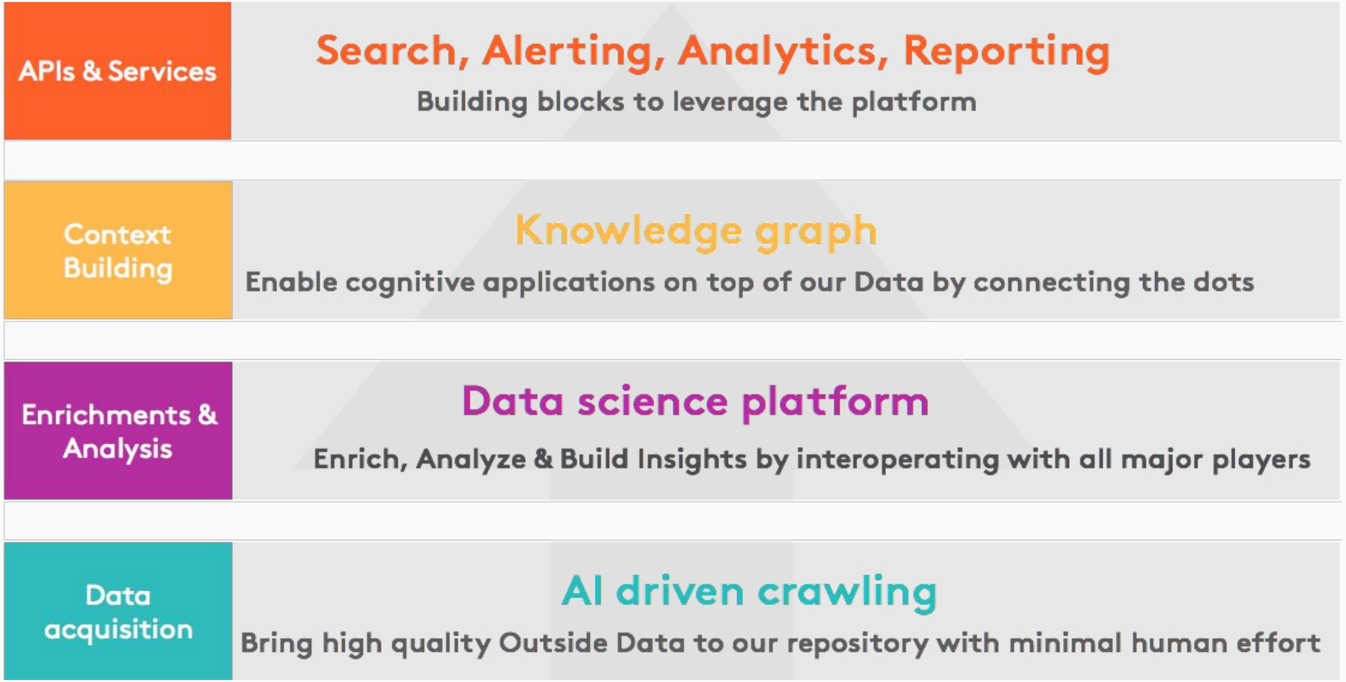
- campuses in Ghana and Nigeria
- it's a school for African entrepreneurs
- it's an incubator (33 startups)
- it's a networking hub

Fairhair’s AI First Approach

Step 1: Outside Data acquisition & making it available in a form that’s crunchable.

Step 2: Make Data Science (data, algorithms, infrastructure, tools) power everything

Step 3: We can’t foresee all uses of data and insights → Developer APIs & integrations



APIs & Services — Search, Alerting, Analytics, Reporting

Building blocks to leverage the platform

Context Building — Knowledge graph

Enable cognitive applications on top of our Data by connecting the dots

Enrichments & Analysis — Data science platform

Enrich, Analyze & Build Insights by interoperating with all major players

Data acquisition — AI driven crawling

Bring high quality Outside Data to our repository with minimal human effort

Fairhair's AI Crawlers

Traditional scraping requires a **huge human effort:**

- **Code** wrappers for each source, e.g., in Scrapy or MW's source configurations
- **Visually** testing and support tool (ala Connotate, Mozenda, ...)
- **Automatic** scraping for small number of fixed data types (ala Diffbot), e.g., Microdata
- Meltwater (old): ~50 "source engineers" maintaining manual wrappers
  - sources failing at a rate of 100's per week, 1-2h to fix each source effectively

80-90% lower

human effort

without loss in quality compared with state-of-the-art

3-10x more

attributes

and domains than existing automated solutions and affordable supervised one

10-100x

more sources

e.g., 100k+ restaurant websites, 300k+ news sources, 1M+ of company websites

Data and Content Lake

Factual information: wherefrom?

Need to **restrict the domain**: focus on the corporate domain, i.e., companies, people, products, ...

Editorial	Social Media	Job Postings
Patents & Trademarks	Gov Data	Company Websites
Credit Ratings	SEO Filings	● ● ●

Disk used on data nodes

513 TiB

elasticsearch

APACHE Spark

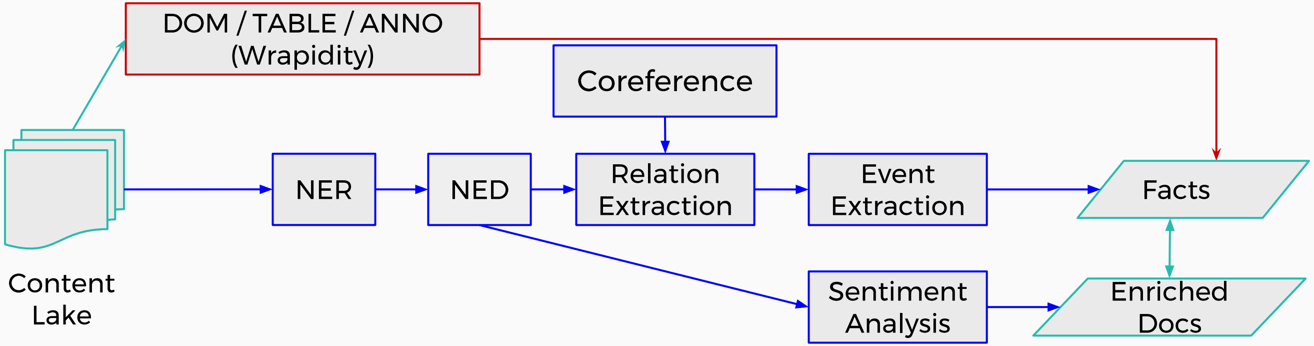
HIVE

WolframAlpha

EC2 cluster with 2.8k vCPU, 21TB RAM, 630TB SSD

Data and Content Lake

Linguistic enrichments to support semantic retrieval and fact extraction



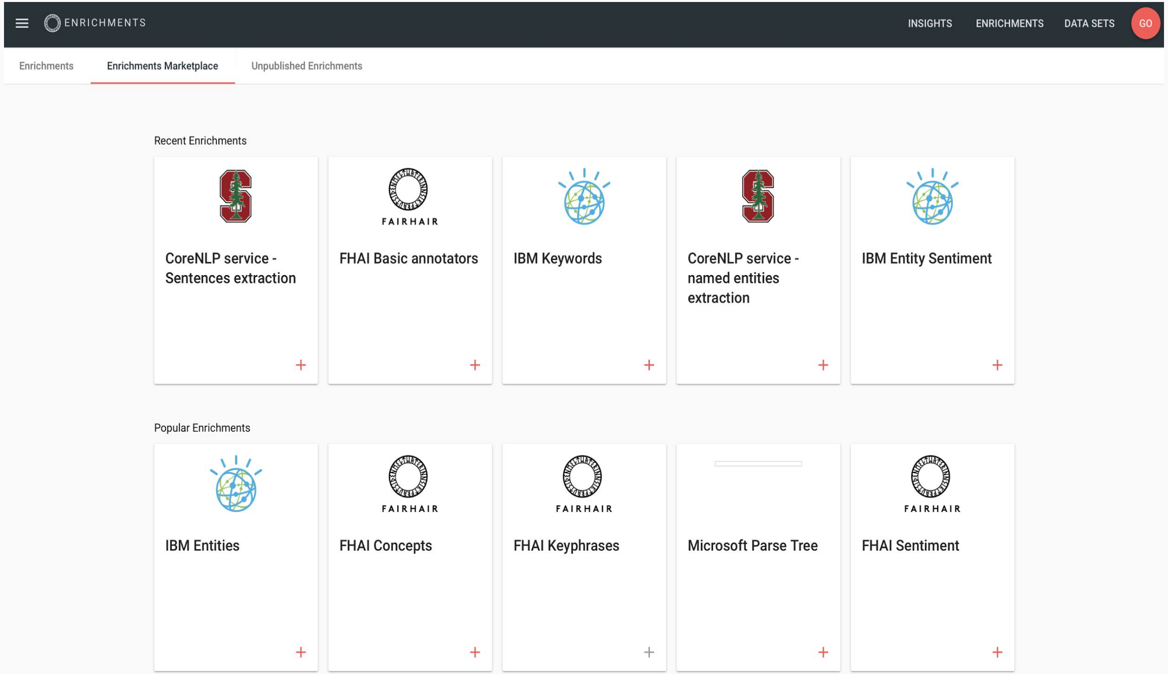
Summary:

- Scalable and distributed dynamic enrichment workflows
- CRFs for NER, PageRank (variant) for NED, CNNs and LSTMs for Relation/Event extraction, sentiment analysis
- TensorFlow and GPUs for training infrastructure

Enrichments

We can’t foresee all uses of our data: Developer APIs to Integrate and orchestrate third party tools.

Personalization is key in Data Science: A flexible data wrangling infrastructure is required.



Recent Enrichments

- CoreNLP service - Sentences extraction
- FHAI Basic annotators
- IBM Keywords
- CoreNLP service - named entities extraction
- IBM Entity Sentiment

Popular Enrichments

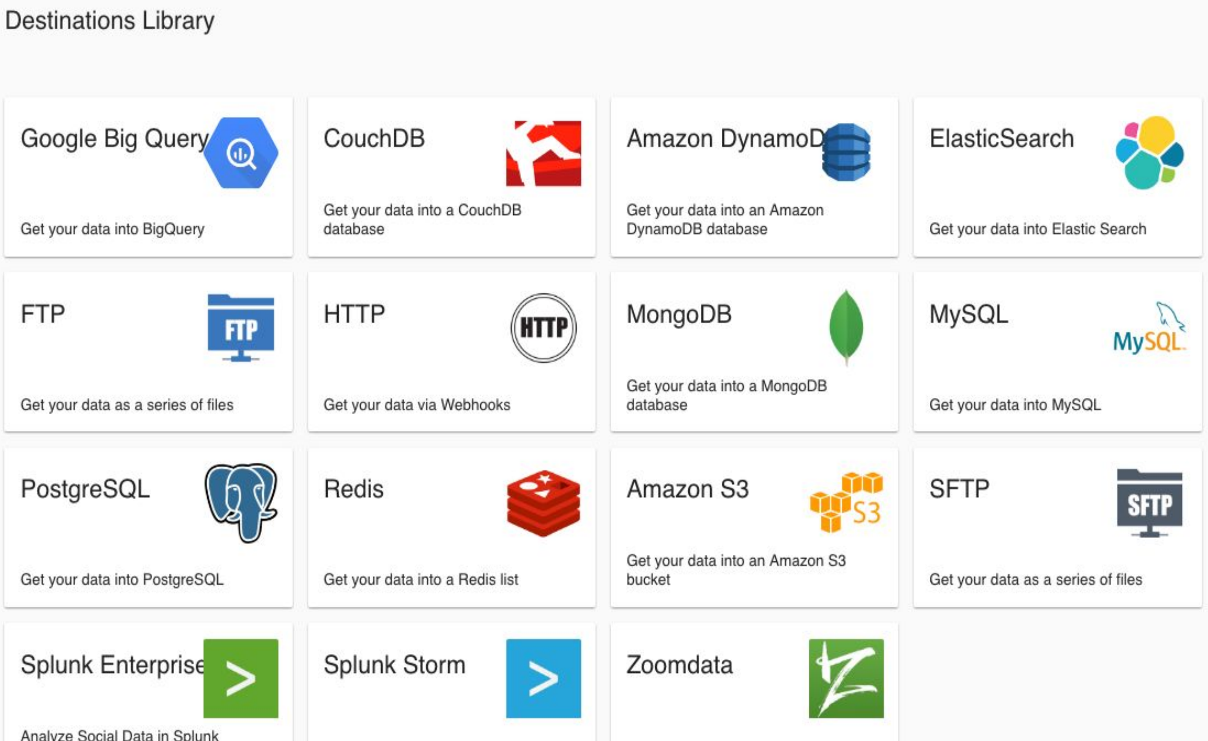
- IBM Entities
- FHAI Concepts
- FHAI Keyphrases
- Microsoft Parse Tree
- FHAI Sentiment

- Interoperate with state-of-the-art external enrichments
- Chain multiple external enrichments
- Train your own models!

Connectors to Internal Systems

Goal is to join Outside Insights with Internal Data and workflows

Data Ingestion & Insights Delivery by setting up simple schema mappers



Destinations Library

Google Big Query
Get your data into BigQuery

CouchDB
Get your data into a CouchDB database

Amazon DynamoDB
Get your data into an Amazon DynamoDB database

ElasticSearch
Get your data into Elastic Search

FTP
Get your data as a series of files

HTTP
Get your data via Webhooks

MongoDB
Get your data into a MongoDB database

MySQL
Get your data into MySQL

PostgreSQL
Get your data into PostgreSQL

Redis
Get your data into a Redis list

Amazon S3
Get your data into an Amazon S3 bucket

SFTP
Get your data as a series of files

Splunk Enterprise
Analyze Social Data in Splunk

Splunk Storm

Zoomdata

Knowledge Graph

Wait… did you say PageRank, triples? So do you have a (Knowledge) Graph?

Content:

- Companies
- Brands
- Products
- Key people
- Influencers

Goals:

- Relate facts
- Data mining
- Cognitive applications (higher-order reasoning)

Challenges:

- Data Cleaning
- Data deduplication / integration
- Truth Finding

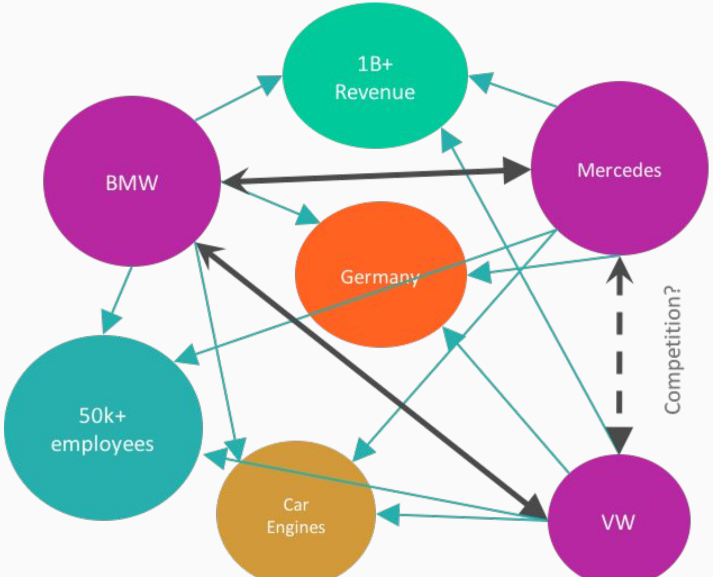
Cognitive Applications

Infer high-level insights from a set of extracted events/facts.

- Competitor
- Customer
- Investment
- Lawsuit/Litigation
- Partnership
- Supplier
- Acquisition
- Out/under performance
- Expanding Operations
- Compliance
- Funding Developments
- Leadership Changes
- New Offerings
- Bankruptcy
- Restructuring, Cost Cutting

Insight discovery:

- Rule/Graph mining (data cleaning)
  - GPAR (VLDB ‘15)
  - RUDIK (internal, paper submitted)
- Link prediction (data enrichment, fact checking)
  - Path Ranking Algorithms (PRA)
  - Probabilistic Soft Logic (PSL)
  - Snorkel (Stanford Collaboration)



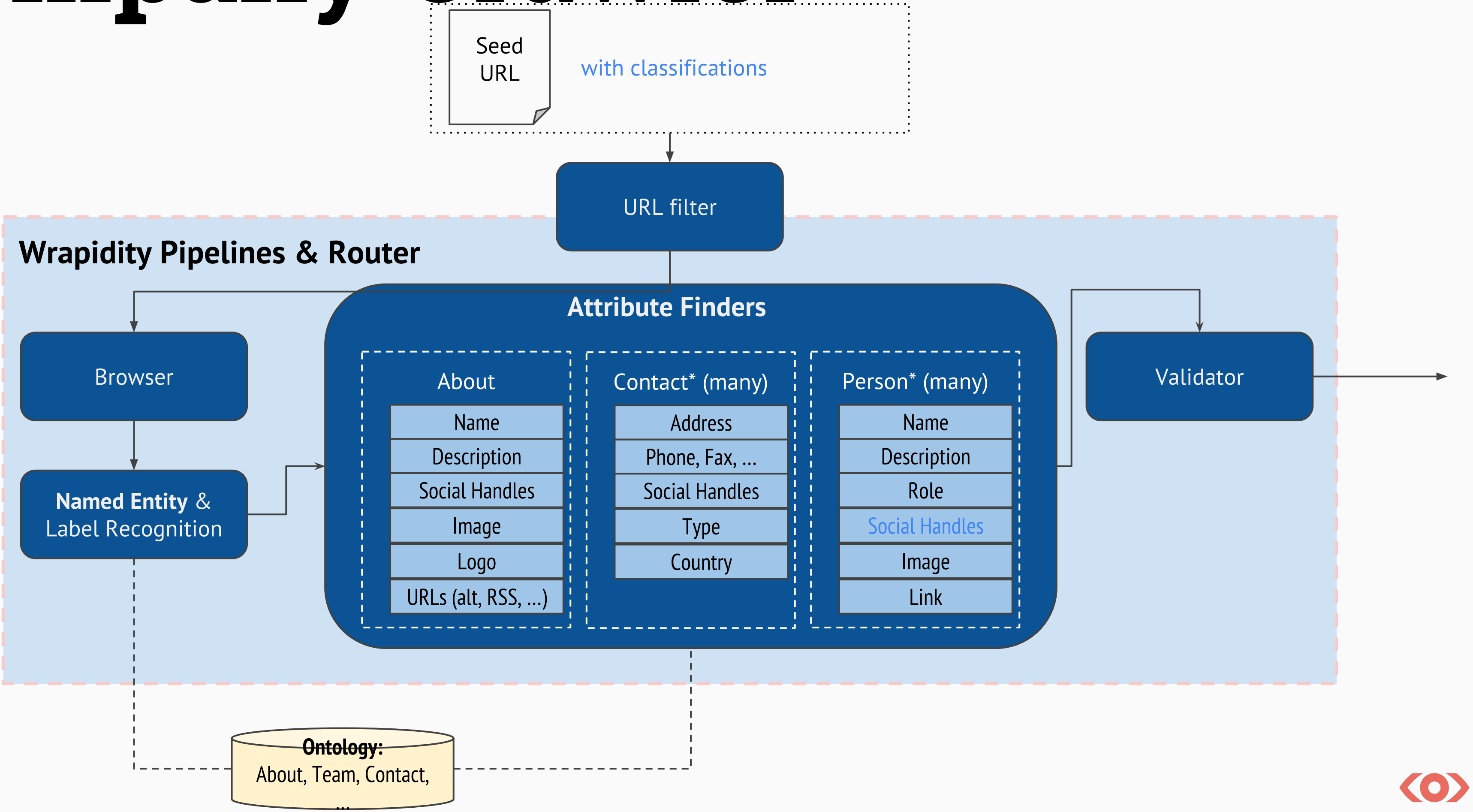
Questions

рахмат
danke
謝謝
ngiyabonga
spas
Баярлалаа
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
nanni
nandri
dhanyavad
thank you
gracias
хвала
asante
manana
obrigada
tenki
murakoze
bedankt
bayarlalaa
gràçie
hvala
mauruuru
köszönöm
aciū
dankon
akun
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
taiku
sobodi
dėkuji
sukriya
kop khun krap
sagolun
arigatô
takk
dakujem
trugarez
obrigado
mèsi
didi madloba
ありがとう
najis tuke
tanemirt
grazie
rahmet
terima kasih
kam sah hamnida
rahmat
dhanyavadagalu
diolch
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci

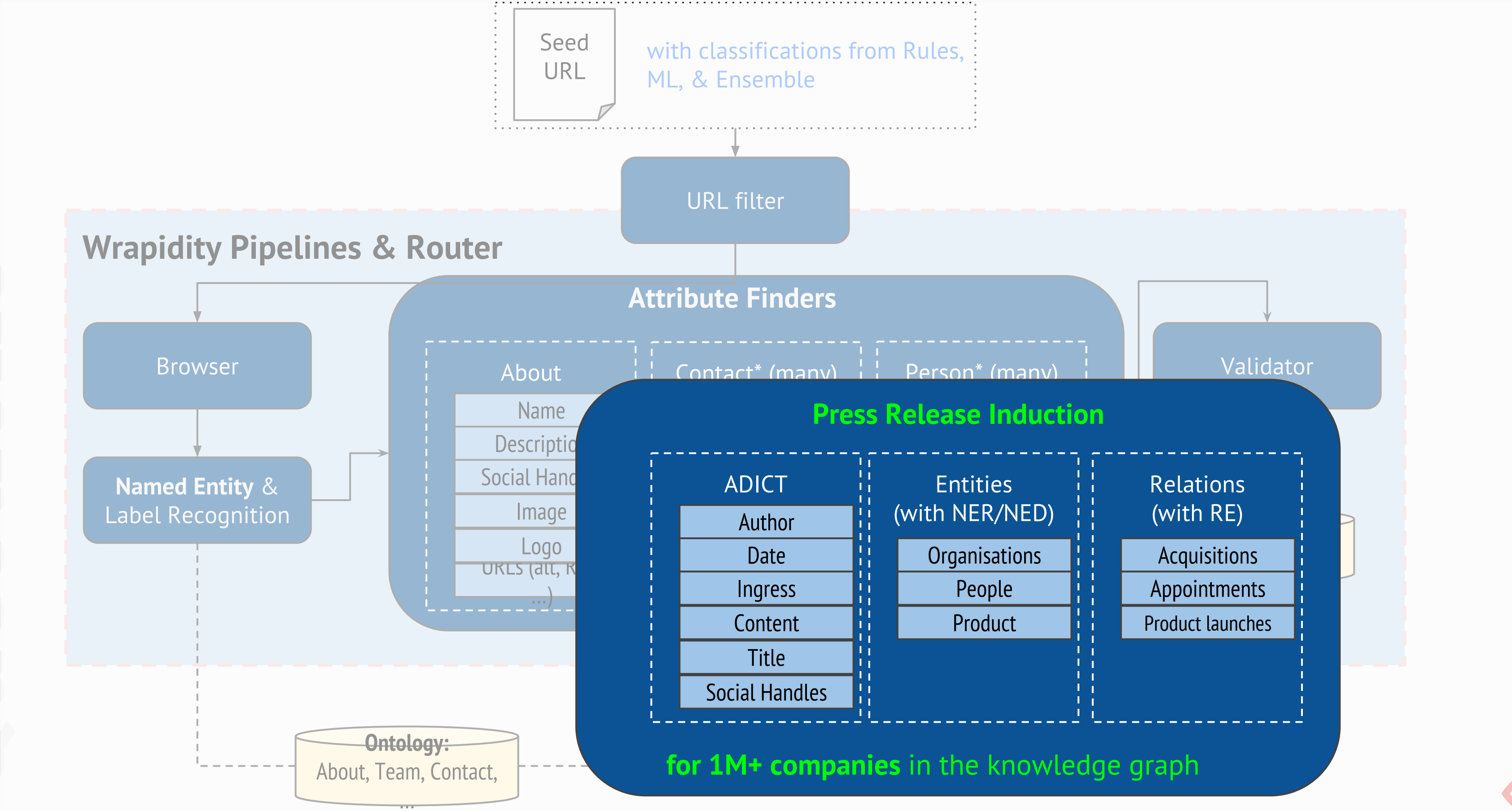
More?

acm
SIGMOD
POD
San Francisco
The VLDB Journal
The International Journal on
Very Large Data Bases
DocEng
2016
CIKM 2017
SINGAPORE
BIG ISLAND • BIG DATA • BIG ISLAND • BIG DATA •
VLDB 2015
www
2013
Rio de Janeiro, Brazil
EDBT
ICDT
Coling 2016
ISWC 2014
Riva del Garda - Trentino, Italy
VLDB 2013
VLDB
2014
WIMS'11
WORLD WIDE WEB
2012
LYON - FRANCE
ISWC 2012
The 11th International
Semantic Web Conference
Web Information System Engineering
WISE
2013
12th International Conference on Web Engineering
ICWE
July 23 - 27, 2012, Berlin, Germany
FoSSaCS
RuleML
Realize your Knowledge
DEXA 2012

Company Crawler



Company Crawler



Press Release Induction

ADICT	Entities (with NER/NED)	Relations (with RE)
Author	Organisations	Acquisitions
Date	People	Appointments
Ingress	Product	Product launches
Content		
Title		
Social Handles		

for 1M+ companies in the knowledge graph