Making knowledge reusable at Yahoo!

A look at the Yahoo! Knowledge Graph

Nicolas Torzec

Semantic Technology & Business Conference
San Francisco
June 2013

Outline

1. **Driving Ideas**
2. **Overview of the Yahoo! Knowledge Graph**

   Knowledge Acquisition

   Knowledge Base

   Knowledge Consumption
3. **Some Applications at Yahoo!**

Introduction

DRIVING IDEAS

Knowledge Projects at Yahoo!

Search, Media, Content, Personalization, etc.

⇒ Various contexts. A few overlapping knowledge use cases.

Content understanding and query interpretation

- Recognize entities/topics using dictionaries and disambiguation features.

Answers, not links

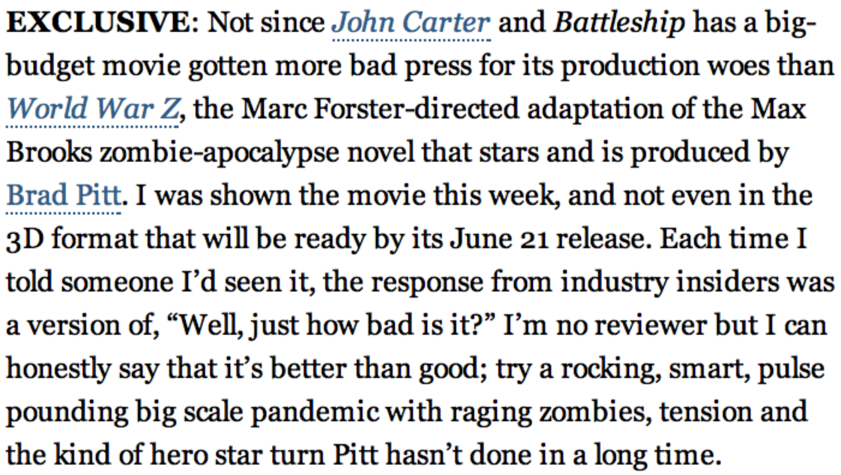
- Provide (semantic) information about entities

Related information in context

- Relate entities to each others, and entities to other types of objects

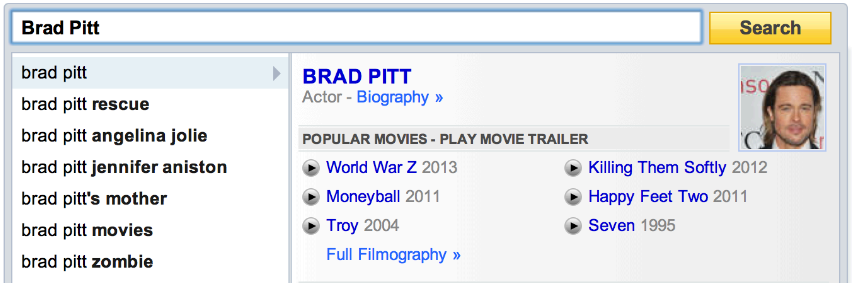
… Some illustrations

Content understanding



EXCLUSIVE: Not since *John Carter* and *Battleship* has a big-budget movie gotten more bad press for its production woes than *World War Z*, the Marc Forster-directed adaptation of the Max Brooks zombie-apocalypse novel that stars and is produced by Brad Pitt. I was shown the movie this week, and not even in the 3D format that will be ready by its June 21 release. Each time I told someone I'd seen it, the response from industry insiders was a version of, "Well, just how bad is it?" I'm no reviewer but I can honestly say that it's better than good; try a rocking, smart, pulse pounding big scale pandemic with raging zombies, tension and the kind of hero star turn Pitt hasn't done in a long time.

Query interpretation



brad pitt
brad pitt **rescue**
brad pitt **angelina jolie**
brad pitt **jennifer aniston**
brad pitt**'s mother**
brad pitt **movies**
brad pitt **zombie**

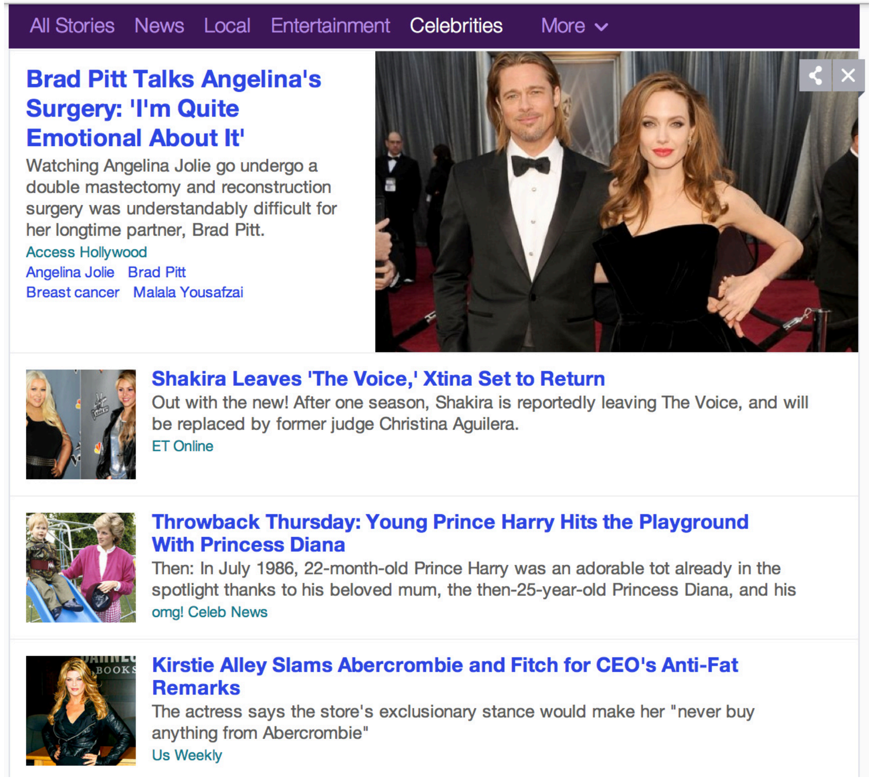
BRAD PITT
Actor - Biography »

POPULAR MOVIES - PLAY MOVIE TRAILER

▶ World War Z 2013	▶ Killing Them Softly 2012
▶ Moneyball 2011	▶ Happy Feet Two 2011
▶ Troy 2004	▶ Seven 1995

Full Filmography »

Personalization



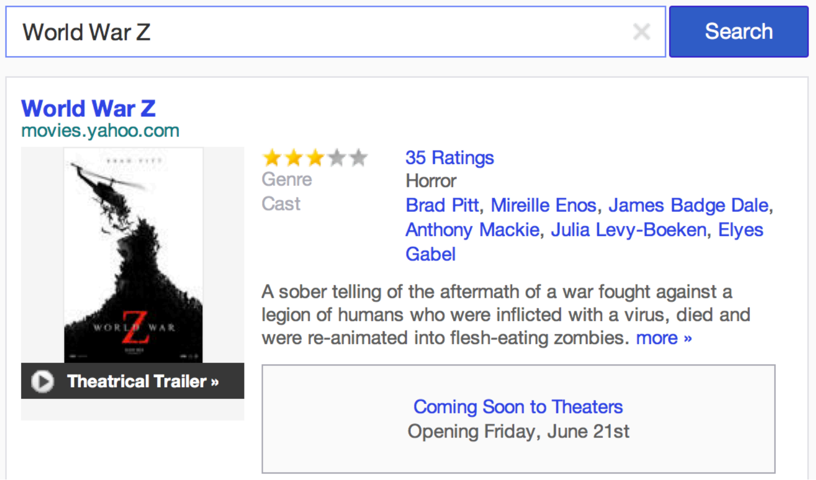
Brad Pitt Talks Angelina's Surgery: 'I'm Quite Emotional About It'
Watching Angelina Jolie go undergo a double mastectomy and reconstruction surgery was understandably difficult for her longtime partner, Brad Pitt.
Access Hollywood
Angelina Jolie Brad Pitt
Breast cancer Malala Yousafzai

Shakira Leaves 'The Voice,' Xtina Set to Return
Out with the new! After one season, Shakira is reportedly leaving The Voice, and will be replaced by former judge Christina Aguilera.
ET Online

Throwback Thursday: Young Prince Harry Hits the Playground With Princess Diana
Then: In July 1986, 22-month-old Prince Harry was an adorable tot already in the spotlight thanks to his beloved mum, the then-25-year-old Princess Diana, and his
omg! Celeb News

Kirstie Alley Slams Abercrombie and Fitch for CEO's Anti-Fat Remarks
The actress says the store's exclusionary stance would make her "never buy anything from Abercrombie"
Us Weekly

Entity-centric modules… Towards answers, not links!

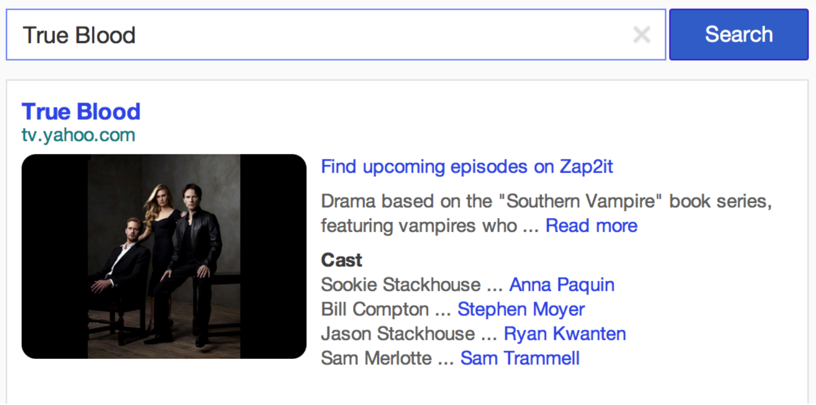


World War Z
movies.yahoo.com

★★★☆☆ 35 Ratings
Genre Horror
Cast Brad Pitt, Mireille Enos, James Badge Dale, Anthony Mackie, Julia Levy-Boeken, Elyes Gabel

A sober telling of the aftermath of a war fought against a legion of humans who were inflicted with a virus, died and were re-animated into flesh-eating zombies. more »

Coming Soon to Theaters
Opening Friday, June 21st

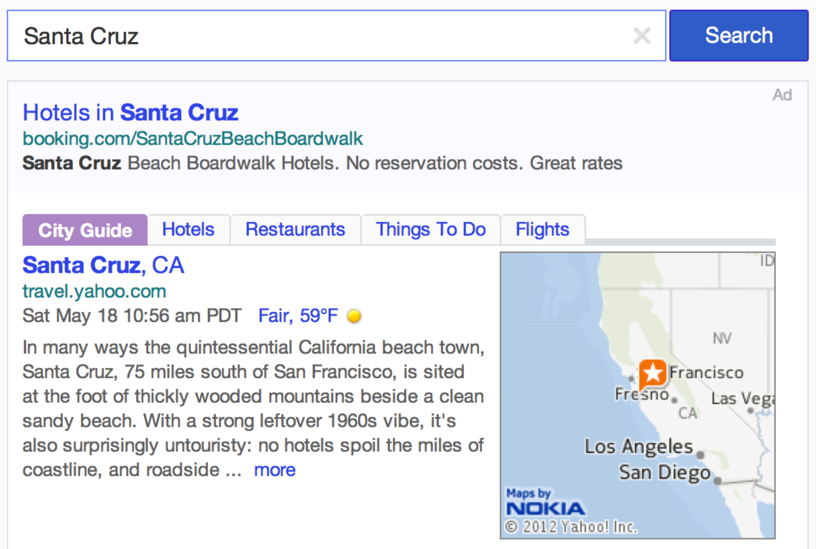


True Blood
tv.yahoo.com

Find upcoming episodes on Zap2it

Drama based on the "Southern Vampire" book series, featuring vampires who ... Read more

Cast
Sookie Stackhouse ... Anna Paquin
Bill Compton ... Stephen Moyer
Jason Stackhouse ... Ryan Kwanten
Sam Merlotte ... Sam Trammell



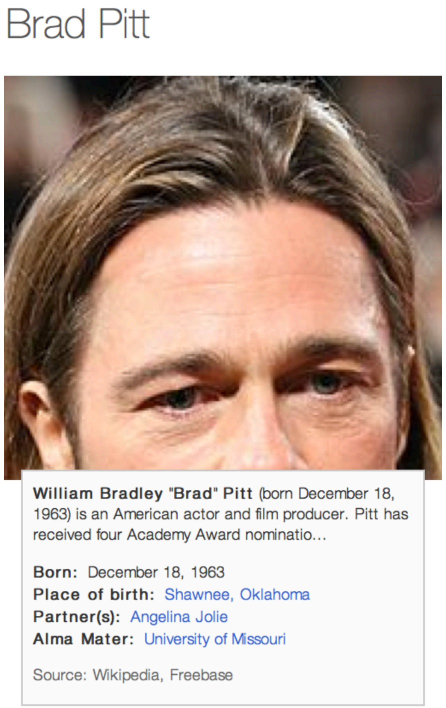
Ad

Hotels in **Santa Cruz**
booking.com/SantaCruzBeachBoardwalk
Santa Cruz Beach Boardwalk Hotels. No reservation costs. Great rates

City Guide | Hotels | Restaurants | Things To Do | Flights

Santa Cruz, CA
travel.yahoo.com
Sat May 18 10:56 am PDT Fair, 59°F

In many ways the quintessential California beach town, Santa Cruz, 75 miles south of San Francisco, is sited at the foot of thickly wooded mountains beside a clean sandy beach. With a strong leftover 1960s vibe, it's also surprisingly untouristy: no hotels spoil the miles of coastline, and roadside ... more



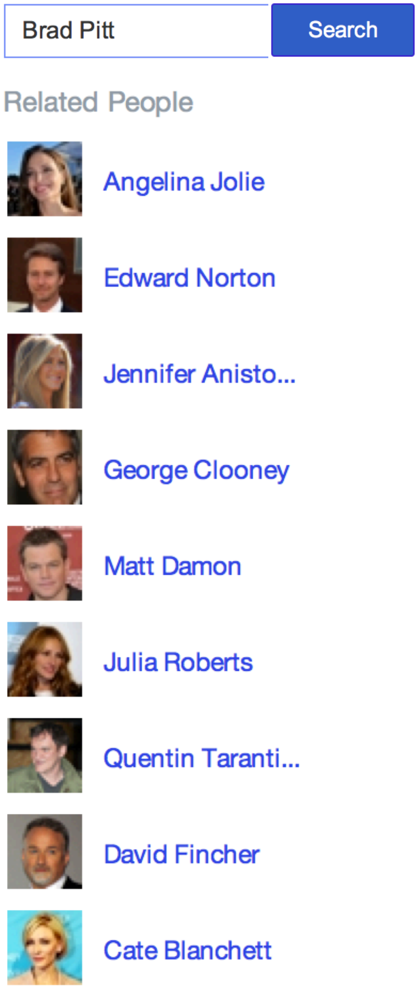
Brad Pitt

William Bradley "Brad" Pitt (born December 18, 1963) is an American actor and film producer. Pitt has received four Academy Award nominatio…

Born: December 18, 1963
Place of birth: Shawnee, Oklahoma
Partner(s): Angelina Jolie
Alma Mater: University of Missouri

Source: Wikipedia, Freebase

Related knowledge in context

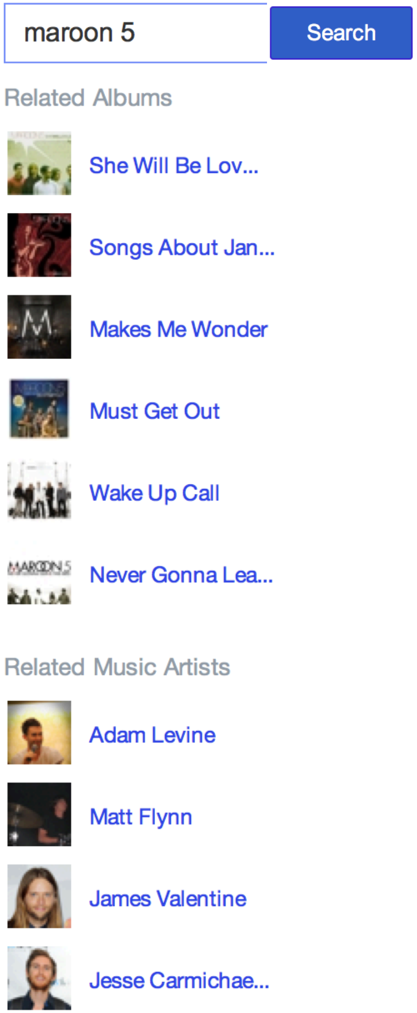


Brad Pitt

Search

Related People

- Angelina Jolie
- Edward Norton
- Jennifer Anisto...
- George Clooney
- Matt Damon
- Julia Roberts
- Quentin Taranti...
- David Fincher
- Cate Blanchett



maroon 5

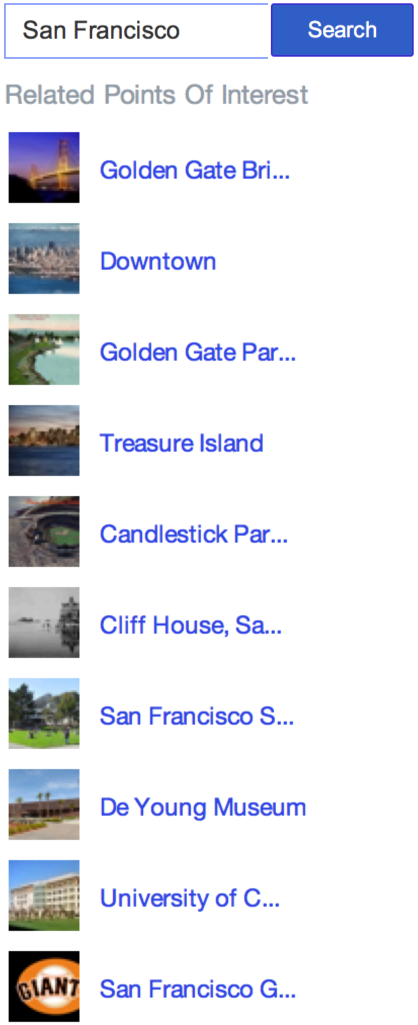
Search

Related Albums

- She Will Be Lov...
- Songs About Jan...
- Makes Me Wonder
- Must Get Out
- Wake Up Call
- Never Gonna Lea...

Related Music Artists

- Adam Levine
- Matt Flynn
- James Valentine
- Jesse Carmichae...



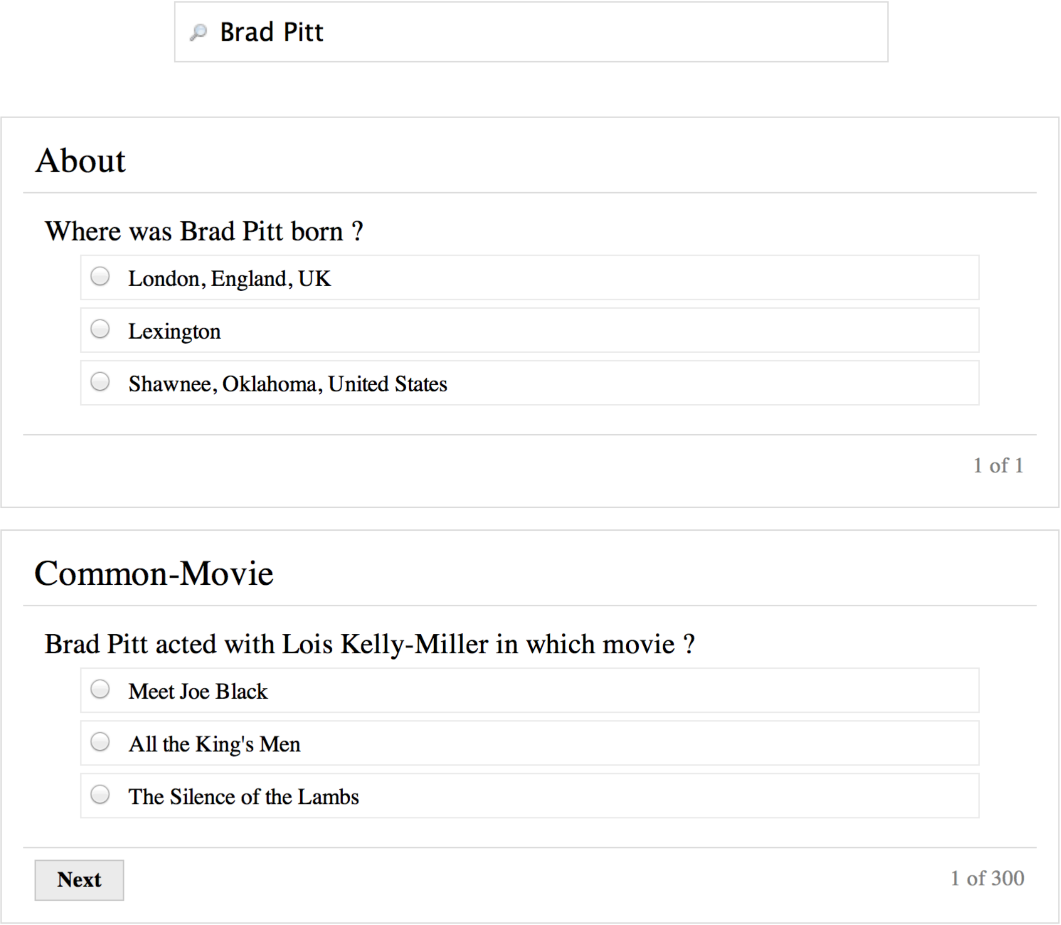
San Francisco

Search

Related Points Of Interest

- Golden Gate Bri...
- Downtown
- Golden Gate Par...
- Treasure Island
- Candlestick Par...
- Cliff House, Sa...
- San Francisco S...
- De Young Museum
- University of C...
- San Francisco G...

Related knowledge in context



Brad Pitt

About

Where was Brad Pitt born ?

○ London, England, UK

○ Lexington

○ Shawnee, Oklahoma, United States

1 of 1

Common-Movie

Brad Pitt acted with Lois Kelly-Miller in which movie ?

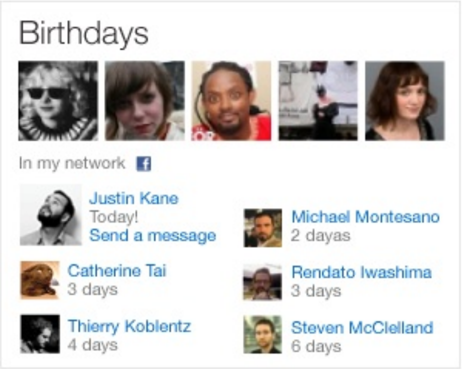
○ Meet Joe Black

○ All the King's Men

○ The Silence of the Lambs

Next

1 of 300



Birthdays

In my network

Justin Kane
Today!
Send a message

Michael Montesano
2 dayas

Catherine Tai
3 days

Rendato Iwashima
3 days

Thierry Koblentz
4 days

Steven McClelland
6 days

… and more.

Knowledge Graph at Yahoo!

Common Requirement

- **need for a central, unified, knowledge repository,**
- with key information about all the entities we care about,
- that can be used across Yahoo!, for various knowledge projects.

Value Proposition

- **Breadth, depth, and accuracy of data** by combining knowledge from multiple complementary sources and domains
- **Agility launching new experiences** by integrating key knowledge from various domains in a central repository, accessible across Yahoo

Overview of our Knowledge Platform

THE YAHOO! KNOWLEDGE GRAPH

The Yahoo! Knowledge Graph

A lightweight, central, unified, Knowledge Graph of all the entities and topics we care about at Yahoo!.

Provide key information about entities, and how they relate to each others.

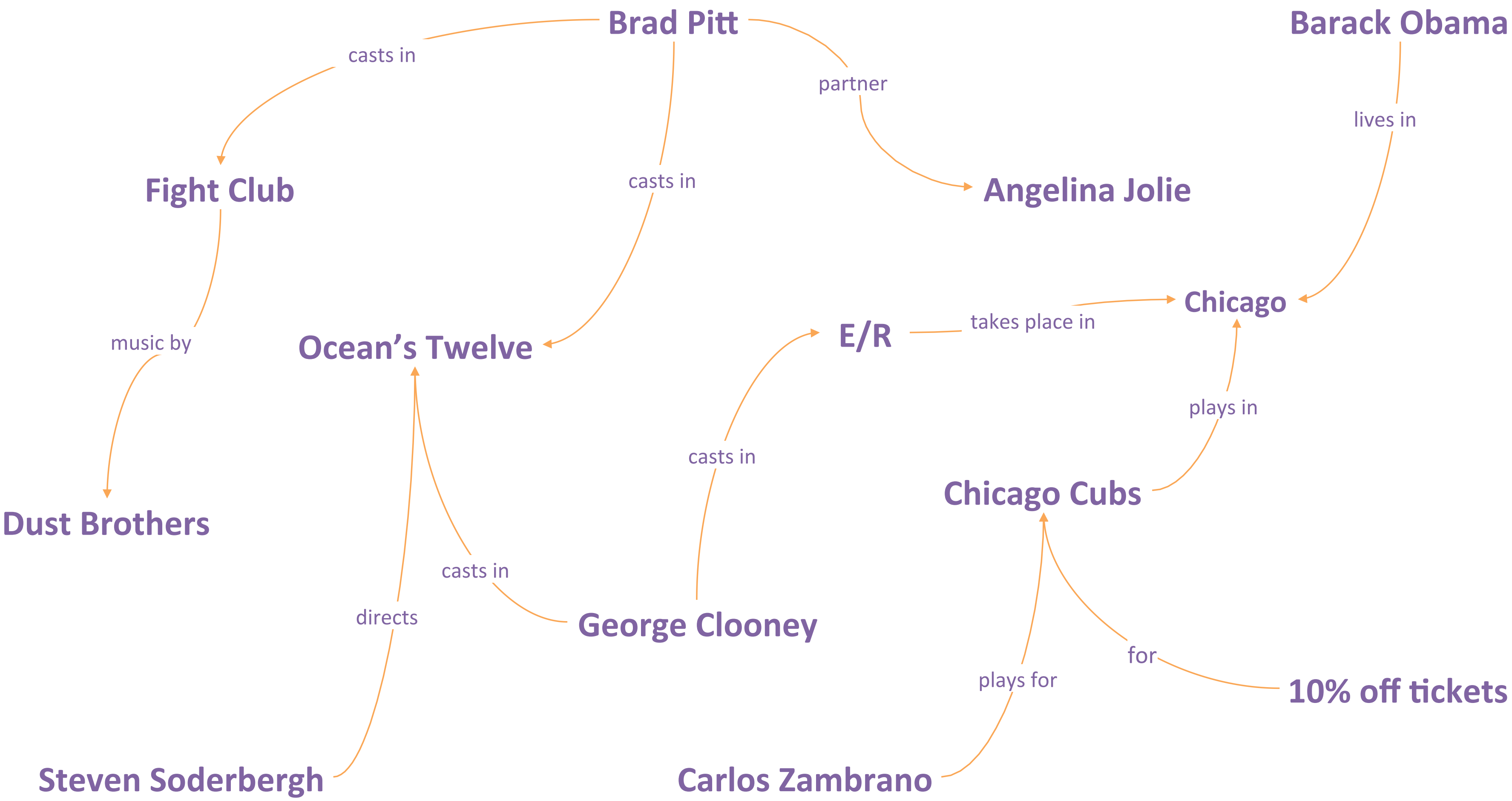
- Shallow and weakly-typed for a wide variety of domains
- Richer and strongly-typed for a few selected domains.

Provide disambiguation and interlinking across objects and data sources

Make this knowledge available within Yahoo!, in a convenient way

… to power knowledge projects in Search, Media, Personalization

Example (1/2)



Brad Pitt

Barack Obama

casts in

partner

lives in

Fight Club

casts in

Angelina Jolie

Chicago

takes place in

E/R

music by

Ocean’s Twelve

plays in

casts in

Dust Brothers

Chicago Cubs

casts in

directs

George Clooney

for

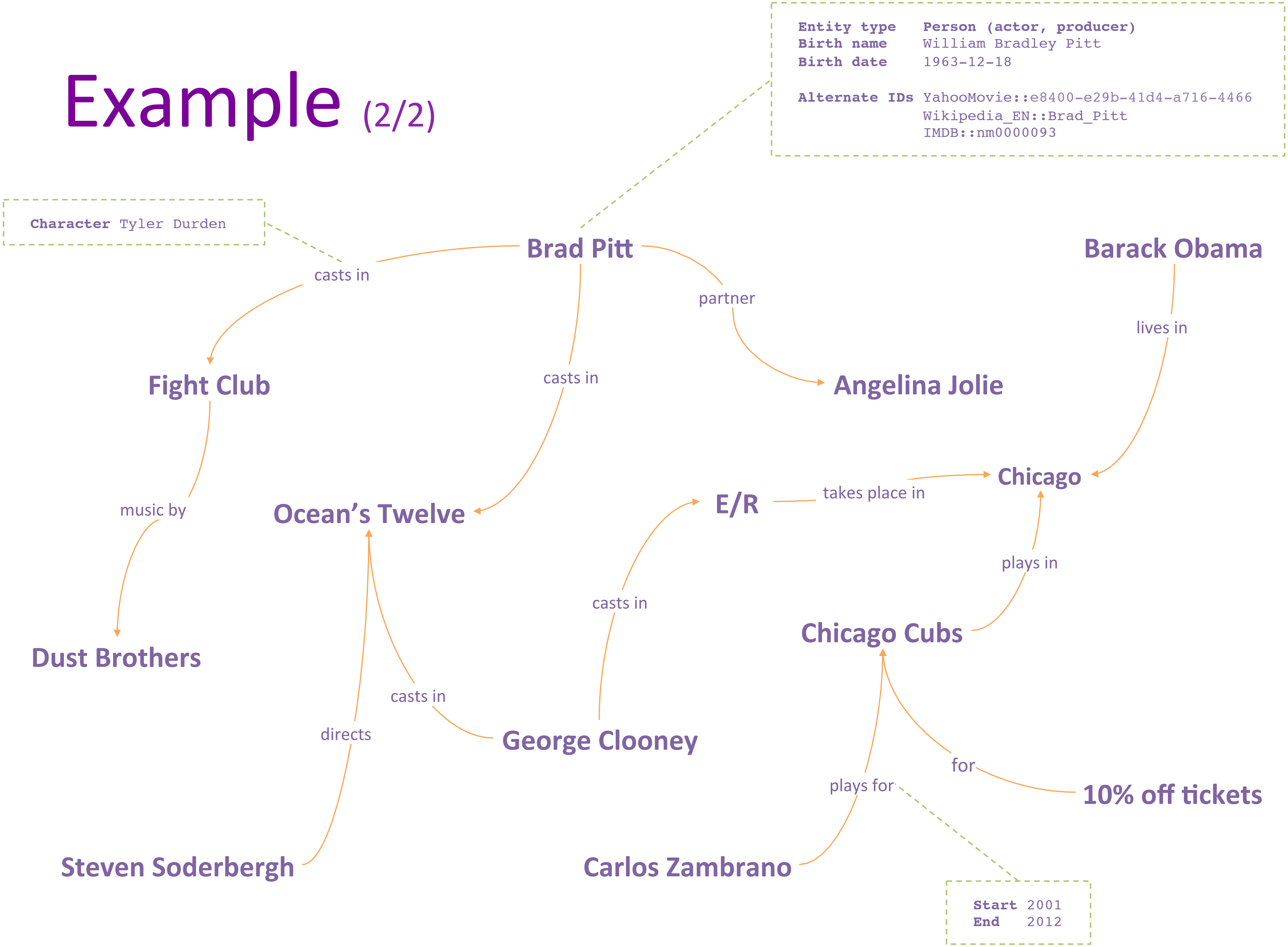
10% off tickets

plays for

Steven Soderbergh

Carlos Zambrano

Example (2/2)



Entity type **Person (actor, producer)**
Birth name William Bradley Pitt
Birth date 1963-12-18

Alternate IDs YahooMovie::e8400-e29b-41d4-a716-4466
Wikipedia_EN::Brad_Pitt
IMDB::nm0000093

Character Tyler Durden

Brad Pitt

casts in

partner

Barack Obama

lives in

Fight Club

casts in

Angelina Jolie

Chicago

music by

Ocean's Twelve

E/R

takes place in

plays in

casts in

Dust Brothers

Chicago Cubs

casts in

directs

George Clooney

for

plays for

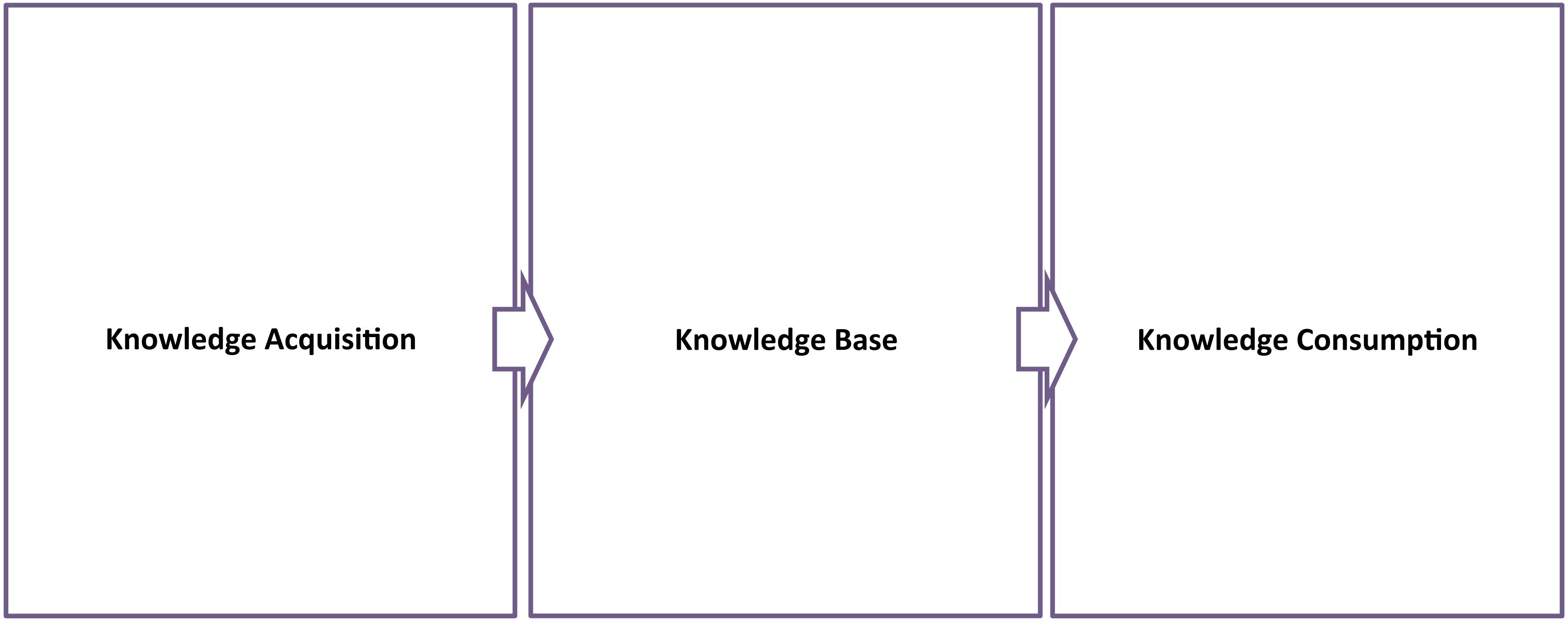
10% off tickets

Steven Soderbergh

Carlos Zambrano

Start 2001
End 2012

Platform Overview

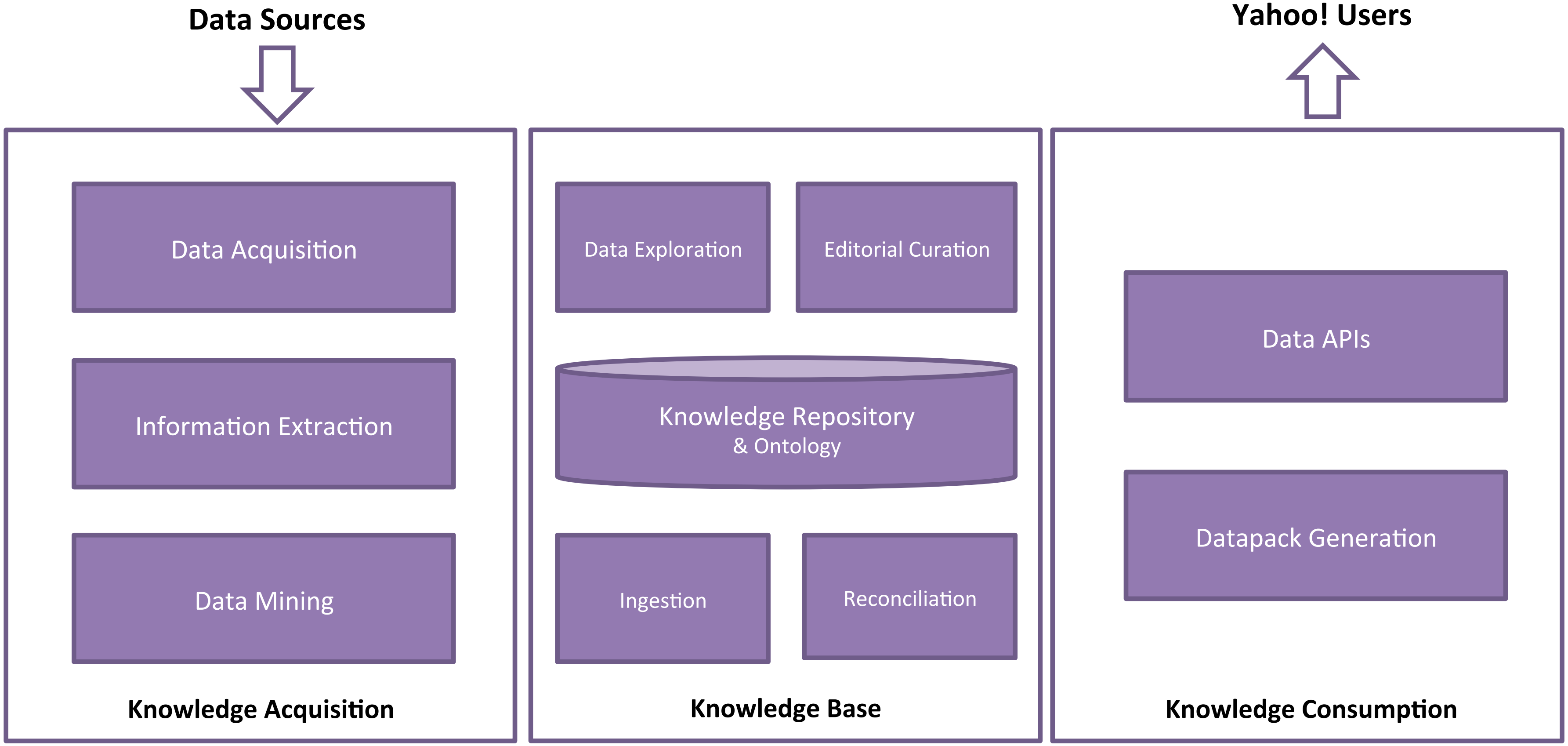


Knowledge Acquisition

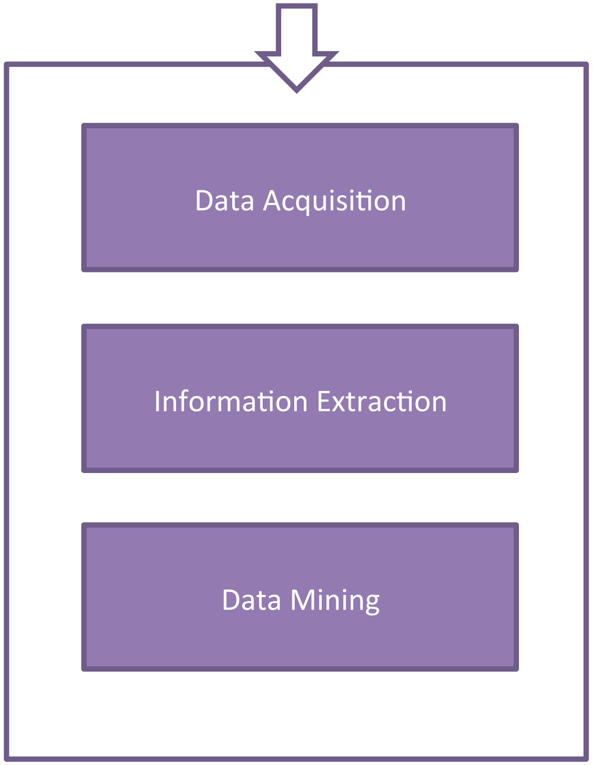
Knowledge Base

Knowledge Consumption

Platform Overview



Knowledge Acquisition



Collect, extract and mine information about entities from multiple complementary sources.

Data Sources

- **Complementary** and **possibly overlapping** sources…
- **Reference data sources** such as Wikipedia form a data backbone that provides (shallow) information for any domain
- **Specialized data sources** provide richer information for the domains we care the most: Finance, Sport, Entertainment, ...

Knowledge Acquisition

1. **Data acquisition**: offline dumps; online fetching; crawling
2. **Information extraction**: wrappers; more complex IE systems
3. **Data mining**: aliases; features for ranking/disambiguation

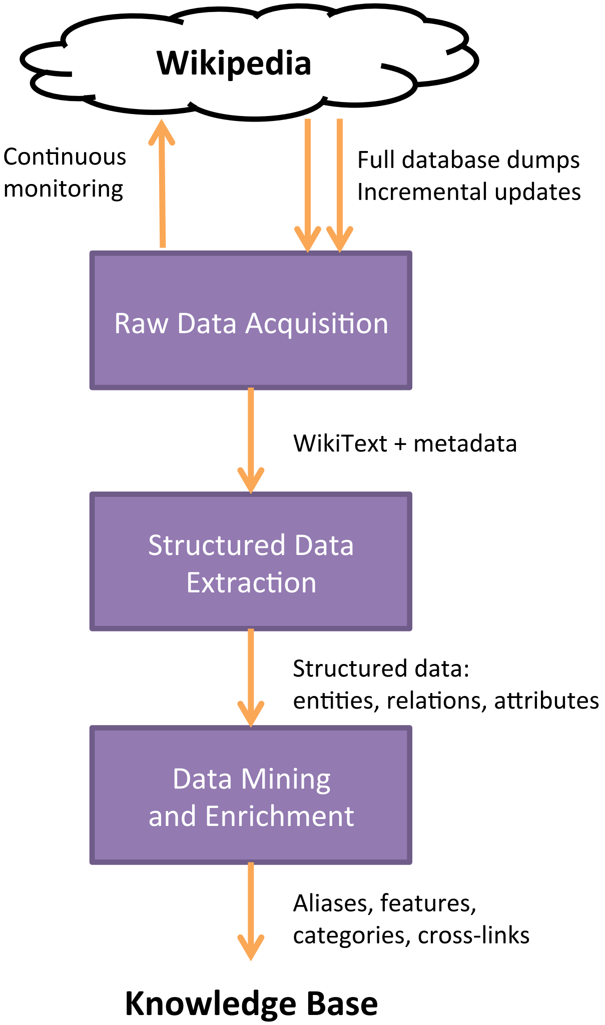
Challenges

- **Scale, volatility, heterogeneity, schema complexity, …**
- => Setup and Maintenance.

Knowledge Acquisition Platform

- Data pipelines run automatically on **Hadoop**…

Zoom: Extraction from Wikipedia


Wikipedia
Continuous monitoring
Full database dumps
Incremental updates
Raw Data Acquisition
WikiText + metadata
Structured Data Extraction
Structured data: entities, relations, attributes
Data Mining and Enrichment
Aliases, features, categories, cross-links
Knowledge Base

Data Acquisition

- **Monitor Wikipedia for new content continuously**
- Fetch new full database dumps when available
- Fetch incremental updates on an ongoing basis

Information Extraction

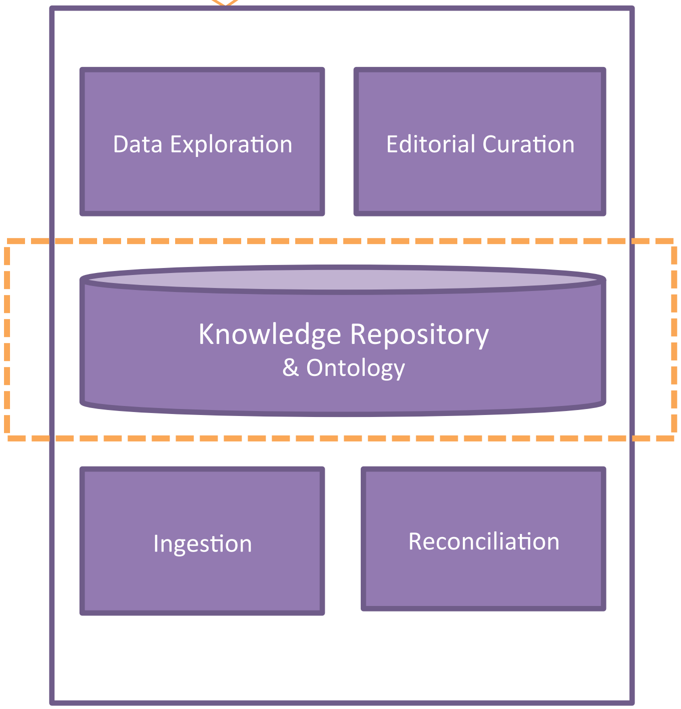
- **Extract Entities, Relations, and Attributes from Wikipedia**
- **DBpedia framework** extracts structured data from articles:
- => URI, label, abstracts, infobox, links, images, categories…
- **Other IE frameworks** do more advanced extraction/cleanup: => template/table/list extraction, type inference, mapping…

Data Mining & Enrichment

- **Derive information and features** from Wikipedia content: => aliases, ranking features, disambiguation features…
- **Enrich Wikipedia entities** with complementary information: => categorization, cross-linking…

Knowledge Base

Knowledge is integrated and managed centrally, in a unified knowledge base with a common ontology.



Knowledge is modeled as a property graph

- **Shallow** and weakly-typed **for a wide variety of domains**
- **Richer** and strongly-typed **for selected domains.**

Knowledge Representation

- **Common ontology and schemas,** aligned with schema.org.
- **250 classes / 800 properties** for modeling entities/relations
- Use case is **representation and validation.** No reasoning.
- Covers News, Finance, Sports, Entertainment, Listings, …

Challenges:

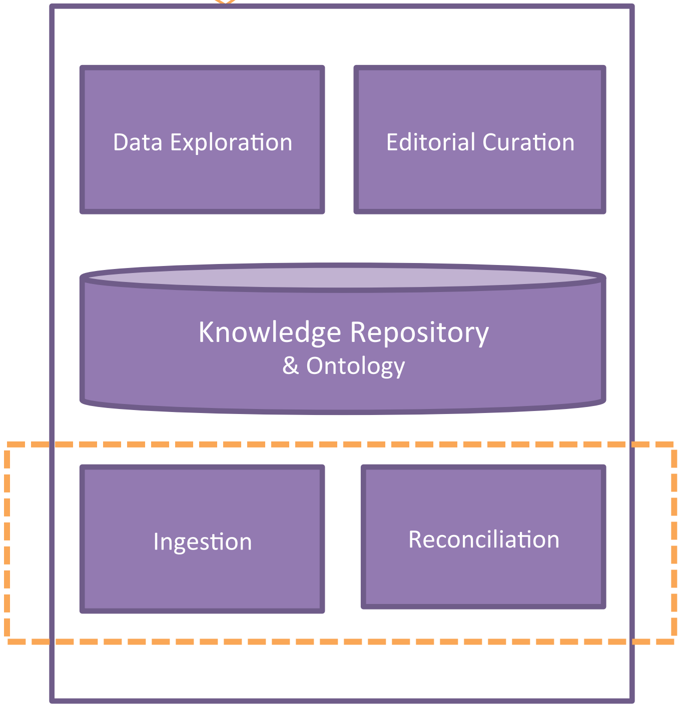
- Modeling **temporality, provenance, localization**, …
- **Trade-off**: expressiveness, comprehensiveness, ease-of-use!

Knowledge Base

- **Graph database** for random access and online processing.
- **Hadoop file system** for offline batch processing.

Knowledge Base

Knowledge is integrated and managed centrally, in a unified knowledge base with a common ontology.



Knowledge Ingestion

- Entities/Relations are **aligned with the common ontology**
- Info are **mapped/normalized to the standard schemas**

Challenges: semantic discrepancies, schema complexity, …

Knowledge Reconciliation = Co-reference Resolution

- **Match/Blend** objects w/ **same referent,** across data sources
- We combine editorial, heuristics, and machine learning

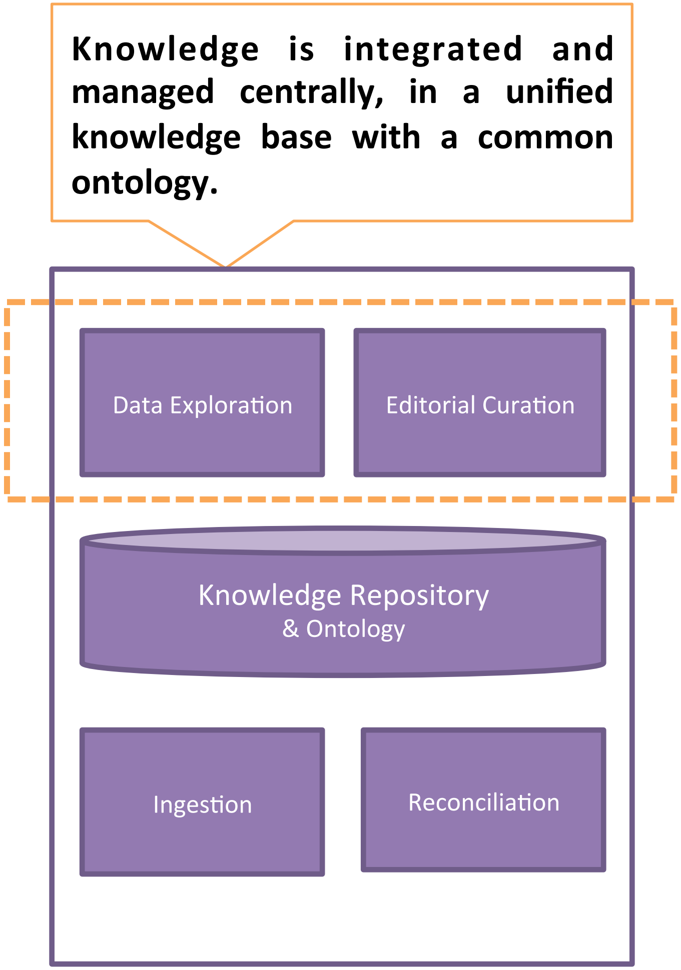
Approach

- Favor **domain-dependent** over domain-agnostic resolution
- Use **blocking + pair-wise matching + merging**
- Use **soft linking and physical blending**
- **Offline reconciliation**, but moving to online reconciliation

Challenges:

1. Scaling reconciliation to **data size, update frequency**
2. Scaling reconciliation to **new domains and entity types**.
3. Making **entity IDs persistent** over full/incremental updates

Knowledge Base


Knowledge is integrated and managed centrally, in a unified knowledge base with a common ontology.
Data Exploration
Editorial Curation
Knowledge Repository
& Ontology
Ingestion
Reconciliation

Editorial Curation

- Create, delete, merge, and split **entities and relations**
- Review, update, delete, and enrich **associated information**
- **Interactive curation** + batch **import/export of small lists**

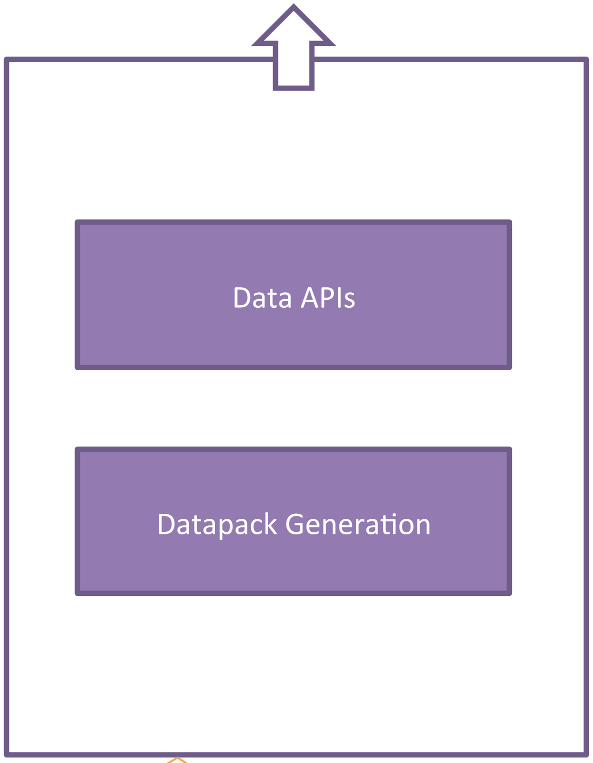
Internal Data APIs

- Low-Level: entity/relation CRUD + Graph Search
- High-Level: Ingest, Reconcile, Curate, Explore, Export

Coverage

- **Focus on most important entities/relations** (i.e. head)
- **Domains: News, Finance, Sport, Movie, TV, Music, and Geo**
- Overall: ~10M entities, ~10M relations, ~30M properties.

Knowledge Consumption



Make knowledge available via custom data packs and data APIs

Public data APIs

1. Export and **index entities/relations in a Search platform**
2. Expose **entities/relations** via **Lookup APIs**
3. Expose **entities/relations** via **Structured Search APIs**

Custom datapack generation

1. Run **offline queries** to **retrieve sub-graphs** from KG
2. Enrich with **complementary data** and/or **indirect relations**
3. Apply **custom filtering, transformation, and formatting**
4. Publish resulting **datasets** for offline batch data processing

Data Quality Monitoring

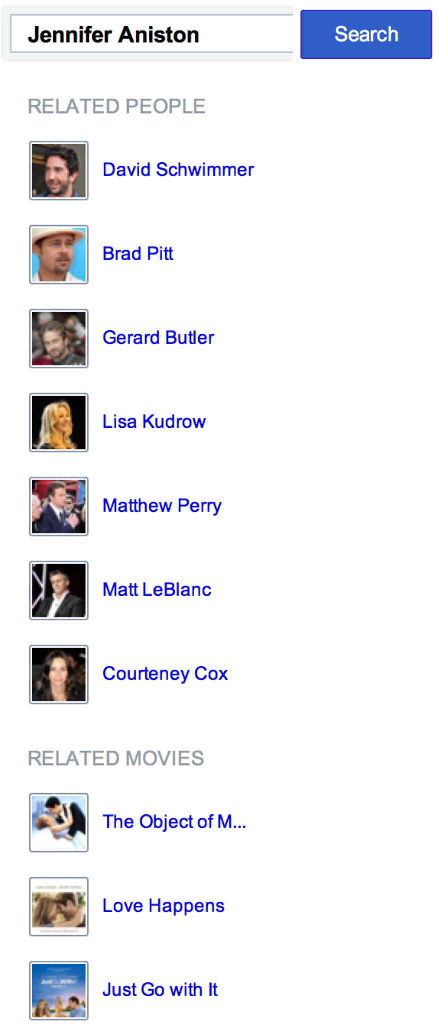
Some metrics we are interested in …

Coverage	Number of domains and entity types covered Number of entities within each domain/type
Richness	Number of attributes and relations for each entity type Number of attributes and relations populated for each entity type
Comprehensiveness	% of important entities/relations/facts found in the knowledge base % of entities/relations/facts mentioned in Search queries and New articles
Correctness	Type correctness: correctness of entity types. Fact correctness: correctness of relations and attribute values
Interlinking	Precision and recall of reconciliation Level of interlinking across internal sources, external sources
Freshness	Freshness of entities/relations/attributes compared to activity about them (popularity, trending/decay, time sensitivity, etc.)
	… evaluated via uniform random sampling, weighted sampling, etc.

(Leveraging the Knowledge Graph)

SOME APPLICATIONS

Recommending Entities on Y! for Web Search



Jennifer Aniston

Search

RELATED PEOPLE

- David Schwimmer
- Brad Pitt
- Gerard Butler
- Lisa Kudrow
- Matthew Perry
- Matt LeBlanc
- Courteney Cox

RELATED MOVIES

- The Object of M...
- Love Happens
- Just Go with It

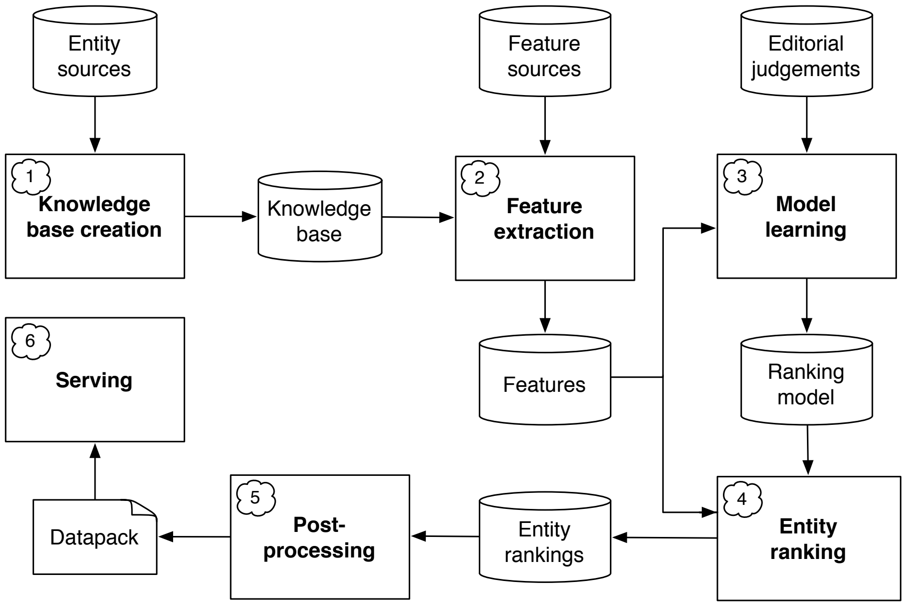


Fig. 1. High-level architecture of the Spark entity recommender system.

More Applications

Content Understanding and Personalization

Yahoo! Verticals

Search

…

The end

QUESTION?

Contacts

About Yahoo! Labs

http://labs.yahoo.com

About the Project

http://labs.yahoo.com/project/knowledge-acquisition-and-management/

About the Author

Labs Page: http://labs.yahoo.com/author/torzecn/

Email: torzecn@yahoo-inc.com

Twitter: @nicolastorzec