Welcome to INFO216:
Knowledge Graphs

Spring 2022

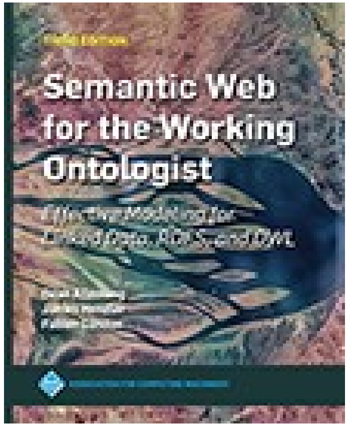
Andreas L Opdahl
<Andreas.Opdahl@uib.no>

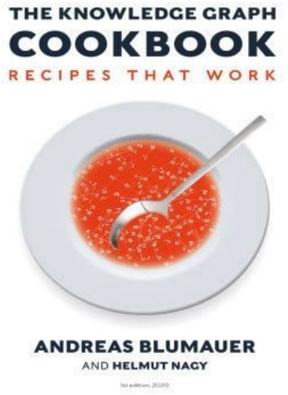
Session 6: Enterprise Knowledge Graphs

- Themes:
  - Open Knowledge Graphs *(← S05)*
    - Linked Open Data resources / datasets
    - Wikidata, DBpedia, GDELT, EventKG GeoNames, WordNet, BabelNet...
  - Enterprise Knowledge Graphs (EKGs) (→ *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





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
ANTONI VAN LEEUWENHOEK FOUNDATION
MAASTRO
DEUTSCHE NATIONAL BIBLIOTHEK
europeana
BBC
The New York Times
GETTY
NXP
REUTERS
KUNGL. BIBLIOTEKET
National Library of Sweden
RENAULT
EPA United States Environmental Protection Agency
IOS Press
European Commission
Walmart
SIEMENS
BEST BUY
Deloitte.
NON SOLUS
ELSEVIER
SPRINGER NATURE
amazon.com
accenture

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

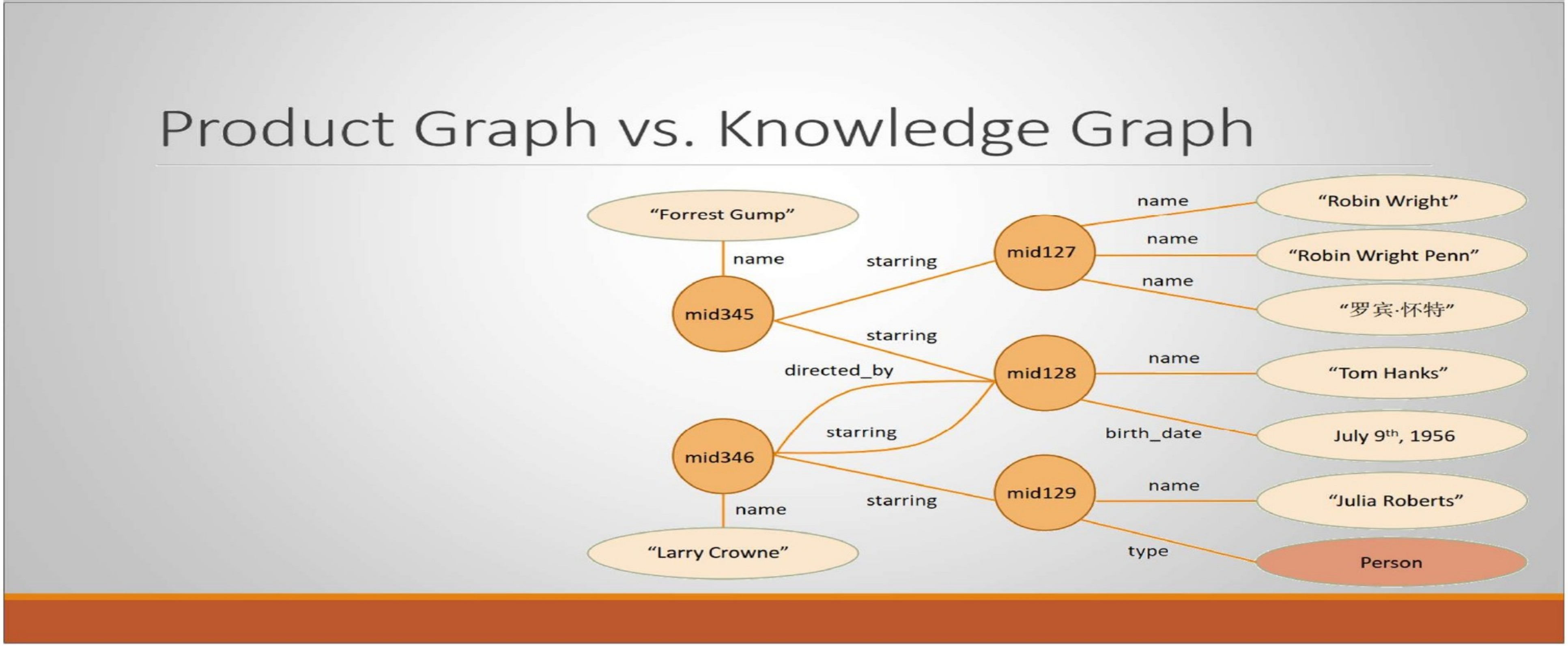
Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*

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

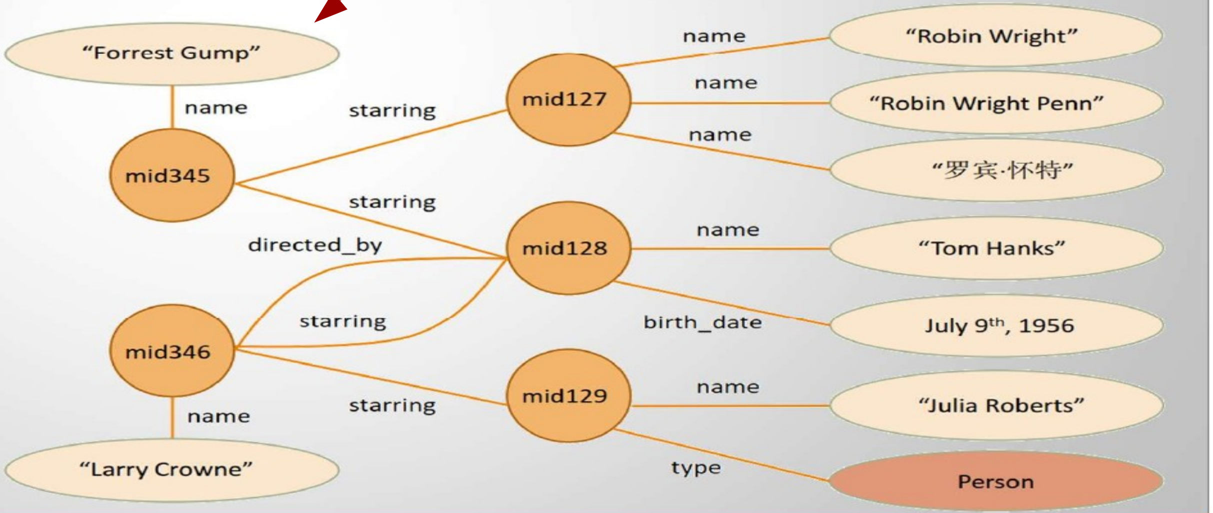
Frank van Harmelen (2018): Keynote at CAiSE'18

Amazon

Product

Product domain

Product Graph vs. Knowledge Graph



Frank van Harmelen (2018): Keynote at CAiSE'18

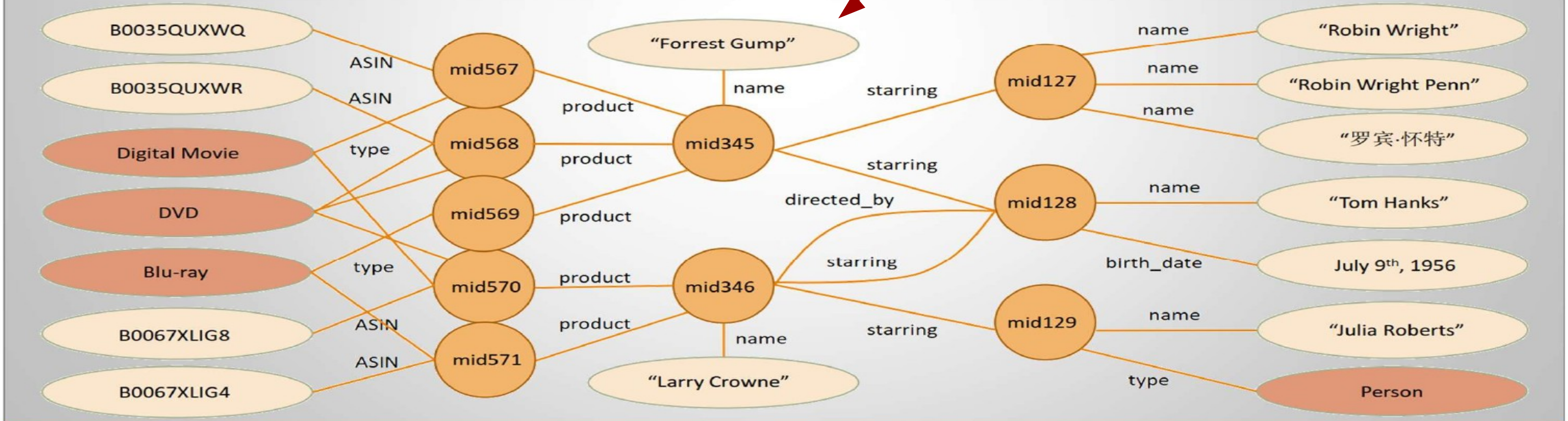
Amazon

Product details

Product

Product domain

Product Graph vs. Knowledge Graph



Frank van Harmelen (2018): Keynote at CAiSE’18

Ratings & reviews

Customers

Amazon

Delivery services

Product details

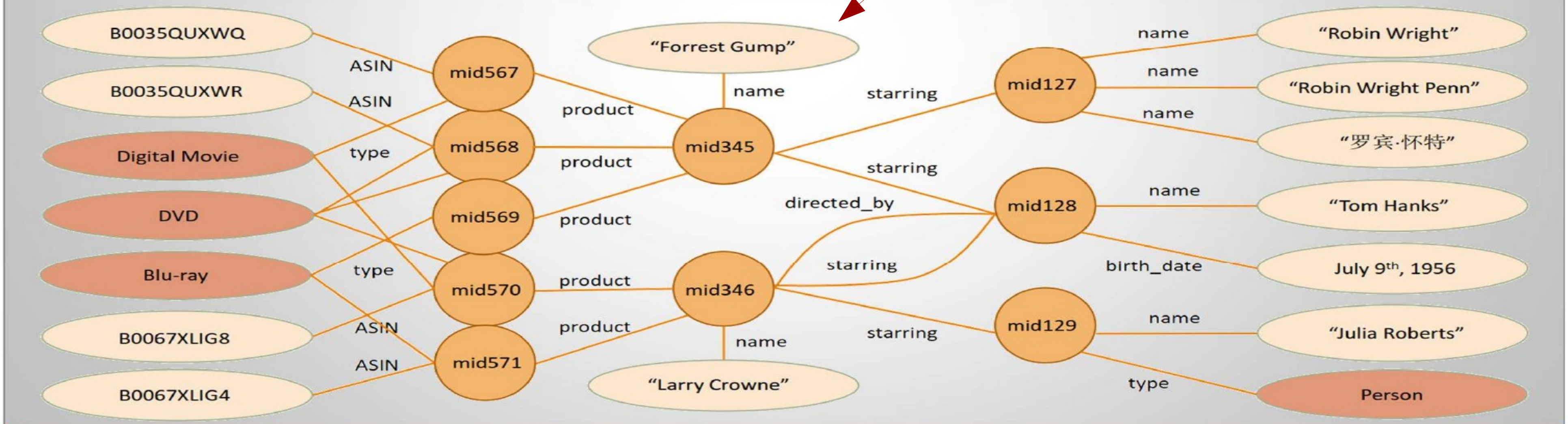
Support

Product

Product domain

Suppliers

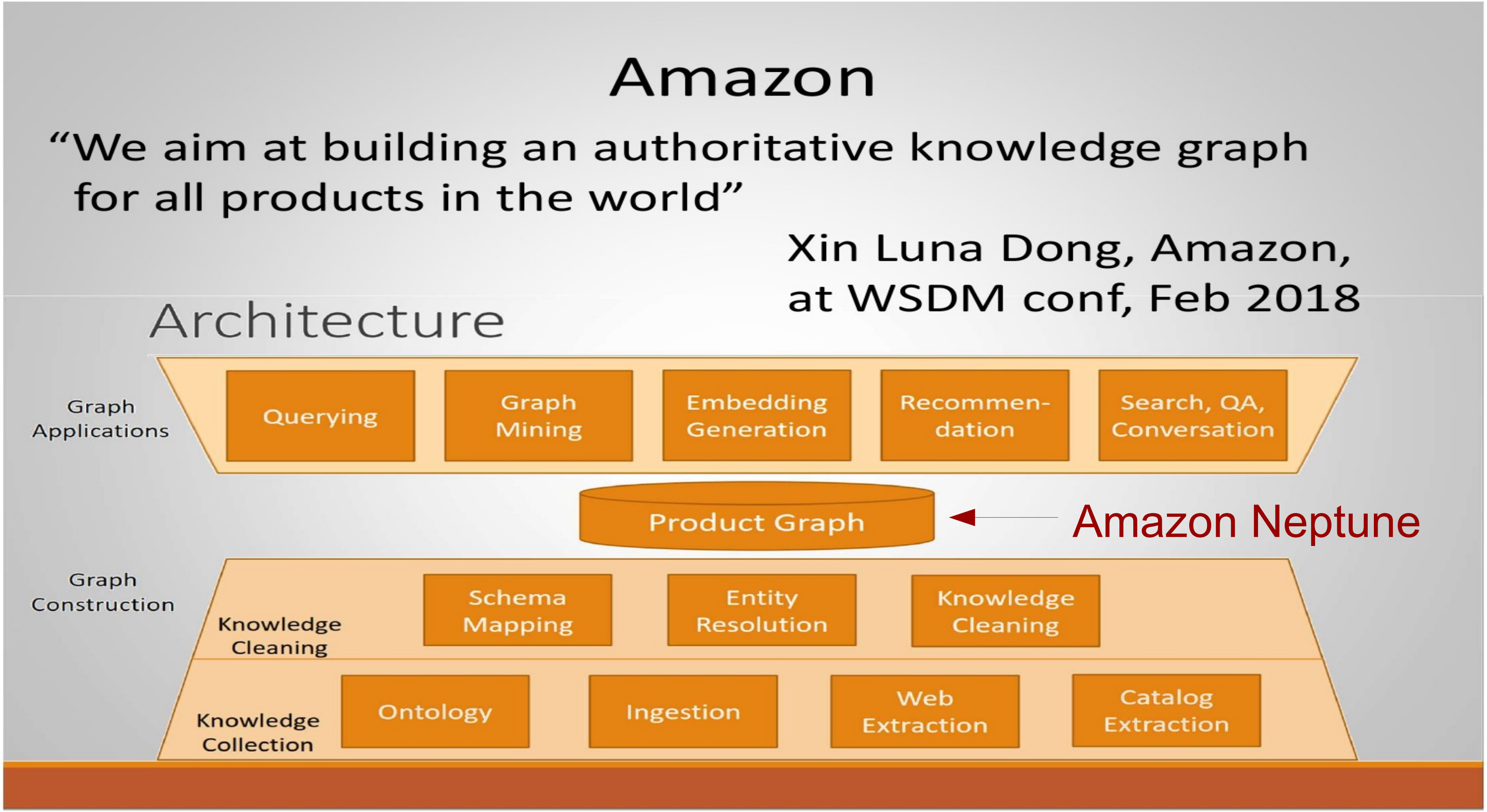
Product Graph vs. Knowledge Graph



Frank van Harmelen (2018): Keynote at CAiSE'18

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 → 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

Graph Applications

Querying

Graph Mining

Embedding Generation

Recommen-dation

Search, QA, Conversation

Product Graph

Amazon Neptune

Graph Construction

Knowledge Cleaning

Schema Mapping

Entity Resolution

Knowledge Cleaning

Knowledge Collection

Ontology

Ingestion

Web Extraction

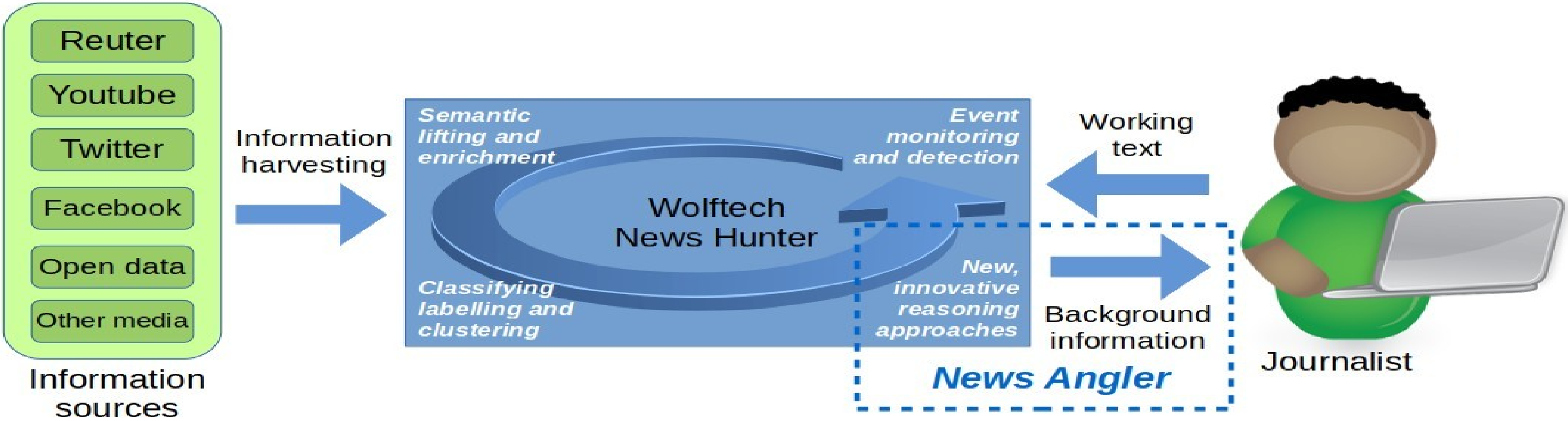
Catalog Extraction

Frank van Harmelen (2018): Keynote at CAiSE’18

The News Hunter Platform

http://newsangler.uib.no

Ongoing project: News Angler



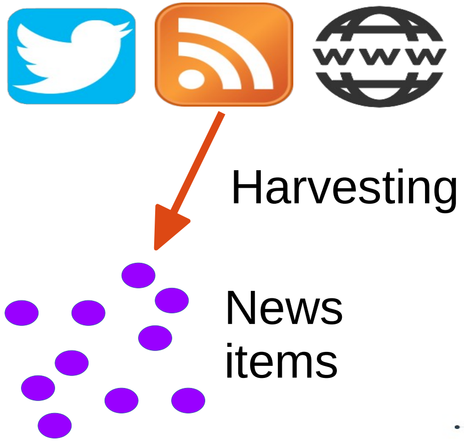
"Wolftech News supports and improves the workflows in a newsroom through mobile solutions for field work that are integrated with central systems for news monitoring, resource management, news editing, and multi-platform publishing"

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

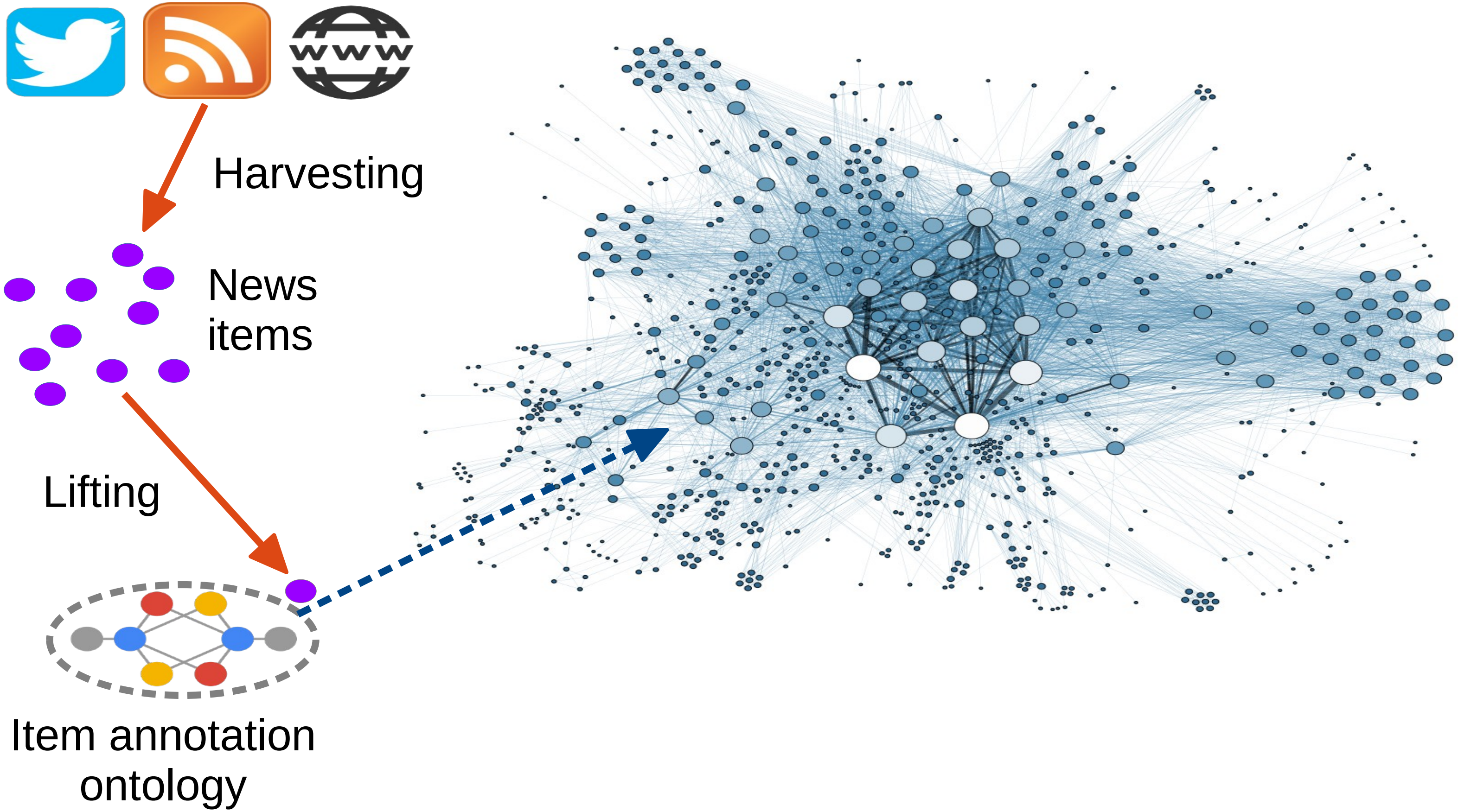
A single central news graph



Harvesting

News items

A single central news graph



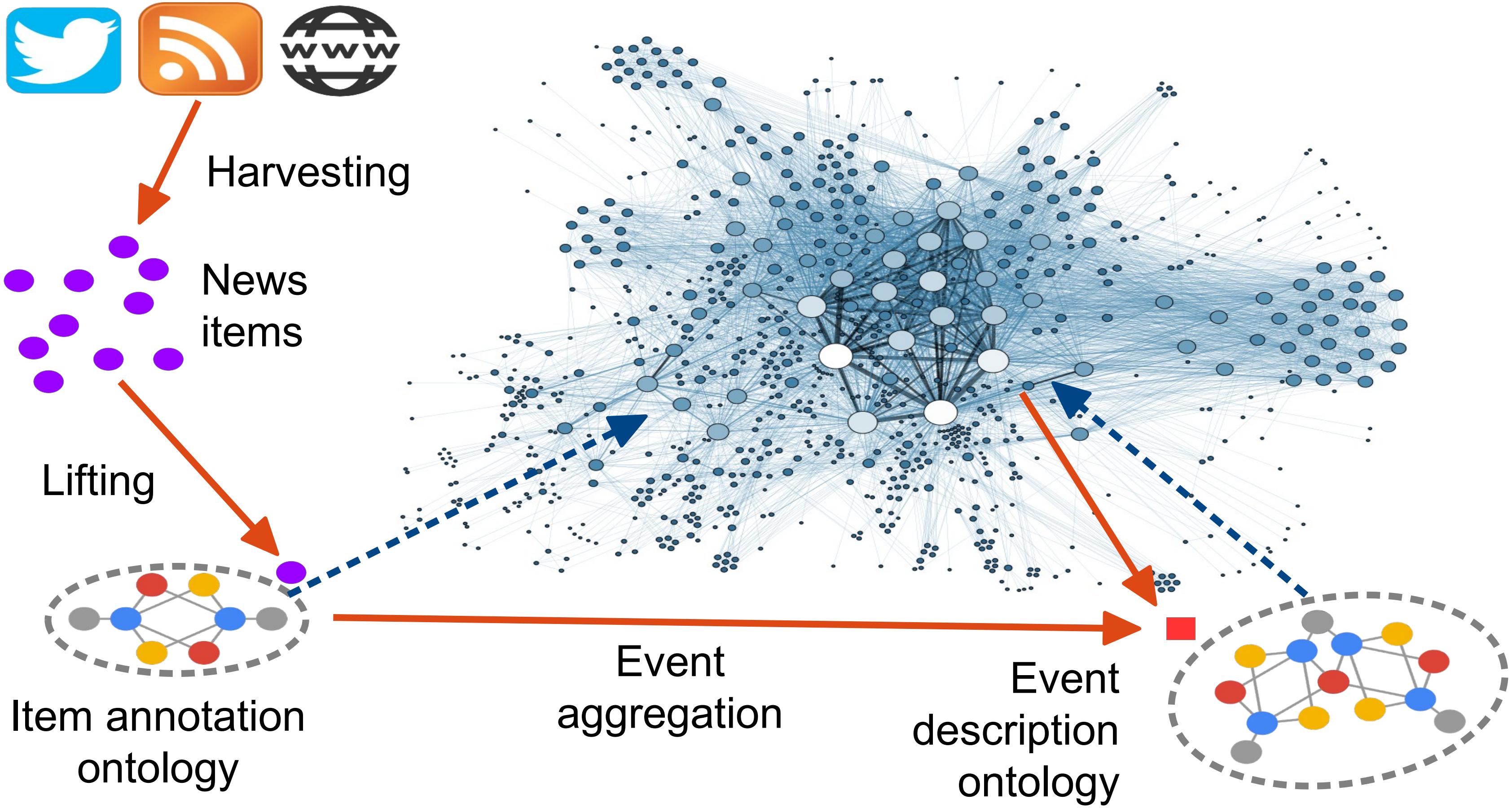
Harvesting

News items

Lifting

Item annotation ontology

A single central news graph



Harvesting

News items

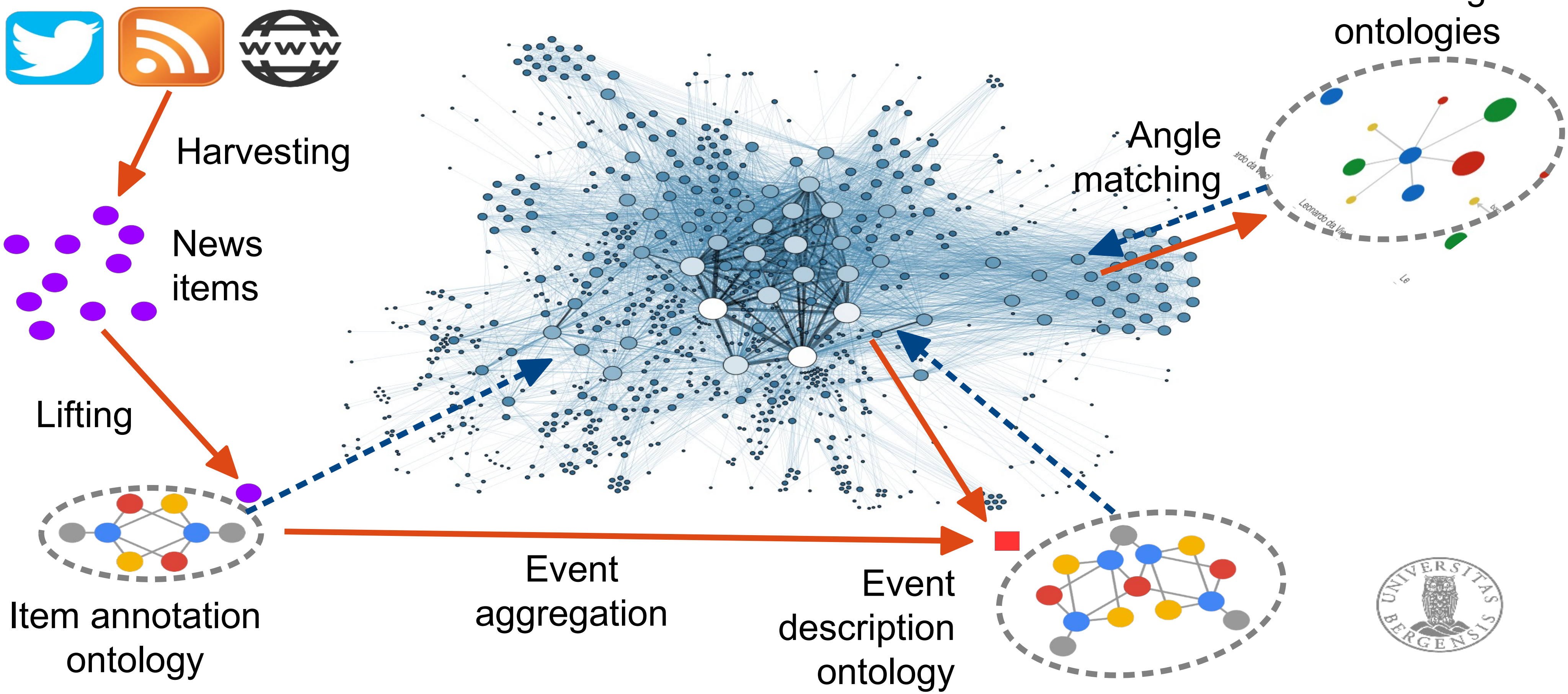
Lifting

Item annotation ontology

Event aggregation

Event description ontology

A single central news graph



Harvesting

News items

Lifting

Item annotation ontology

Event aggregation

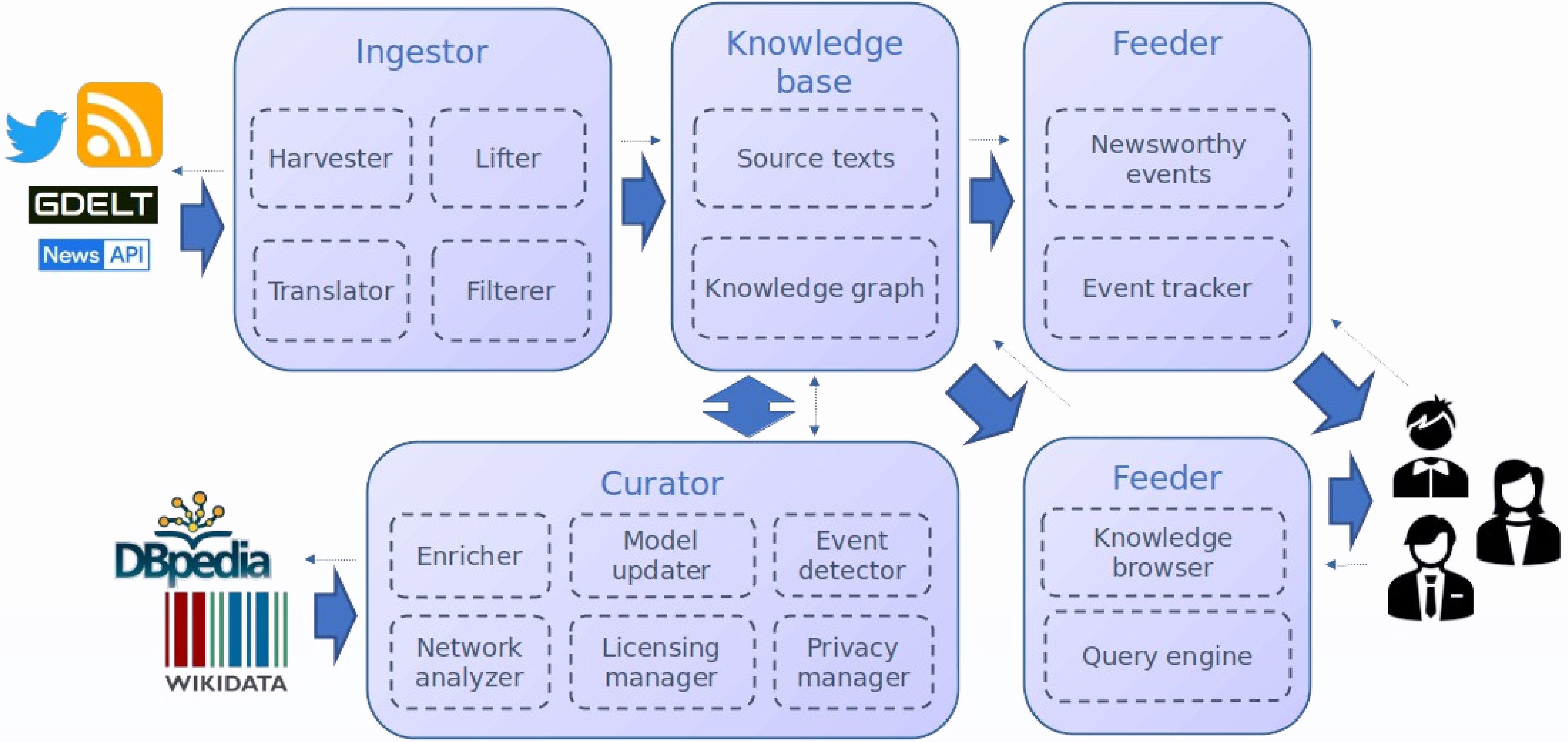
Event description ontology

Angle matching

News Angle ontologies

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.



Ingestor

- Harvester
- Lifter
- Translator
- Filterer

Knowledge base

- Source texts
- Knowledge graph

Feeder

- Newsworthy events
- Event tracker

Curator

- Enricher
- Model updater
- Event detector
- Network analyzer
- Licensing manager
- Privacy manager

Feeder

- Knowledge browser
- Query engine

M. Gallofré Ocaña & A.L. Opdahl (2021)

Services

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

Slide by Marc Gallofré Ocaña

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

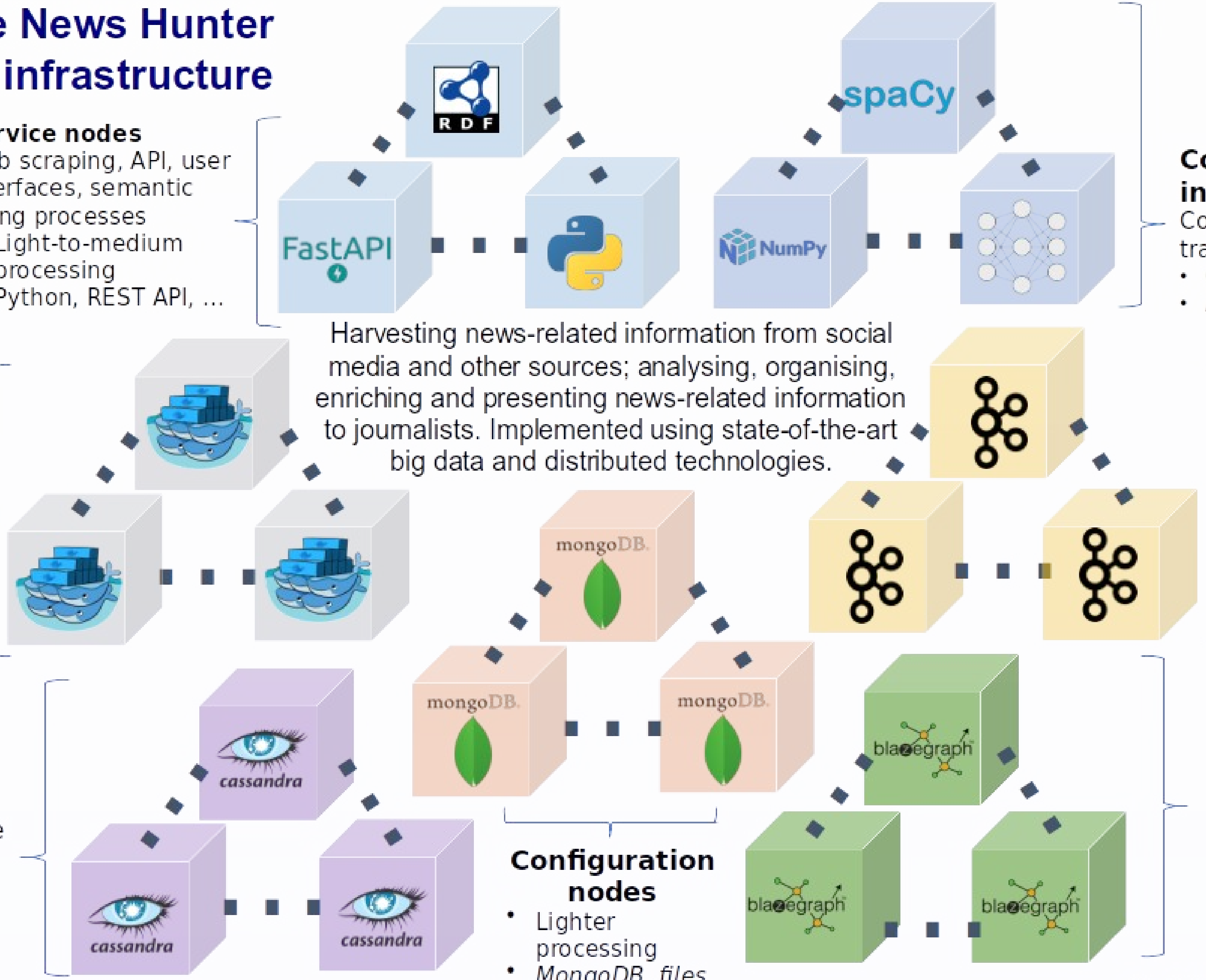
Services - lifters

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

Slide by Marc Gallofré Ocaña

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, …*

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.

Management nodes
Service orchestration and monitoring

- Lighter processing
- Docker Swarm

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*

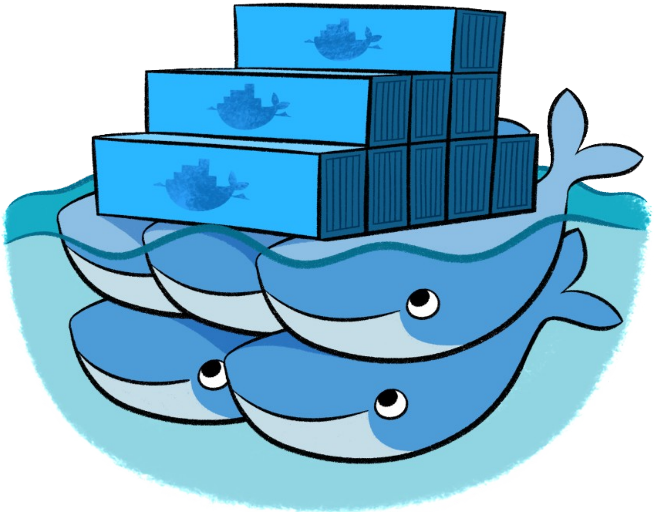
M. Gallofré Ocaña & A.L. Opdahl (2021)

Cloud infrastructure deployment tools

openstack®

HashiCorp
Terraform

ANSIBLE



Slide by Marc Gallofré Ocaña

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 of news and events)
- 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

Slide by Marc Gallofré Ocaña

Next week:

Rules (RDFS)