SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE
10th ANNUAL
August 19-21, 2014 | San Jose, California USA

YAHOO! LABS

Yahoo Knowledge Graph

Making Knowledge Reusable at Yahoo

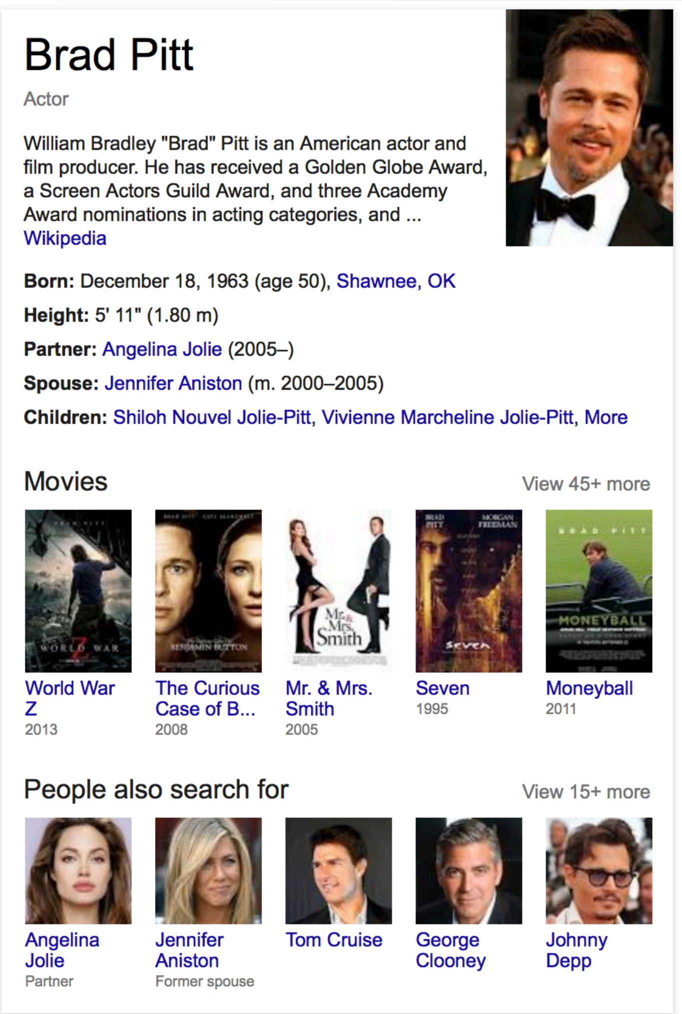
PRESENTED BY **Nicolas Torzec** | August 20, 2014

Background & Context

Google Knowledge Graph

Introducing the Knowledge Graph: Things, Not Strings.

May 16th, 2012



Brad Pitt

Actor

William Bradley "Brad" Pitt is an American actor and film producer. He has received a Golden Globe Award, a Screen Actors Guild Award, and three Academy Award nominations in acting categories, and ...
Wikipedia

Born: December 18, 1963 (age 50), Shawnee, OK

Height: 5' 11" (1.80 m)

Partner: Angelina Jolie (2005–)

Spouse: Jennifer Aniston (m. 2000–2005)

Children: Shiloh Nouvel Jolie-Pitt, Vivienne Marcheline Jolie-Pitt, More

Movies — View 45+ more

World War Z 2013

The Curious Case of B... 2008

Mr. & Mrs. Smith 2005

Seven 1995

Moneyball 2011

People also search for — View 15+ more

Angelina Jolie Partner

Jennifer Aniston Former spouse

Tom Cruise

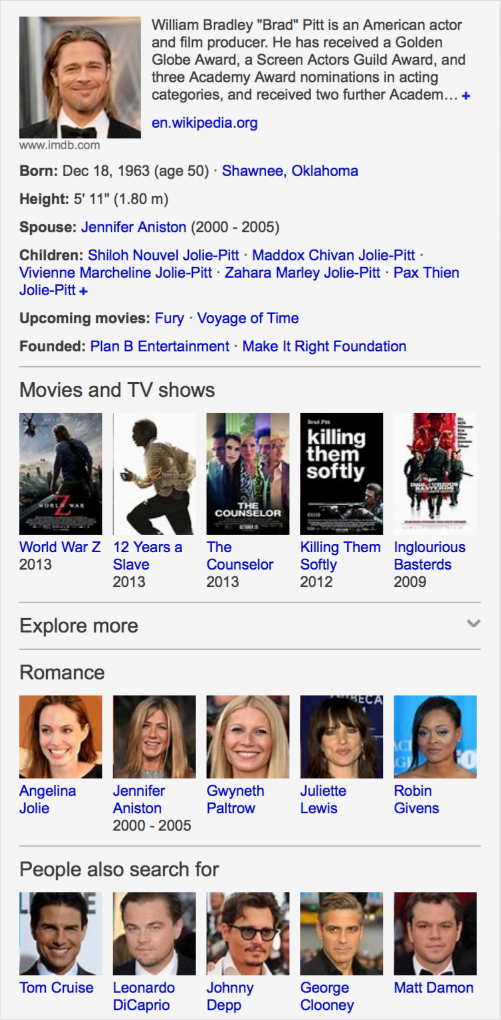
George Clooney

Johnny Depp

Bing Knowledge Graph

Understand Your World with Bing.

March 21st, 2013



William Bradley "Brad" Pitt is an American actor and film producer. He has received a Golden Globe Award, a Screen Actors Guild Award, and three Academy Award nominations in acting categories, and received two further Academ… +

en.wikipedia.org

www.imdb.com

Born: Dec 18, 1963 (age 50) · Shawnee, Oklahoma

Height: 5' 11" (1.80 m)

Spouse: Jennifer Aniston (2000 - 2005)

Children: Shiloh Nouvel Jolie-Pitt · Maddox Chivan Jolie-Pitt · Vivienne Marcheline Jolie-Pitt · Zahara Marley Jolie-Pitt · Pax Thien Jolie-Pitt +

Upcoming movies: Fury · Voyage of Time

Founded: Plan B Entertainment · Make It Right Foundation

Movies and TV shows

World War Z 2013 · 12 Years a Slave 2013 · The Counselor 2013 · Killing Them Softly 2012 · Inglourious Basterds 2009

Explore more

Romance

Angelina Jolie · Jennifer Aniston 2000 - 2005 · Gwyneth Paltrow · Juliette Lewis · Robin Givens

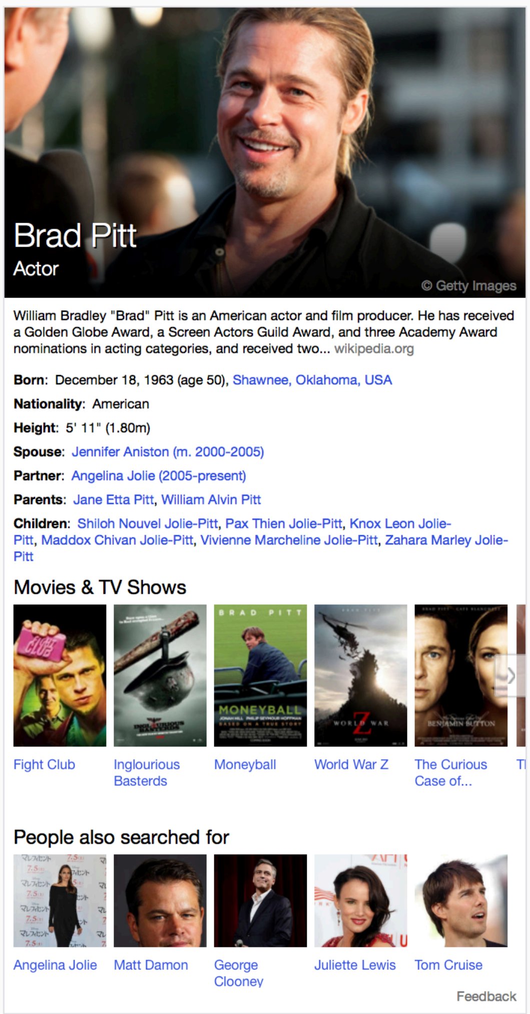
People also search for

Tom Cruise · Leonardo DiCaprio · Johnny Depp · George Clooney · Matt Damon

Yahoo Knowledge Graph

Yahoo Entity Search.

Soft launch in Nov. 2013



Brad Pitt

Actor

© Getty Images

William Bradley "Brad" Pitt is an American actor and film producer. He has received a Golden Globe Award, a Screen Actors Guild Award, and three Academy Award nominations in acting categories, and received two... wikipedia.org

Born: December 18, 1963 (age 50), Shawnee, Oklahoma, USA

Nationality: American

Height: 5' 11" (1.80m)

Spouse: Jennifer Aniston (m. 2000-2005)

Partner: Angelina Jolie (2005-present)

Parents: Jane Etta Pitt, William Alvin Pitt

Children: Shiloh Nouvel Jolie-Pitt, Pax Thien Jolie-Pitt, Knox Leon Jolie-Pitt, Maddox Chivan Jolie-Pitt, Vivienne Marcheline Jolie-Pitt, Zahara Marley Jolie-Pitt

Movies & TV Shows

Fight Club, Inglourious Basterds, Moneyball, World War Z, The Curious Case of...

People also searched for

Angelina Jolie, Matt Damon, George Clooney, Juliette Lewis, Tom Cruise

Other “Knowledge Graphs"

Rich, domain-specific, graphs	Wolfram Alpha, BBC, Rovi, TMS, Baseline, Gracenote, Amazon, Walmart Labs
Interest graphs	Gravity Adchemy
Social graphs	Facebook LinkedIn
Reference knowledge graphs	Freebase + Yago, Wikidata, DBpedia, and other Wikipedia-based projects…
Common-sense knowledge graphs	Cyc

Scope

Vision

- **A unified knowledge graph for Yahoo**
  - All entities and topics relevant to Yahoo (users)
  - Rich information about entities: facts, relationships, features
  - Identifiers, interlinking across data sources, and links to relevant services

- **To power knowledge-based services at Yahoo**
  - **Search:** display, and search for, information about entities
  - **Discovery:** relate entities, interconnect data sources, link to relevant services
  - **Understanding:** recognize entities in queries and text

- **Managed and served by a central knowledge team / platform**

Value Proposition

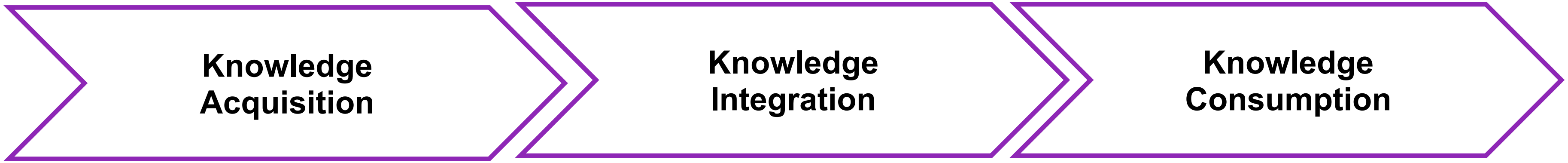
- **Data breadth, depth, and accuracy**
  - Combine information from multiple complementary/overlapping data sources

- **Centralized expertise**
- **Common technologies**
- **Same knowledge graph**

leveraged across the company

- **Speed and agility at launching new and richer experiences**

In a Nutshell



Knowledge Acquisition

Knowledge Integration

Knowledge Consumption

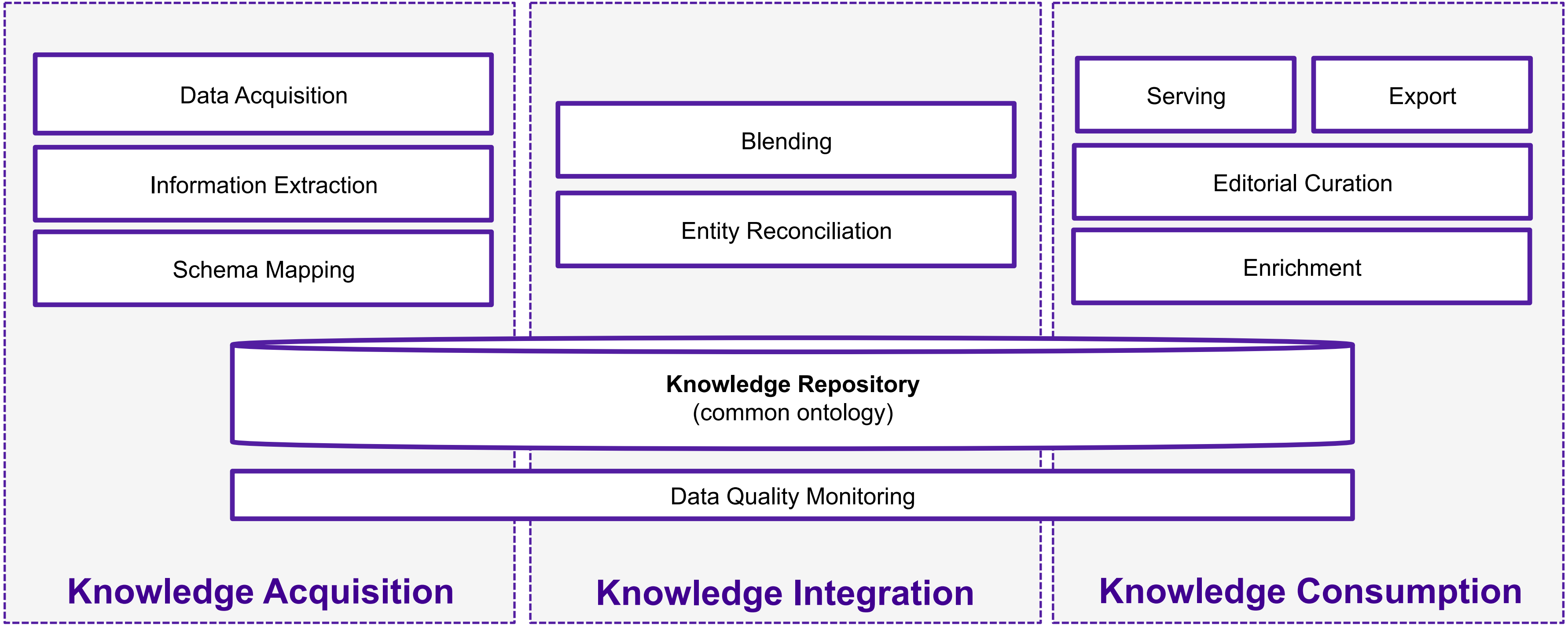
Ongoing information extraction, from complementary sources.

Reconciliation into a unified knowledge repository.

Enrichment and serving…

Making knowledge reusable at Yahoo

Key Tasks



Data Acquisition

Information Extraction

Schema Mapping

Blending

Entity Reconciliation

Serving

Export

Editorial Curation

Enrichment

Knowledge Repository
(common ontology)

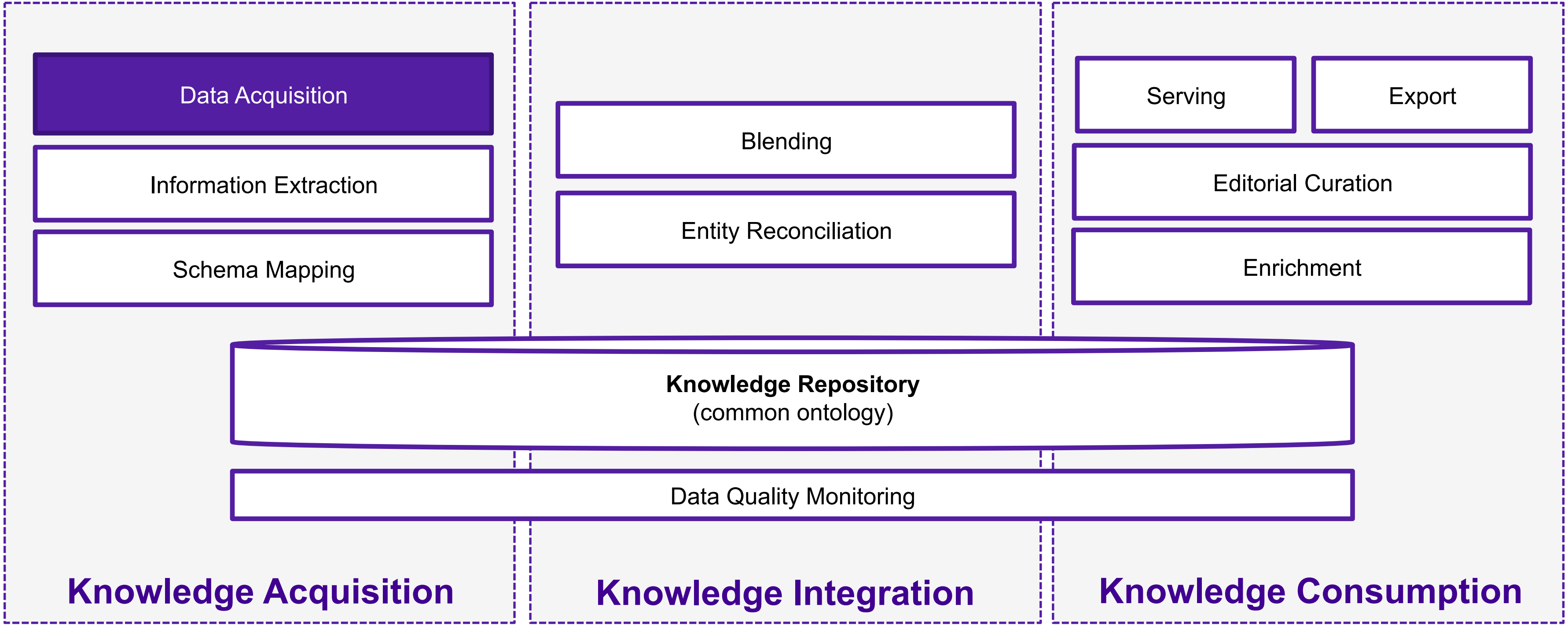
Data Quality Monitoring

Knowledge Acquisition

Knowledge Integration

Knowledge Consumption

Data Acquisition



Data Acquisition

Information Extraction

Schema Mapping

Blending

Entity Reconciliation

Serving

Export

Editorial Curation

Enrichment

Knowledge Repository
(common ontology)

Data Quality Monitoring

Knowledge Acquisition

Knowledge Integration

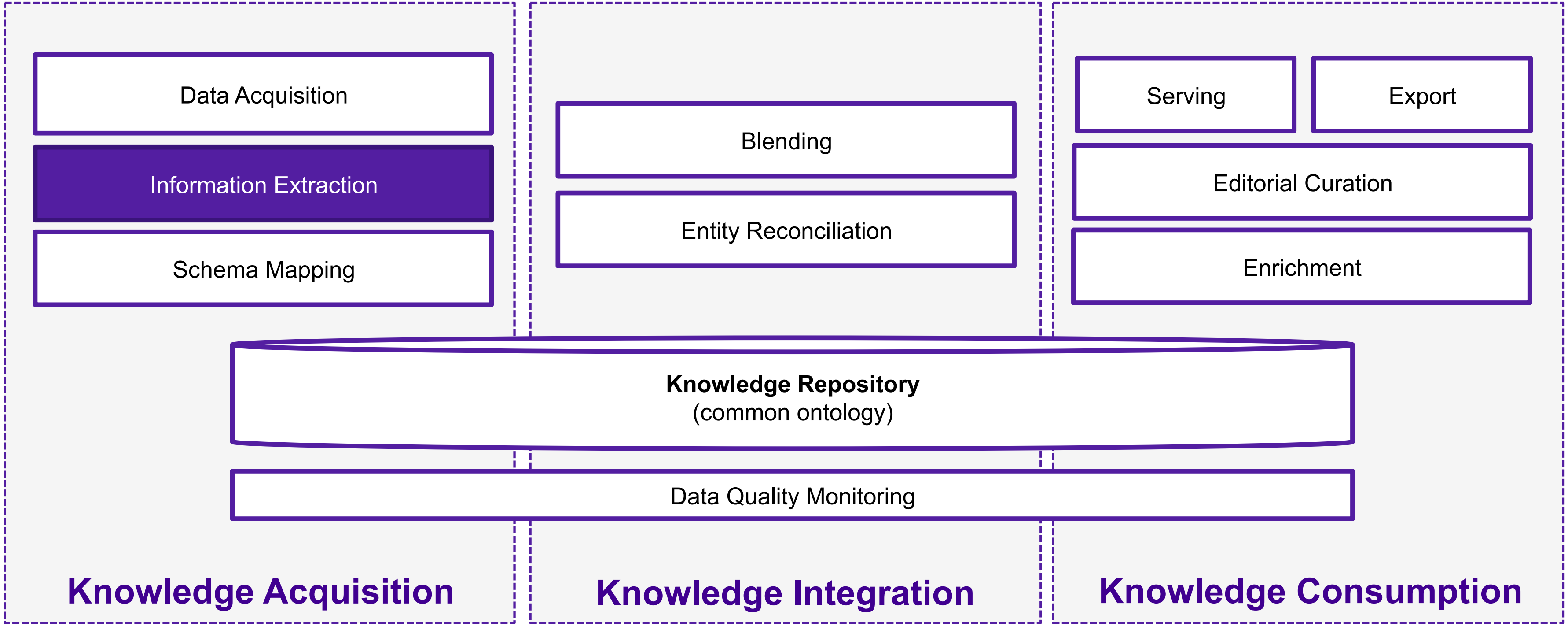
Knowledge Consumption

Data Acquisition

- **Multiple complementary data sources**
  - Combine and cross-validate data from *authoritative** sources
  - Reference data sources such as Wikipedia and Freebase form our backbone
  - Specialized data sources such as TMS and Music Brainz adds breadth/depth
  - Optimize for relevance, comprehensiveness, correctness, freshness, consistency

- **Ongoing acquisition of raw data**
  - Feed acquisition from open data sources and paid providers
  - Web/Targeted crawling, online fetching, ad hoc acquisition (e.g. Wikipedia monitoring)
  - Deal w/ operational complexity: data size, bandwidth, update frequency, license, ©

Information Extraction

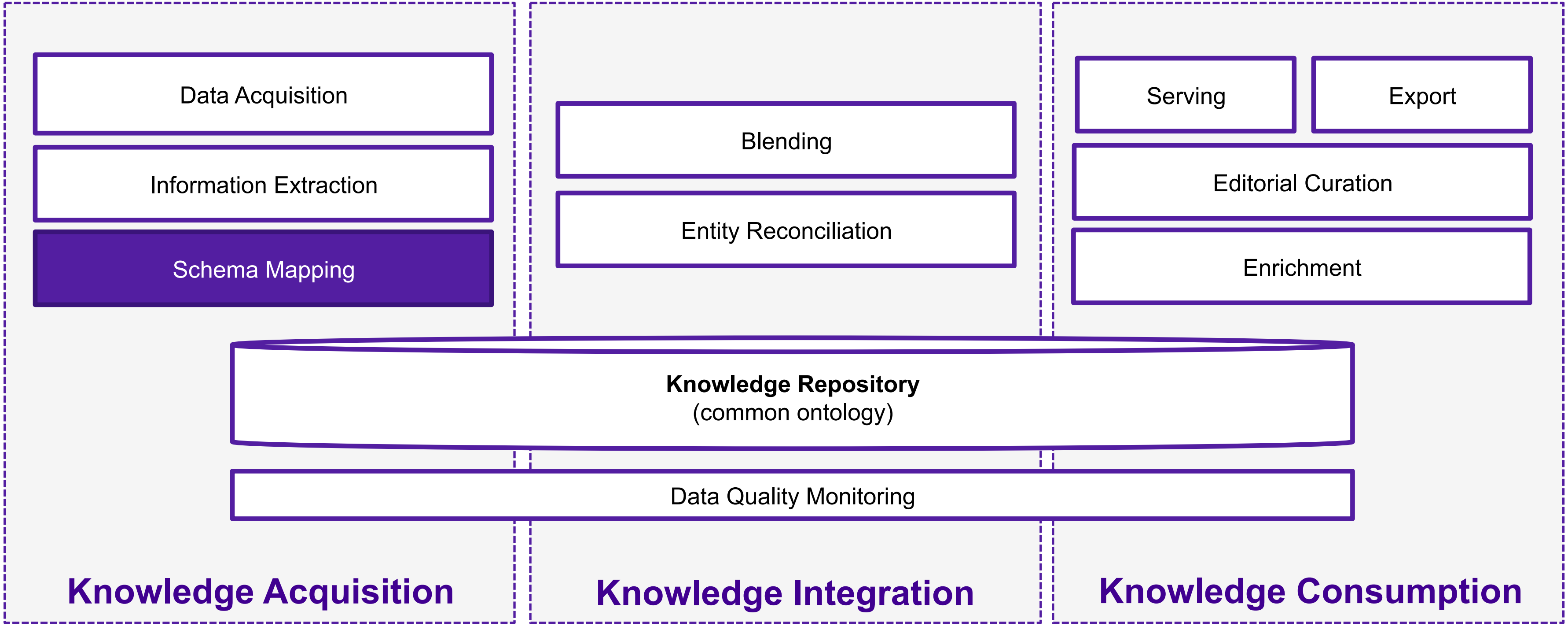


Information Extraction

- **Extraction of entities, attributes, relationships, features**
  - Deal w/ scale, volatility, heterogeneity, inconsistency, schema complexity, breakage
  - Expensive to build and maintain (i.e. declarative rules, expert’s knowledge, ML…)
  - Being able to measure and monitor data quality is key

- **Mixed approach**
  1. Parsing of large data feeds and online data APIs
  2. Structured data extraction on the Web: markup, Web scraping, Wrapper induction,
  3. Wikipedia mining, Web mining, News mining, open information extraction

Schema Mapping



Data Acquisition

Information Extraction

Schema Mapping

Blending

Entity Reconciliation

Serving

Export

Editorial Curation

Enrichment

Knowledge Repository
(common ontology)

Data Quality Monitoring

Knowledge Acquisition

Knowledge Integration

Knowledge Consumption

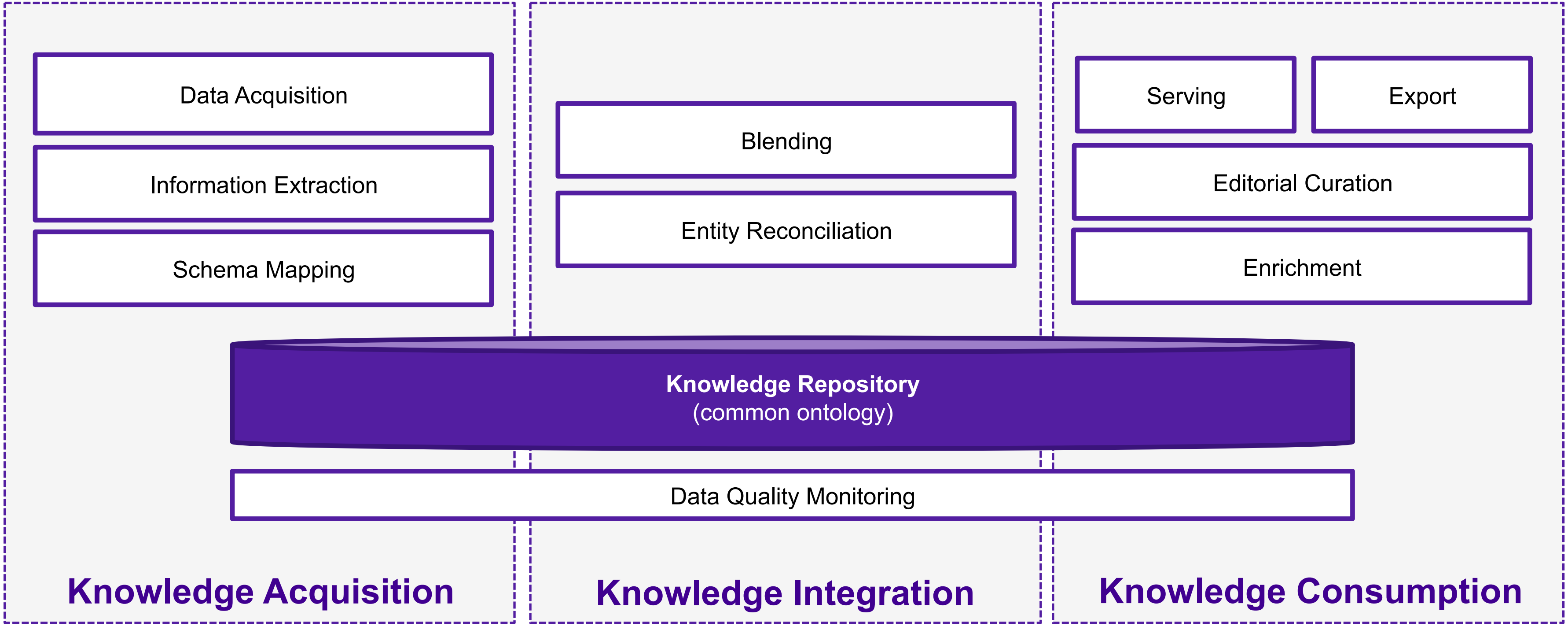
Schema Mapping

- **Normalization to common ontology, schemas, and data types/units**
  - Upfront normalization: uniform data facilitate downstream usage
  - Validation against the ontology to ensure well-formedness, validity, and consistency

Ontology alignment	`<Mad_Men, isA, TVSeries>`	Classifiers: heuristics + ML
Schema mapping	`<Jon_Hamm, birthplace, St._Louis>`	Template-driven ; mostly declarative
Data normalization	`<Jon_Hamm, birthdate, “1971-03-10”>`	Common plugins

- **Challenges**
  - Noisy information extraction: e.g. strong types vs. inferred types
  - Discrepancies between source/target ontologies: e.g. can Pal_(dog) be an actor?
  - Schema complexity and schema evolutions…

Knowledge Representation



Knowledge Representation

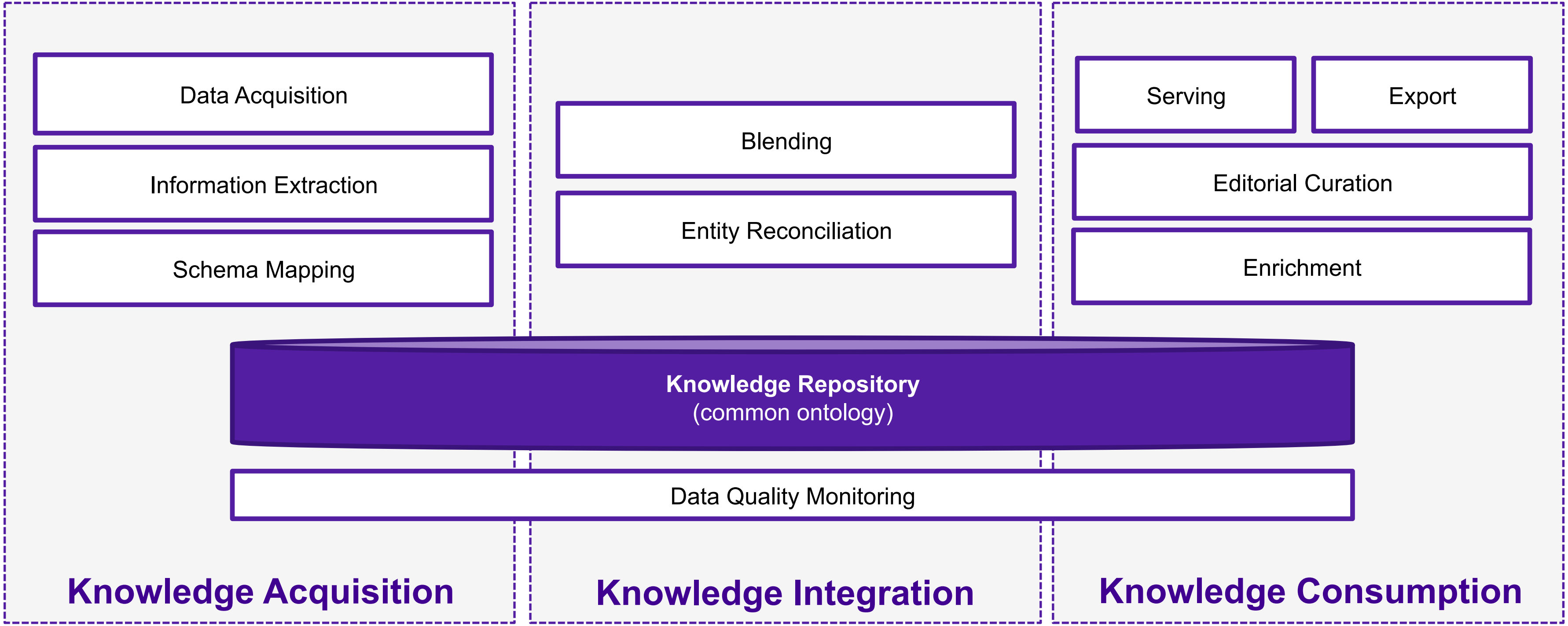
- **Property Graph data model**
  - JSON-LD serialization when needed

- **Common ontology**
  - OWL ontology. Focuses on representation & validation, not reasoning
  - Covers domains relevant to Yahoo: 300 classes, 500 object properties, 800 data prop.

Challenges

- Modeling/managing temporality, provenance, license, localization
- Soundness, expressiveness and comprehensiveness … vs. practicality
- Collaborative development, conflicting modeling, schema evolution over time

Knowledge Repository



Knowledge Repository

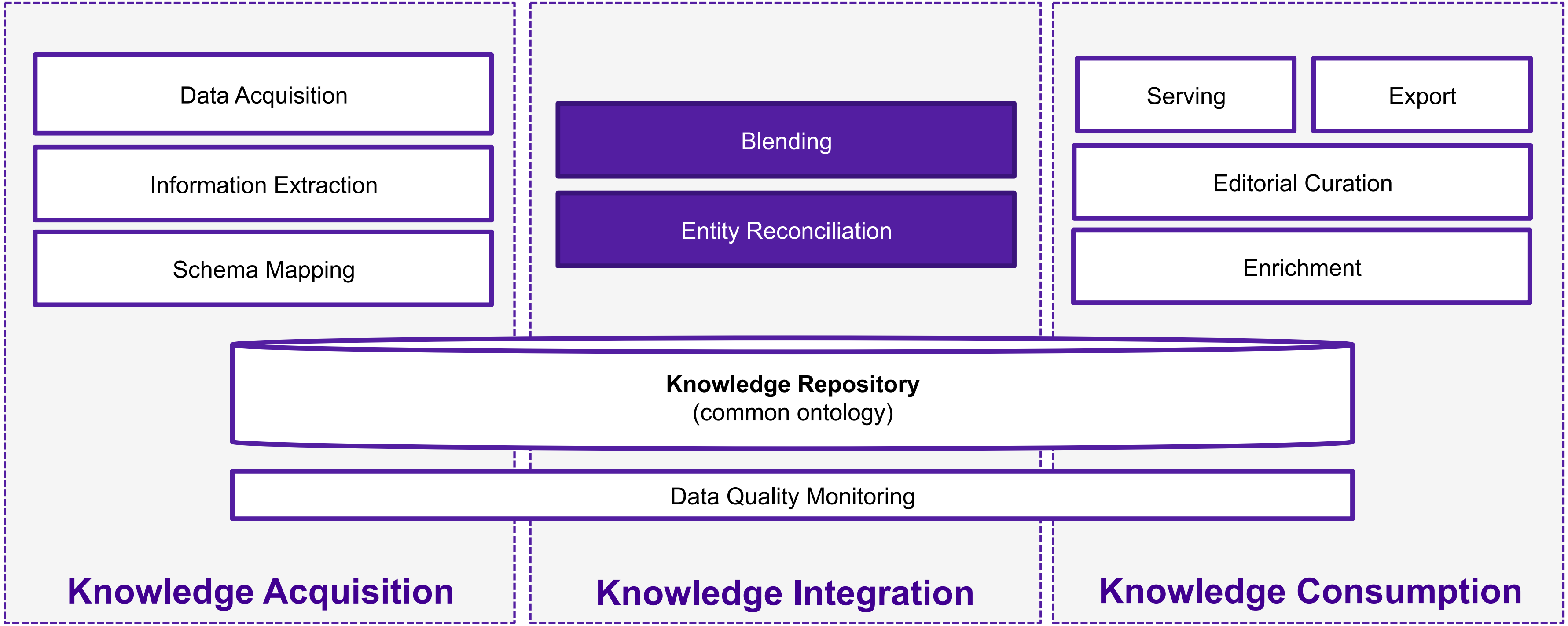
- **Present knowledge repository backed by a column-oriented store :-(**
  - Store de-normalized graph persistently and provide some random access via 2^{ndary} indices
  - Scale out nicely and smooth integration with Hadoop workflows
  - But simplistic data model and limited API make working with graph data tedious

- **Moving to a graph-oriented repository and workflow engine :-)**

  Challenges

  - Scale to 100s of millions of nodes and billions of facts? (processing, storage, retrieval)
  - Mix large record-oriented ETL workflows and distributed graph processing?
  - Efficient graph traversal and query? Built-in inference mechanism?
  - Schema-less? Data versioning?

Entity Reconciliation & Blending



Data Acquisition

Information Extraction

Schema Mapping

Blending

Entity Reconciliation

Serving

Export

Editorial Curation

Enrichment

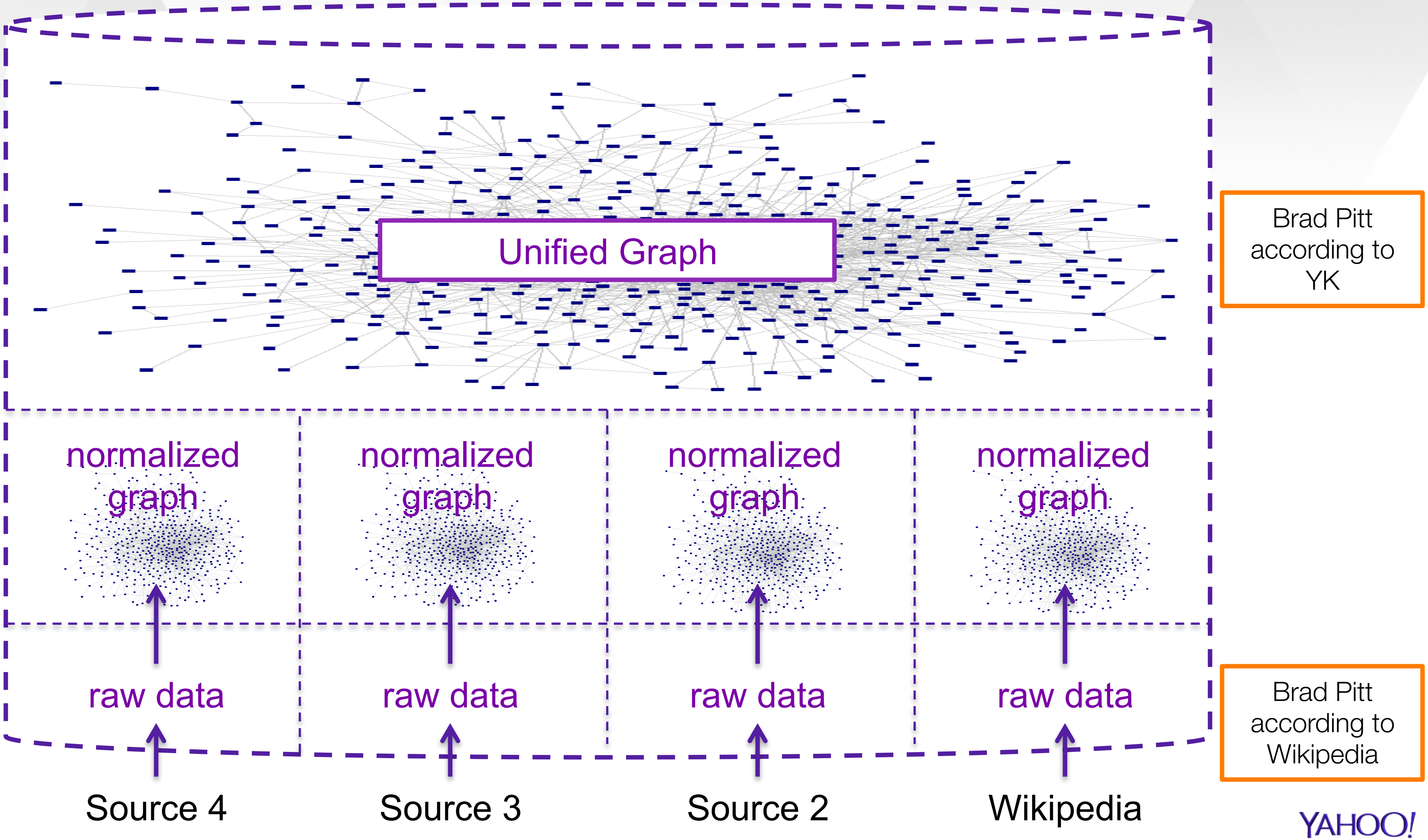
Knowledge Repository
(common ontology)

Data Quality Monitoring

Knowledge Acquisition

Knowledge Integration

Knowledge Consumption



Unified Graph

normalized graph

normalized graph

normalized graph

normalized graph

raw data

raw data

raw data

raw data

Source 4

Source 3

Source 2

Wikipedia

Brad Pitt according to YK

Brad Pitt according to Wikipedia

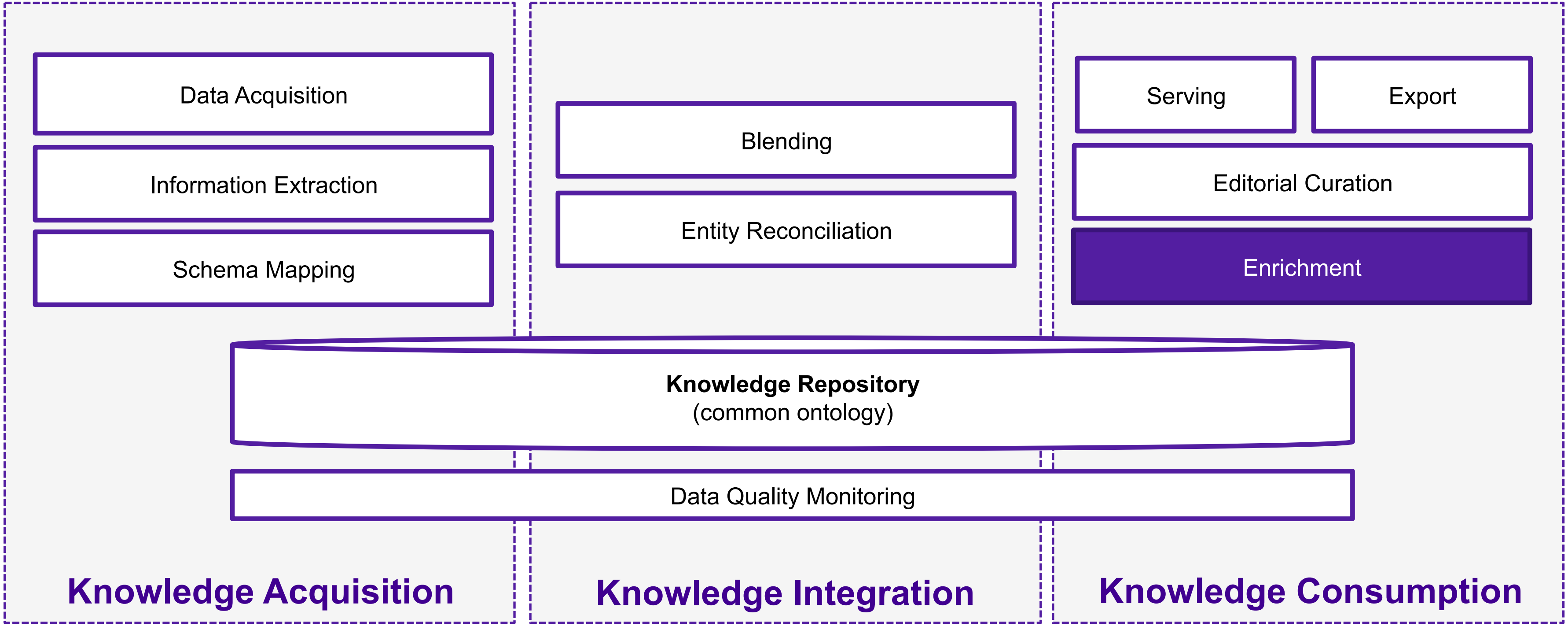
Entity Reconciliation & Blending

- **Disambiguate and merge entities across/within data sources**

Blocking	Select candidates most likely to refer to the same real world entity	Fast approximate similarity search Hashing techniques
Scoring	Compute similarity score between all pair of candidates	ML classifier or heuristics
Clustering	Decide which candidates refer to the same entity and interlink them	ML clustering or heuristics
Merging	Build a unified object for each cluster. Populate with *best* properties	ML selection or heuristics

- **Challenges**
  - Hard Science and Tech problems !
  - Scale and adapt to new entity types, data sources, data sizes, update frequencies…
  - Ongoing reconciliation/blending/evaluation. Need for consistent entity IDs. Provenance.

Enrichment


Data Acquisition
Information Extraction
Schema Mapping
Blending
Entity Reconciliation
Serving
Export
Editorial Curation
Enrichment
Knowledge Repository
(common ontology)
Data Quality Monitoring
Knowledge Acquisition
Knowledge Integration
Knowledge Consumption

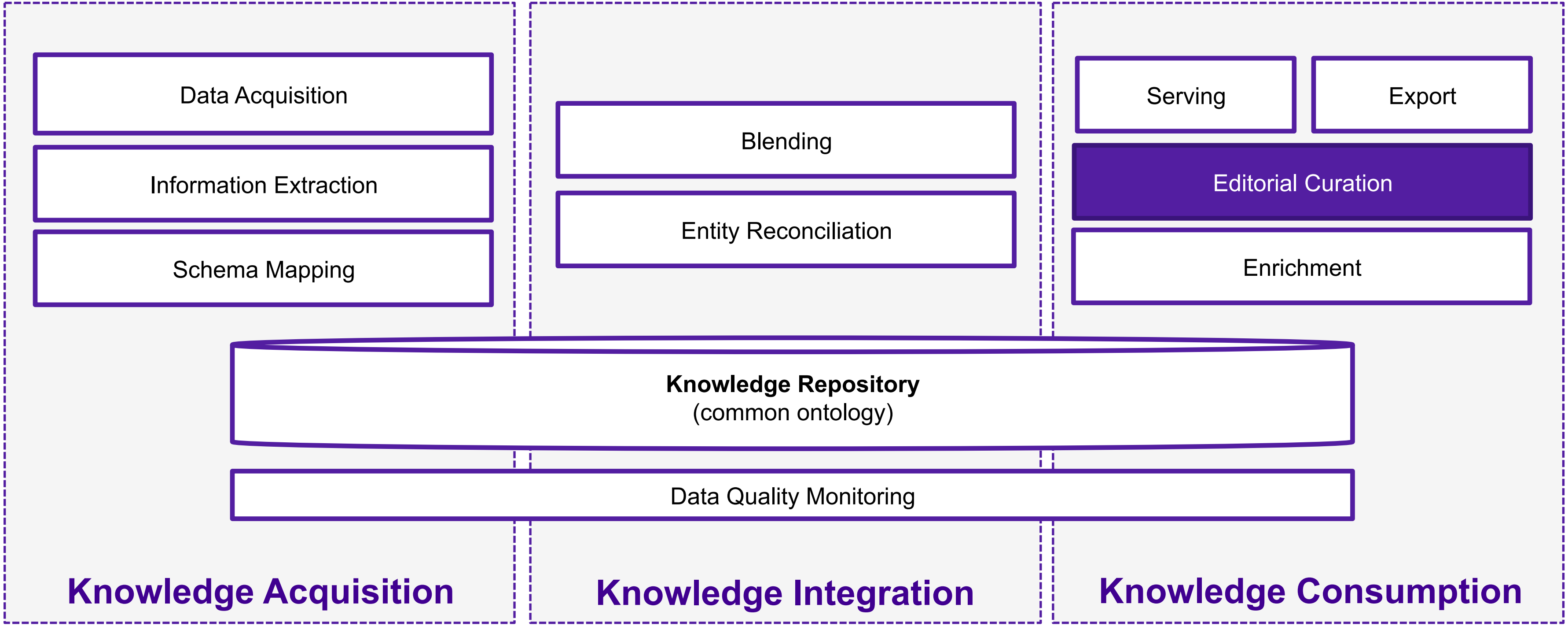
Enrichment

- **Enrich the graph with complementary and/or inferred information**
  - Generic enrichments vs. context-specific and application-specific enrichments

- **Examples:**
  - Entity description cleanup and summarization
  - Ranking of related entities
  - Entity categorization

- **Challenges**
  - Integrating, managing, and running a large number of, possibly conflicting, enrichers.

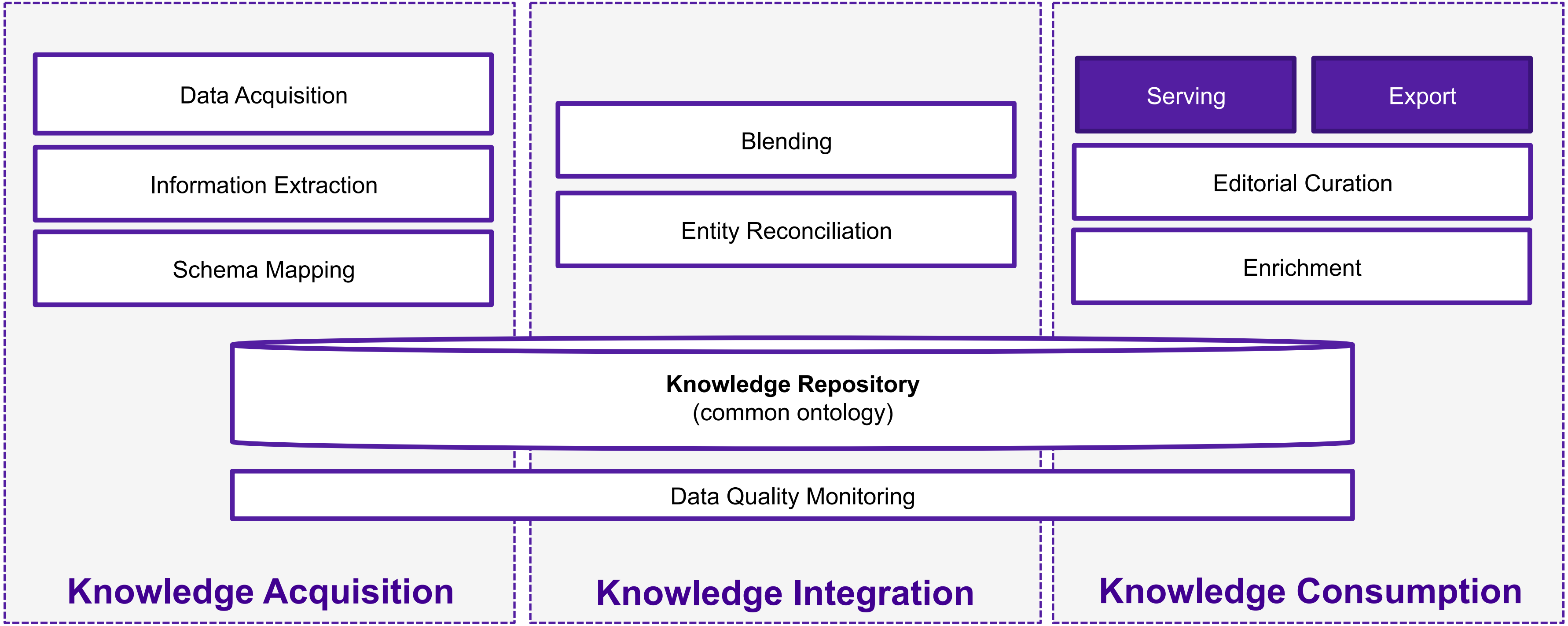
Editorial Curation



Editorial Curation

- **Enable editors to perform hot fixes**
  - Interactive (and safe) GUI for updating entities and associated information
- **Internal Wall of Shame**
  - Typical issues: incorrect/outdated facts, images, categorization (examples below)
  - Occasionally some reconciliation issues: Frankenstein objects!
- **Challenges**
  - Instantly reflect editorial updates in knowledge graph and consuming systems
  - Re-evaluate and manage editorial updates over time since they typically blindly overwrite
  - Manage multiple concurrent, and possibly conflicting, editorial updates.

Serving & Publishing



Data Acquisition

Information Extraction

Schema Mapping

Blending

Entity Reconciliation

Serving

Export

Editorial Curation

Enrichment

Knowledge Repository
(common ontology)

Data Quality Monitoring

Knowledge Acquisition

Knowledge Integration

Knowledge Consumption

Serving & Publishing

- **Online serving**
  - Dedicated serving infrastructure powering various online data APIs
  - Search layer provides efficient random access to the graph (and limited traversal)
  - Federation layer integrates transient info from connected services at query time
  - Customization layer provides attribute-level filtering, transformation, formatting

- **Datapack generation**
  - Regular datapack generation for offline batch consumption
  - Typically one single generic datapack with all the data

Knowledge-based services at Yahoo

- **Search:**
  - display, and search for, information about entities
- **Discovery:**
  - relate entities, interconnect data sources, link to relevant services
- **Understanding:**
  - recognize entities in queries and text

Yahoo Knowledge Graph

Making Knowledge Reusable

Thank you.

torzecn@yahoo-inc.com
Twitter: nicolastorzec