Wrapping Millions of Documents Per Day

and How that’s Just the Beginning
Background: About me

Lecturer in Databases and Co-investigator VADA
Fellow of Oxford Martin School & OMI

Co-founder and CTO

Senior Research Manager
Meltwater
Mission
<table>
<thead>
<tr>
<th>NAME</th>
<th>STREET ADDRESS</th>
<th>LOCALITY</th>
<th>STATE</th>
<th>POSTCODE</th>
<th>PHONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEACH HAVEN II</td>
<td>600 N. Bay Avenue</td>
<td>Beach Haven</td>
<td>NJ</td>
<td>08008</td>
<td>609-492-6300</td>
</tr>
<tr>
<td>BORGATA</td>
<td>1 Borgata Way</td>
<td>Atlantic City</td>
<td>NJ</td>
<td>08401</td>
<td>609-317-8206</td>
</tr>
<tr>
<td>30TH STREET STATION</td>
<td>30TH STREET STATION</td>
<td>Philadelphia</td>
<td>PA</td>
<td>19104</td>
<td>215-222-2996</td>
</tr>
</tbody>
</table>
White House vows to fight media 'tooth and nail' over Trump coverage; says it presented 'alternative facts'

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" over the media, which he says attacks his administration and his family. The White House had previously dismissed reports of amidst low crowds during the inauguration as "fake news."
DEMO GODS
HELP US PLEASE
World class **crawling** platform to largely automate **outside content** collection
Crawling Space & Volume

Information Extraction

TXT (301M)

DOM (1280M)

110M

TBL (10M)

ANNO (145M)

1.1M

0.3M

13K

1.5M

1.7M

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

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
<table>
<thead>
<tr>
<th></th>
<th>effective</th>
<th>wrong or missing data</th>
<th>no data</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK real estate</td>
<td>91%</td>
<td>7%</td>
<td>2%</td>
</tr>
<tr>
<td>Oxford real estate</td>
<td>90%</td>
<td>6%</td>
<td>4%</td>
</tr>
<tr>
<td>ViNTs^10</td>
<td>4%</td>
<td>5%</td>
<td>91%</td>
</tr>
<tr>
<td>UK used cars</td>
<td>93%</td>
<td>4%</td>
<td>3%</td>
</tr>
<tr>
<td>US real estate</td>
<td>90%</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>
Vertical 3:

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

By Doina Chiacu and Jason Lange

**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” with the media. "We're going to be fighting that," he said.

The government later apologized for the confusion over crowd size.

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 just days after the US government released a report dismissing the scientific consensus on the dangers of climate change.

The UK report says that doubling the number of deaths when temperatures hit 30C could occur as soon as 2050, while the number of deaths when temperatures reach 40C could double by 2100.

A government spokesperson said the report was "outdated and based on out-of-date science and data."
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)
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

Jorn Lyseggen
Executive Officer and Founder

Martin Hernandez
Chief Financial Officer

Kaveh Rostamposhti
Executive Director, Americas

Pål Larsen
Director, EMEA

John Bok
Director, APAC

Niklas de Besche
Director, Products
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

<table>
<thead>
<tr>
<th></th>
<th>% of companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEO &amp; execs</td>
<td>85%</td>
</tr>
<tr>
<td>description</td>
<td>74%</td>
</tr>
<tr>
<td>location</td>
<td>82%</td>
</tr>
<tr>
<td>logo</td>
<td>89%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>present</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEO &amp; execs</td>
<td>58%</td>
</tr>
<tr>
<td>description</td>
<td>79%</td>
</tr>
<tr>
<td>location</td>
<td>69%</td>
</tr>
<tr>
<td>logo</td>
<td>74%</td>
</tr>
</tbody>
</table>
THIS SEEMS LIKE BULLSHIT
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
## Results

**25 Locations Found Near You**

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

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

---

### Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Street Address</th>
<th>Locality</th>
<th>State</th>
<th>Postcode</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEACH HAVEN II</td>
<td>600 N. Bay Avenue</td>
<td>Beach Haven</td>
<td>NJ</td>
<td>08008</td>
<td>609-492-6300</td>
</tr>
<tr>
<td>BORGATA</td>
<td>1 Borgata Way</td>
<td>Atlantic City</td>
<td>NJ</td>
<td>08401</td>
<td>609-317-8206</td>
</tr>
<tr>
<td>30TH STREET STATION</td>
<td>30TH STREET STATION</td>
<td>Philadelphia</td>
<td>PA</td>
<td>19104</td>
<td>215-222-2996</td>
</tr>
</tbody>
</table>
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

- But: they still didn’t believe
  - hired Accenture to assess quality of the data
  - result: over 97% precision
## Restaurant location: Independent Evaluation

<table>
<thead>
<tr>
<th>ASG</th>
<th>Present</th>
<th>and correct data &amp; extraction</th>
<th>but wrong extraction</th>
<th>but wrong data</th>
<th>but raters disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>city</td>
<td>100%</td>
<td>99.3%</td>
<td>0.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>street</td>
<td>100%</td>
<td>96.4%</td>
<td>1.7%</td>
<td>1.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>postcode</td>
<td>99%</td>
<td>97.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>2.8%</td>
</tr>
<tr>
<td>latlong</td>
<td>89%</td>
<td>99.7%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td>hours</td>
<td>47%</td>
<td>98.2%</td>
<td>0.0%</td>
<td>1.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>name</td>
<td>100%</td>
<td>99.5%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>phone</td>
<td>86%</td>
<td>98.7%</td>
<td>1.3%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>category</td>
<td>100%</td>
<td>98.9%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>98.5%</td>
<td>0.5%</td>
<td>0.4%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

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

~50k sources, 30 min recrawl interval

87%

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

90%

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

91%
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, …
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)

```html
<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

- 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F₁ for labeling</td>
<td>92%</td>
<td>96%</td>
<td>96%</td>
<td>98%</td>
</tr>
</tbody>
</table>
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

### Table 1: Address Information

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

### Diagram 1: Address Graph

- **Locality**: San Diego, CA 92101, Boise, ID 83702, Portland, OR 97209
- **Postcode**: 92101, 83702, 97209
- **State**: CA, ID, OR
- **Street Address**: 471 Horton Plaza, 103 North 10th Street, 301 NW 10th Avenue
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
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>ROSeAnn</strong> <em>(VLDB’14)</em></td>
<td><strong>Entity extraction</strong> from DOMs</td>
</tr>
<tr>
<td>2</td>
<td><strong>OPAL</strong> <em>(WWW’12, VLDBJ’13)</em></td>
<td><strong>Form understanding &amp; filling</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>AMBER</strong> <em>(ICWE’11)</em></td>
<td><strong>Record identification</strong> for lists</td>
</tr>
<tr>
<td>4</td>
<td><strong>OXPath</strong> <em>(VLDB’11, VLDBJ’13)</em></td>
<td><strong>Extraction language</strong></td>
</tr>
<tr>
<td>5</td>
<td><strong>Robust XPaths</strong> <em>(SIGMOD’16)</em></td>
<td><strong>Change-resistant</strong> wrappers</td>
</tr>
<tr>
<td>6</td>
<td><strong>Oxtractor</strong> <em>(Coling’16)</em></td>
<td><strong>Attribute extraction</strong></td>
</tr>
<tr>
<td>7</td>
<td><strong>WaDaR</strong> <em>(ICDE’16)</em></td>
<td><strong>Joint wrapper &amp; relation repair</strong></td>
</tr>
<tr>
<td>8</td>
<td><strong>VADA</strong> <em>(EDBT’16)</em></td>
<td><strong>Wrangling</strong> of extracted data (in progress)</td>
</tr>
</tbody>
</table>

**DIADEM** *(VLDB’14)*
World-first accurate, automatic **full-site extraction system**
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
Meltwater: In Numbers

- 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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Language</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notiradar</td>
<td>Mexico</td>
<td>Spanish</td>
<td>40700</td>
</tr>
<tr>
<td>福建东南新闻网</td>
<td>China</td>
<td>Chinese (simpl.)</td>
<td>23182</td>
</tr>
<tr>
<td>中工网</td>
<td>China</td>
<td>Chinese (simpl.)</td>
<td>20953</td>
</tr>
<tr>
<td>메일경제</td>
<td>Korea, Republic Of</td>
<td>Korean</td>
<td>20191</td>
</tr>
<tr>
<td>جستجوگ اخبار تی نیوز</td>
<td>Iran (islamic Republic Of)</td>
<td>Persian</td>
<td>18055</td>
</tr>
<tr>
<td>Match 生活網</td>
<td>Taiwan</td>
<td>Chinese (trad.)</td>
<td>17512</td>
</tr>
<tr>
<td>47NEWS</td>
<td>Japan</td>
<td>Japanese</td>
<td>16966</td>
</tr>
<tr>
<td>Nambia Press Agency</td>
<td>Namibia</td>
<td>English</td>
<td>9521</td>
</tr>
<tr>
<td>中金在线 - 外汇网</td>
<td>China</td>
<td>Chinese (simpl.)</td>
<td>7957</td>
</tr>
<tr>
<td>Onet.pl</td>
<td>Poland</td>
<td>English</td>
<td>7127</td>
</tr>
</tbody>
</table>
Meltwater: Exisiting 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: People & Community

5 Data Science Research offices

6 Data Science Hubs (co-working spaces)
- London
- San Francisco
- Singapore
- Sydney
- Berlin
- New York

University collaborations

- Stanford University
- Carnegie Mellon University
- 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**

<table>
<thead>
<tr>
<th>APIs &amp; Services</th>
<th>Search, Alerting, Analytics, Reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context Building</td>
<td>Knowledge graph</td>
</tr>
<tr>
<td>Enrichments &amp; Analysis</td>
<td>Data science platform</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>AI driven crawling</td>
</tr>
</tbody>
</table>

- Search, Alerting, Analytics, Reporting: Building blocks to leverage the platform
- Knowledge graph: Enable cognitive applications on top of our Data by connecting the dots
- Data science platform: Enrich, Analyze & Build Insights by interoperating with all major players
- 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, ...

---

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

**Disk used on data nodes**

**513 TiB**

**elasticsearch**

**Spark**

**WolframAlpha**

**Hive**
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
Fairhair

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.

- 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 BigQuery
  - 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
  - Get your data into a Splunk Storm

- Zoomdata
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
Company Crawler

Wrapidity Pipelines & Router

Attribute Finders

Browser

Named Entity & Label Recognition

Seed URL

with classifications from Rules, ML, & Ensemble

Attribute Finders

About

Contact* (many)

Person* (many)

Validator

URL filter

Browser

URL filter

Named Entity & Label Recognition

Ontology:

About, Team, Contact,

Press Release Induction

ADICT

Author

Date

Ingress

Content

Title

Social Handles

Entities (with NER/NED)

Organisations

People

Product

Relations (with RE)

Acquisitions

Appointments

Product launches

for 1M+ companies in the knowledge graph

Questions