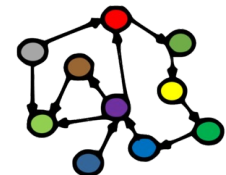


Welcome to INFO216:
Knowledge Graphs
Spring 2022

Andreas L Opdahl
<Andreas.Opdahl@uib.no>

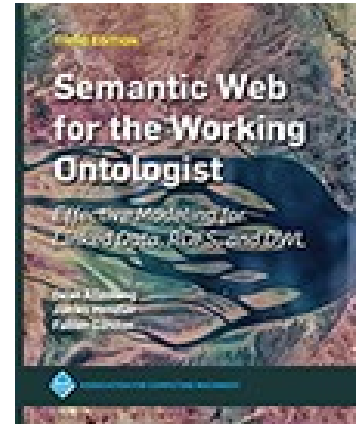
Session 6: Enterprise Knowledge Graphs

- Themes:
 - Open Knowledge Graphs (\leftarrow S05)
 - Linked Open Data resources / datasets
 - Wikidata, DBpedia, GDELT, EventKG
GeoNames, WordNet, BabelNet...
 - Enterprise Knowledge Graphs (EKGs) (\rightarrow S06)
 - Google's knowledge graph
 - Amazon's product graphs
 - the News Hunter infrastructure and architecture



Readings

- Sources (suggested):
 - Blumauer & Nagy (2020):
Knowledge Graph Cookbook – Recipes that Work (parts 2 and 4)
- Material at <http://wiki.uib.no/info216>:
 - *Introducing the Knowledge Graph: Things not Strings*, Amit Singhal, Google (2012).
 - *A reintroduction to our Knowledge Graph and knowledge panels*, Danny Sullivan, Google (2020).
 - *How Amazon’s Product Graph is helping customers find products more easily*, Arun Krishnan, Amazon (2018).
 - lecture slides

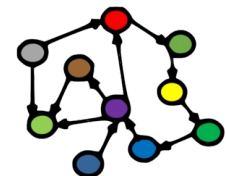


THE KNOWLEDGE GRAPH
COOKBOOK
RECIPES THAT WORK



ANDREAS BLUMAUER
AND HELMUT NAGY

1st edition, 2020

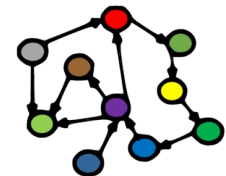


Is anyone really using
Knowledge Graphs?

Is anyone really using this?

Yes!

- But...



Tencent 腾讯

UniProt USGS

Google
Bing

Alibaba.com

Baidu 百度

PubMed

facebook

DEUTSCHE
NATIONAL
BIBLIOTHEK

ANTONI
VAN
LEEUVENHOEK
FOUNDATION



The
New York
Times

BBC

européana

NXP



REUTERS



National Library
of Sweden



EPA
United States
Environmental Protection
Agency

IOS
Press

Walmart

SIEMENS



Deloitte.



SPRINGER NATURE

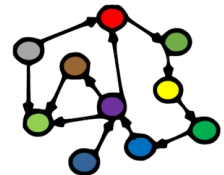
accenture

amazon.com

Is anyone really using this?

Yes!

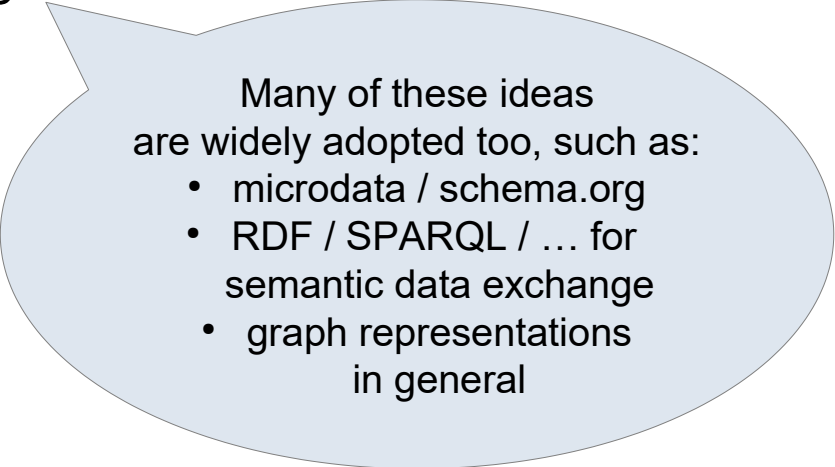
- **But...**
 - not quite as in the semantic web vision
 - not quite as in the LOD vision either
- Knowledge graphs are (additionally) becoming:
 - company internal
 - based on other technologies
 - such as general graph databases
 - not always linked to the LOD cloud



Is anyone really using this?

Yes!

- **But...**
 - not quite as in the semantic web vision
 - not quite as in the LOD vision either
- Knowledge graphs are (additionally) becoming:
 - company internal
 - based on other technologies
 - such as general graph databases
 - not always linked to the LOD cloud



Many of these ideas are widely adopted too, such as:

- microdata / schema.org
- RDF / SPARQL / ... for semantic data exchange
- graph representations in general

Is anyone really using this?

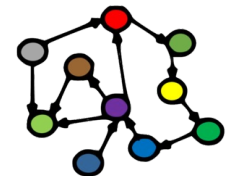
Yes!

- **But...**
 - not quite as in the semantic web vision
 - not quite as in the LOD vision either
- Knowledge graphs are (additionally) becoming:
 - company internal
 - based on other technologies
 - such as general graph databases
 - not always linked to the LOD cloud



Similar ideas,
adapted to new uses
and business contexts,
using a combination of
standard and other
technologies

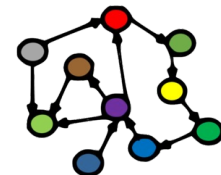
Google's Knowledge Graph



Google's Knowledge Graph

- Google Knowledge Graph (from 2012)
 - “Things, not Strings”
 - seeded from Freebase
 - facts from Wikipedia, Wikidata, CIA World Factbook
 - a growing number of other sources
 - enriched by natural-language parsing (NLP)
 - Google’s Knowledge Vault
 - used internally for many purposes
 - visible in Google Search results (Knowledge Panels)
 - question answering in Google Assistant / Home

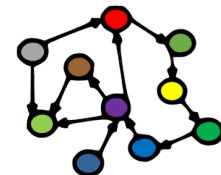
Caution: The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.



Google's Knowledge Graph

- Coverage:
 - claimed
 - 18 billion facts (18G, norsk: 18 milliarder)
about 570 million entities *soon after start*
 - 70 billion facts claimed in (2016)
 - 500 billion facts about five billion entities (2020)
 - ...perhaps 3 times the size of the LOD cloud
 - from English to multiple languages
- Critiques:
 - source attribution, incl. Wikipedia / Wikidata

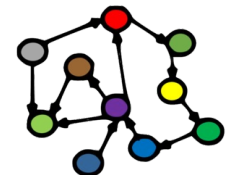
Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*



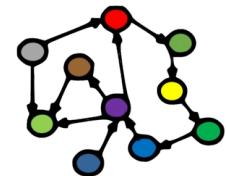
Google's Knowledge Vault Project

- Google Knowledge Vault
 - extends the Knowledge Graph
 - covers resources not from open semantic datasets
 - facts extracted from the whole web
 - NLP of text documents
 - HTML trees and tables
 - human annotated pages (e.g., schema.org)
 - probabilistic reasoning
 - graph-based priors
 - knowledge fusion

Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*

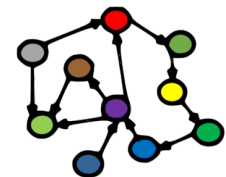


Amazon's Knowledge Graph



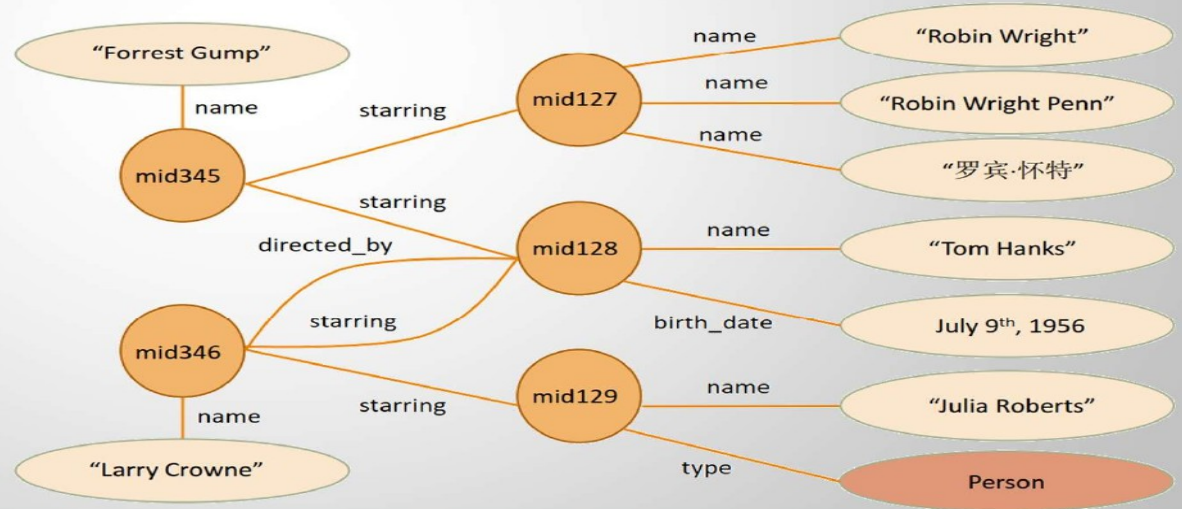
Amazon's ambition

- Let shoppers find the best products that fit their needs
 - allow greater variation in search terms
 - allow complex queries
- Structure all of the world's information as it relates to everything available on Amazon
- Describe every product on Amazon
 - concrete and abstract concepts
 - products and non-products
 - link different entities
- Enriched customer experience
 - visit Amazon to see what's new or interesting
 - discover ways to simplify and enrich their lives



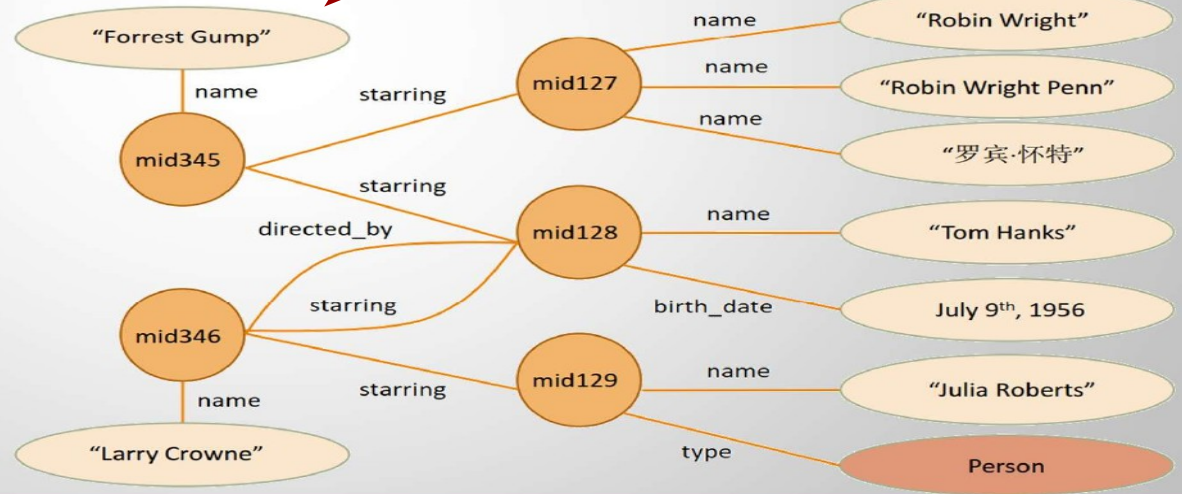
Amazon

Product Graph vs. Knowledge Graph



Amazon

Product Graph vs. Knowledge Graph



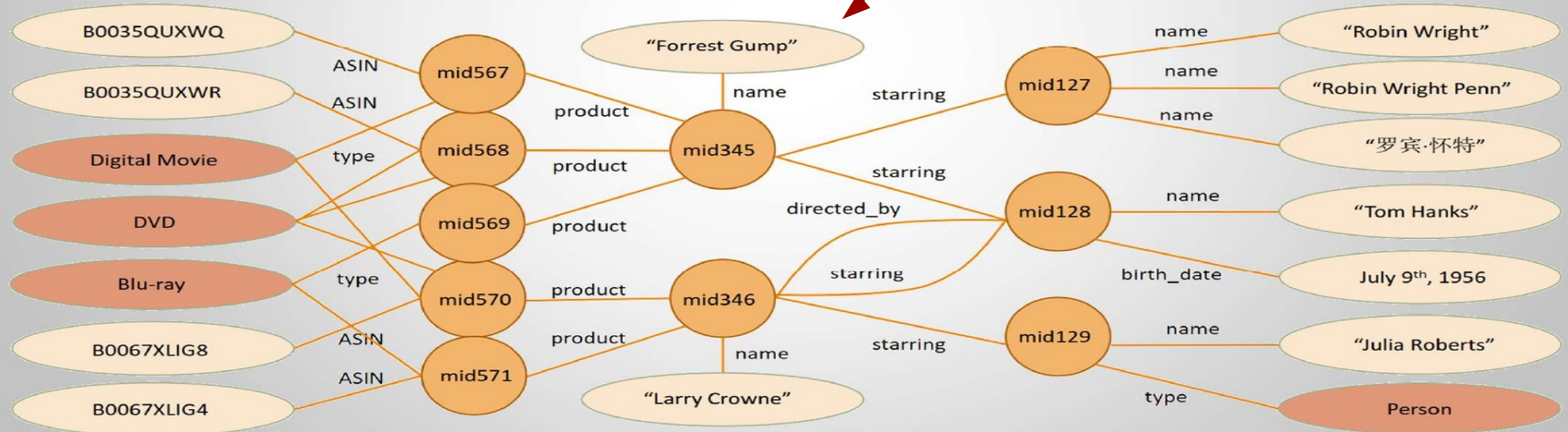
Amazon

Product details

Product

Product domain

Product Graph vs. Knowledge Graph



Ratings & reviews

Amazon

Delivery services

Customers

Product details

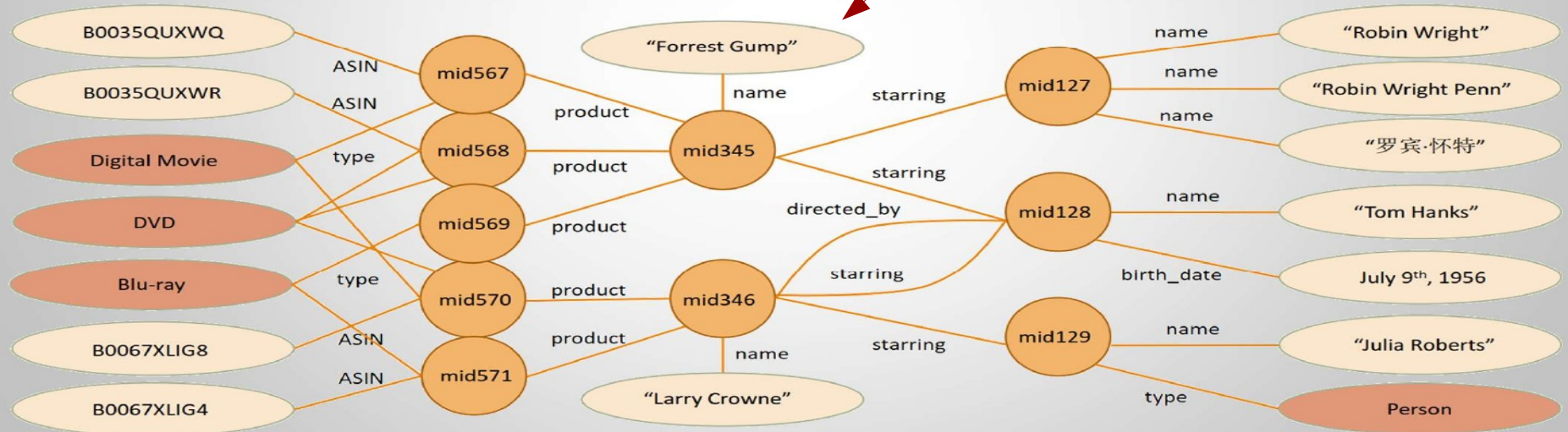
Suppliers

Support

Product

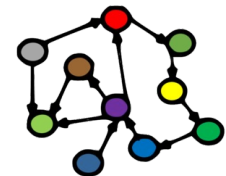
Product domain

Product Graph vs. Knowledge Graph



Challenges

- Ingest product-related information from Amazon's detail pages and from the Internet at large
 - product information is largely unstructured
 - trustworthiness of sources
- Machine learning techniques for
 - knowledge extraction, linkage and cleaning
 - distantly supervised learning
 - train on more structured subset of data
 - run on larger unstructured data space
 - open information extraction
 - graph mining techniques to identify interesting hidden patterns (buying product-X \rightarrow buying product-Y)

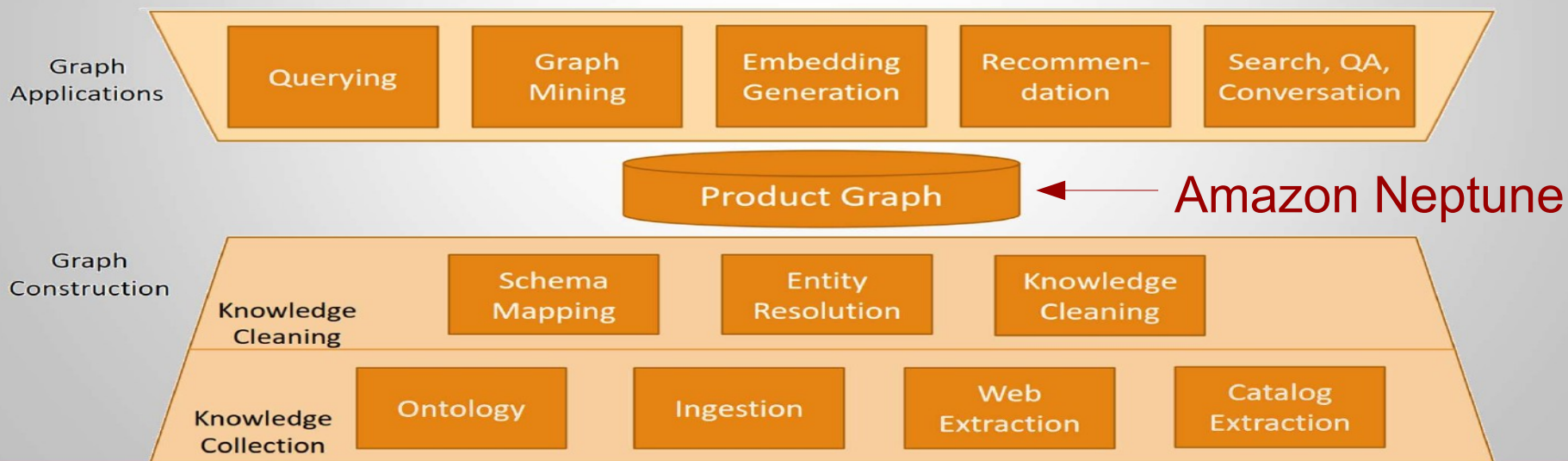


Amazon

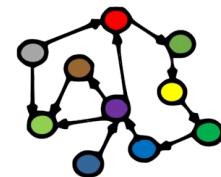
“We aim at building an authoritative knowledge graph for all products in the world”

Xin Luna Dong, Amazon,
at WSDM conf, Feb 2018

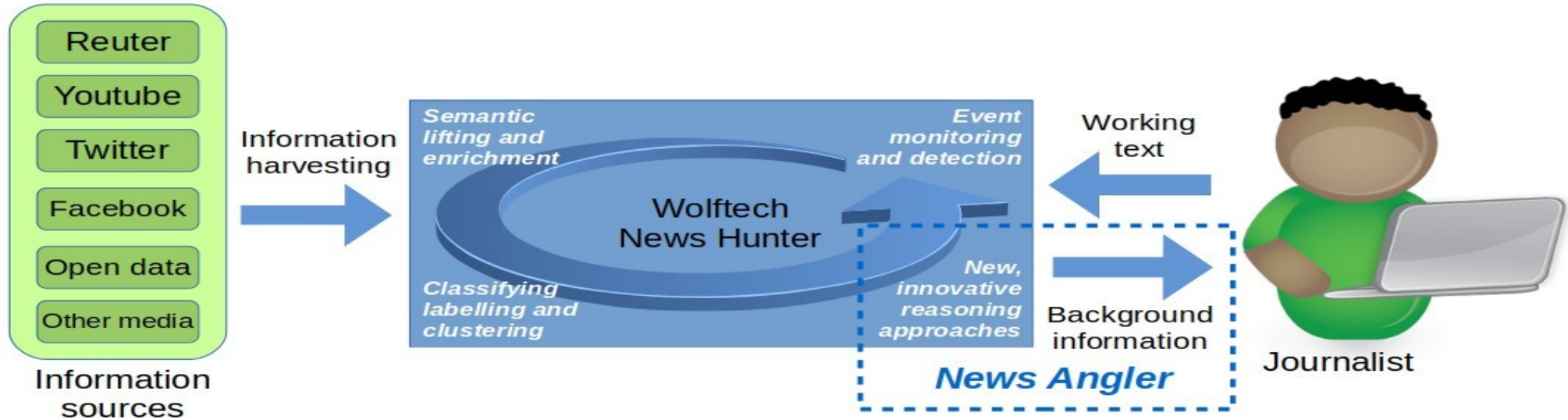
Architecture



The News Hunter Platform



Ongoing project: News Angler



“Wolftech News supports and improves the workflows in a newsroom through mobile solutions for field work that are integrated with central systems for news monitoring, resource management, news editing, and multi-platform publishing”

- 1) Harvesting and analysing messages
 - 2) Growing a semantic news graph
 - concepts, named entities, context...
 - 3) Analysing working texts (stories)
 - 4) Identifying background information
 - 5) Prioritising and preparing
 - 6) Journalistic and editorial preferences
- Research:* graph, searches, preparation, preferences, language, scaling

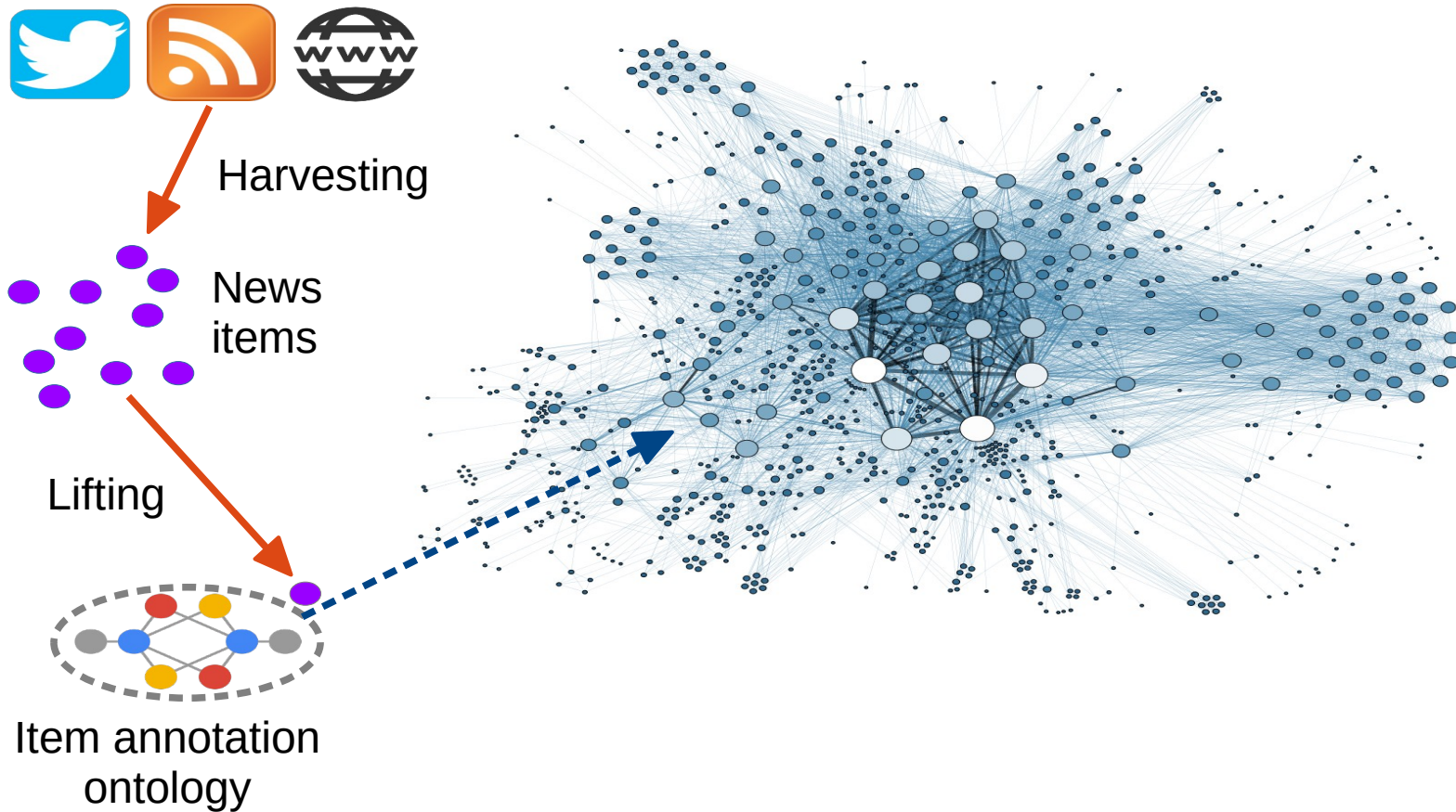
A single central news graph



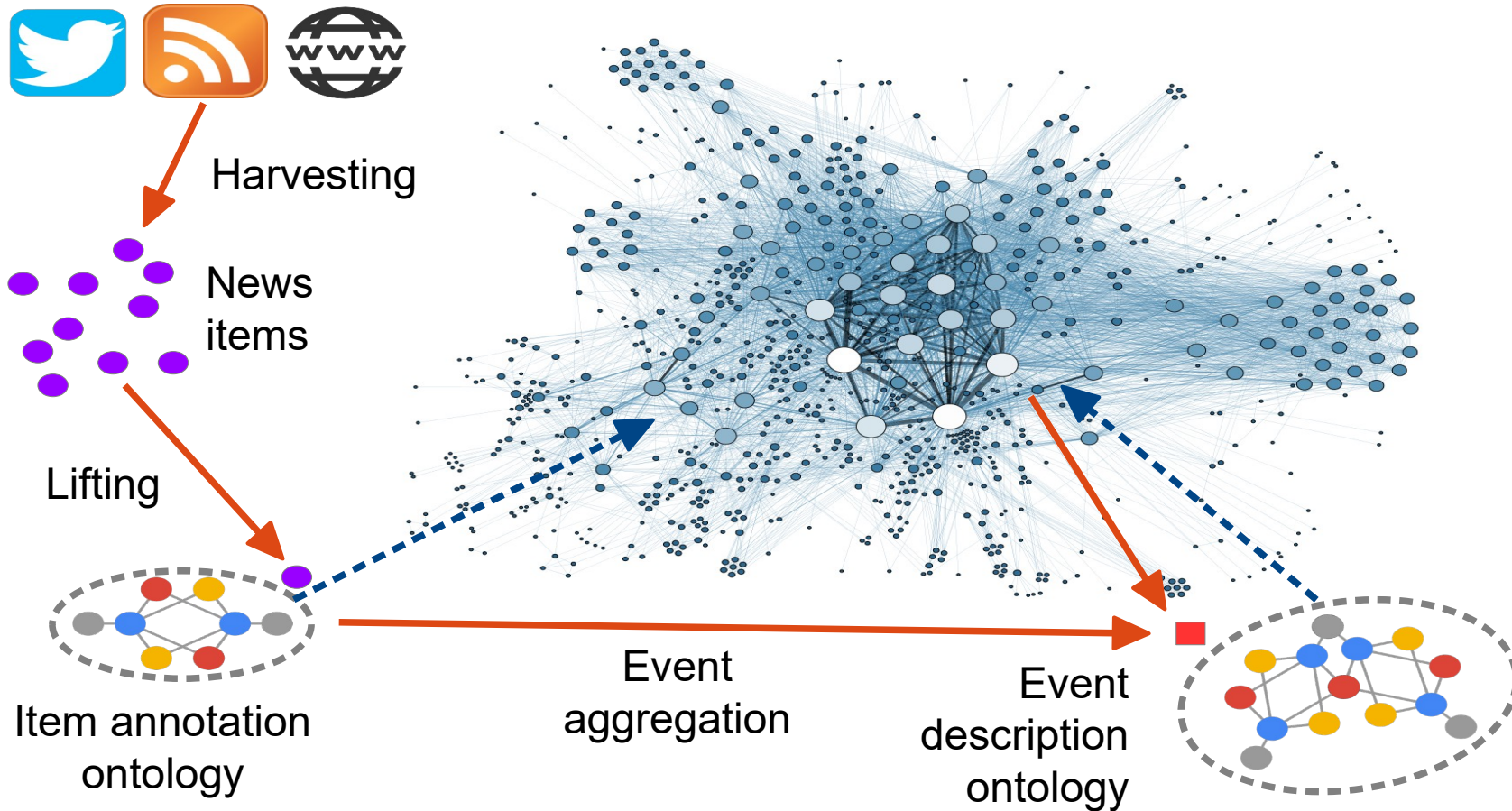
A single central news graph



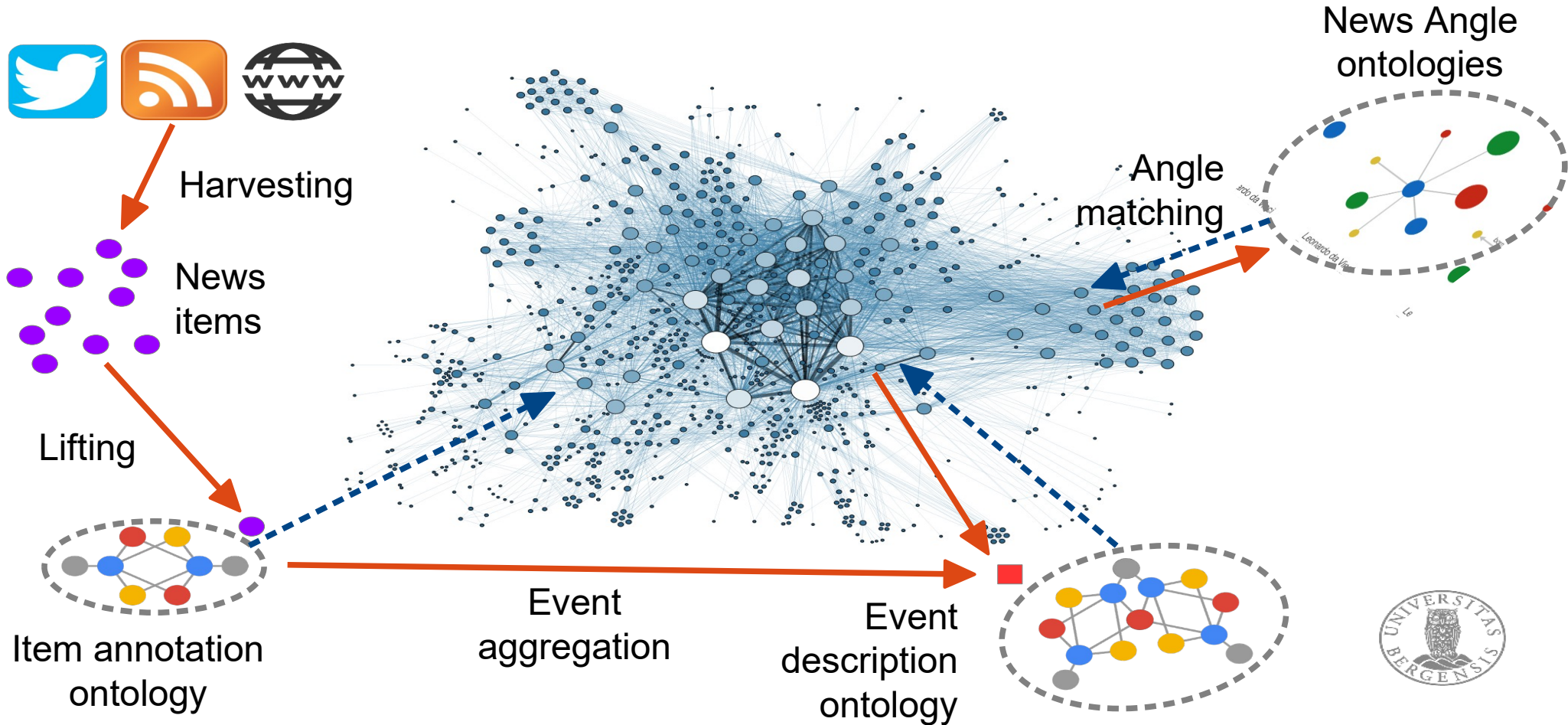
A single central news graph



A single central news graph



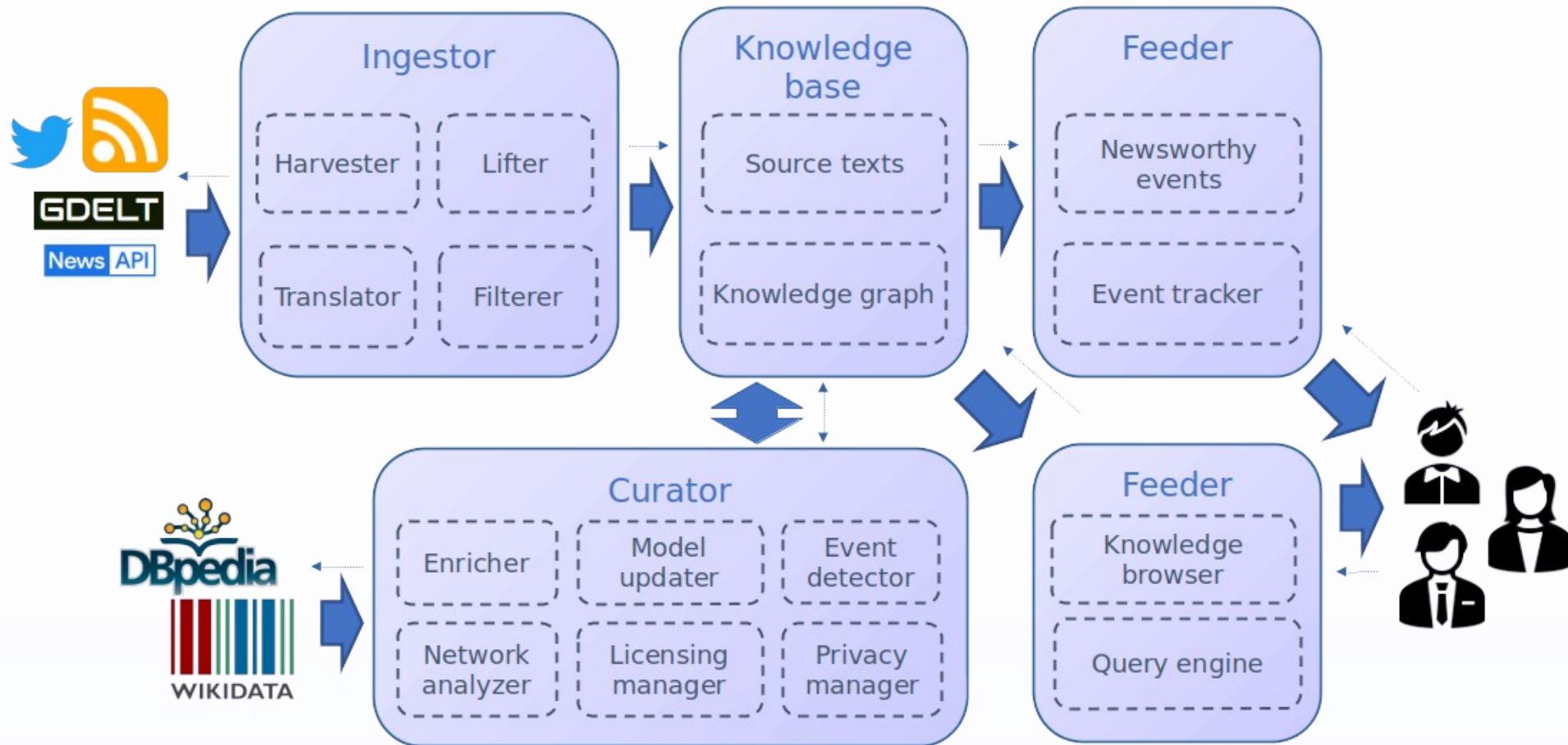
A single central news graph





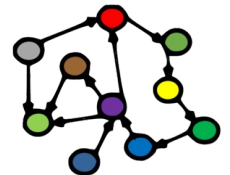
The News Hunter architecture

Harvesting news-related information from social media and other sources; analysing, organising, enriching and presenting news-related information to journalists. Implemented state-of-the-art big data and distributed technologies.



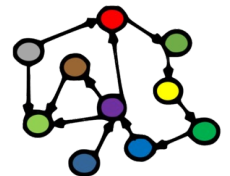
Services

- Written in Python 3.8-3.9
- All services are deployed in docker containers
- FastAPI as the main python library for writing APIs



Services - harvesters

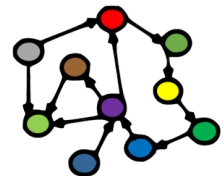
- Twitter harvester: connects to the Twitter API to read streams of tweets from news organizations accounts
- RSS harvester: downloads RSS feeds from news organisations
- GDELT harvester: gets the events and GKG datasets from GDELT projects
- NewsAPI harvester: use NewsAPI.org API to get real-time feeds of news from thousands of news outlets



Services - lifters

Lifters for news and GDELT that use NER to represent the information into knowledge graphs

- DbpediaSpotlight NEL: using DBpediaSpotlight for named entity linking
- SpaCy NEL: using SpaCy for named entity linking
- Kolitsas NEL: using Kolitsas algorithm for named entity linking





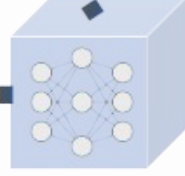
The News Hunter infrastructure



Service nodes

Web scraping, API, user interfaces, semantic lifting processes

- Light-to-medium processing
- Python, REST API, ...



Computation-intensive nodes

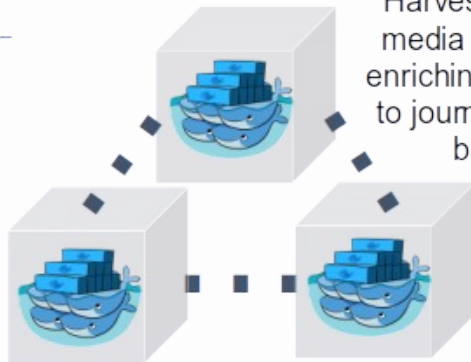
Complex AI services and training processes.

- CPU, RAM, GPU intensive
- *Python, spaCy, ...*

Management nodes

Service orchestration and monitoring

- Lighter processing
- Docker Swarm

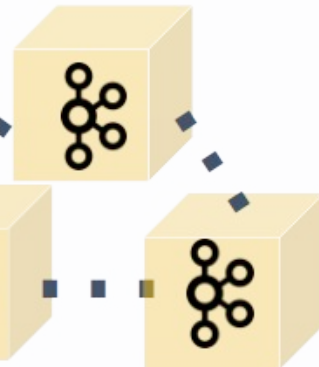


Harvesting news-related information from social media and other sources; analysing, organising, enriching and presenting news-related information to journalists. Implemented using state-of-the-art big data and distributed technologies.

Message queue nodes

Message exchange, queueing (TBD)

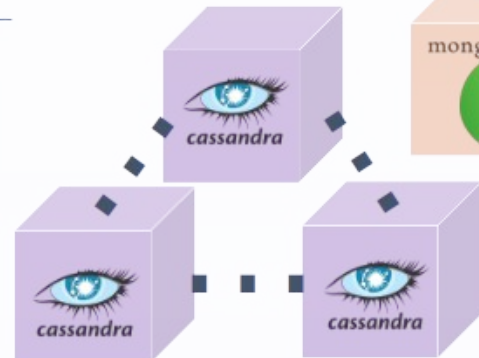
- Lighter processing
- Kafka



Raw data nodes

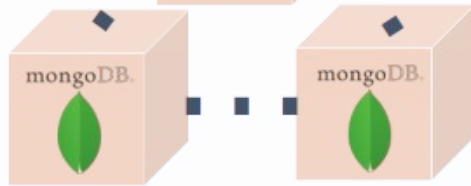
Distributed storage for raw data files (textual, multimedia)

- Disk intensive
- *Cassandra, ...*



Configuration nodes

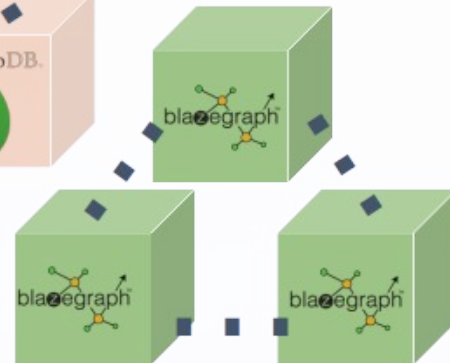
- Lighter processing
- *MongoDB, files*



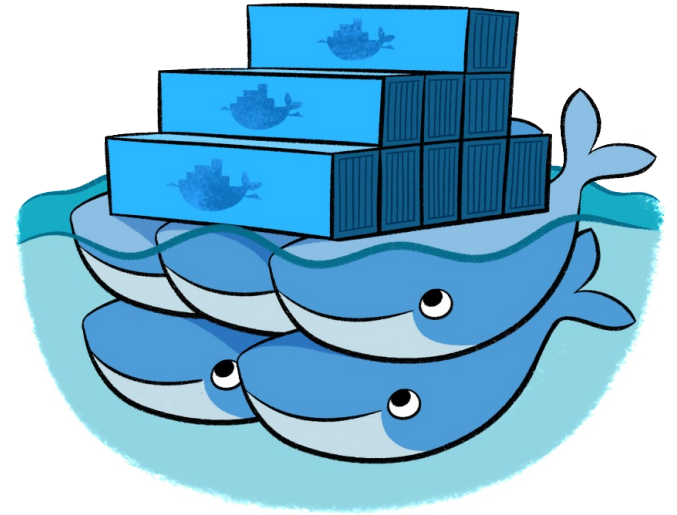
Knowledge graph nodes

News semantic representation storage.

- Disk, CPU and RAM intensive
- *Blazegraph*



Cloud infrastructure deployment tools



Technologies

- Docker Swarm
- Kafka (as pub/sub message queue to communicate between all services in the platform)
- Zookeeper
- Cassandra (storing raw data in a distributed cluster)
- Blazegraph (Knowledge graph with news and events representations)
- MongoDB (configuration and metadata)

* All of them have been deployed using Docker containers



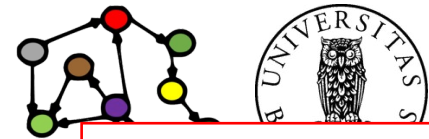
News Hunter Platform:

- 38 vCPUs
- 152GB RAM
- 20TB Disk
- 17 Instances

+

**1 Launcher instance for deploying
the cloud infrastructure:**

- 1 vCPU
- 4 GB RAM



1 vCPU = 0.5CPU

Next week:
Rules (RDFS)