Interlinking Unstructured and Structured Knowledge in an Integrated Framework

Marco Rospocher

Fondazione Bruno Kessler,
Data and Knowledge Management Unit
Trento, Italy
rospocher@fbk.eu :: https://dkm.fbk.eu/rospocher

joint work with:

Francesco Corcoglioniti, Roldano Cattoni,
Bernardo Magnini, Luciano Serafini

Seventh IEEE International Conference on Semantic Computing (ICSC 2013), Irvine, CA, US

Introduction

- The rate of growth of digital data and information is nowadays continuously increasing
  - large amount of data and information is available in structured form
    - e.g., the Linked Data Initiative
  - a huge amount of content (>90% of the digital resources) is still available in an unstructured form
    - e.g., textual document, web pages, and multimedia material

Introduction

- Different format, but very similar content
  - they speak about entities of the world (e.g., PER, ORG, GPE/LOC, EVN), their properties, and relations among them
  - may contain coinciding, contradictory, and complementary facts about them
- Focusing on the content distributed in only one of these two forms may not be appropriate, especially in applications that require complete knowledge
  - e.g., decision making, question answering
- Frameworks enabling the seamless integration and linking of knowledge coming both from structured and unstructured content are still lacking.

Our Contribution

The Knowledge Store

- A framework enabling to jointly store, manage, retrieve, and semantically query, both unstructured ands structured content

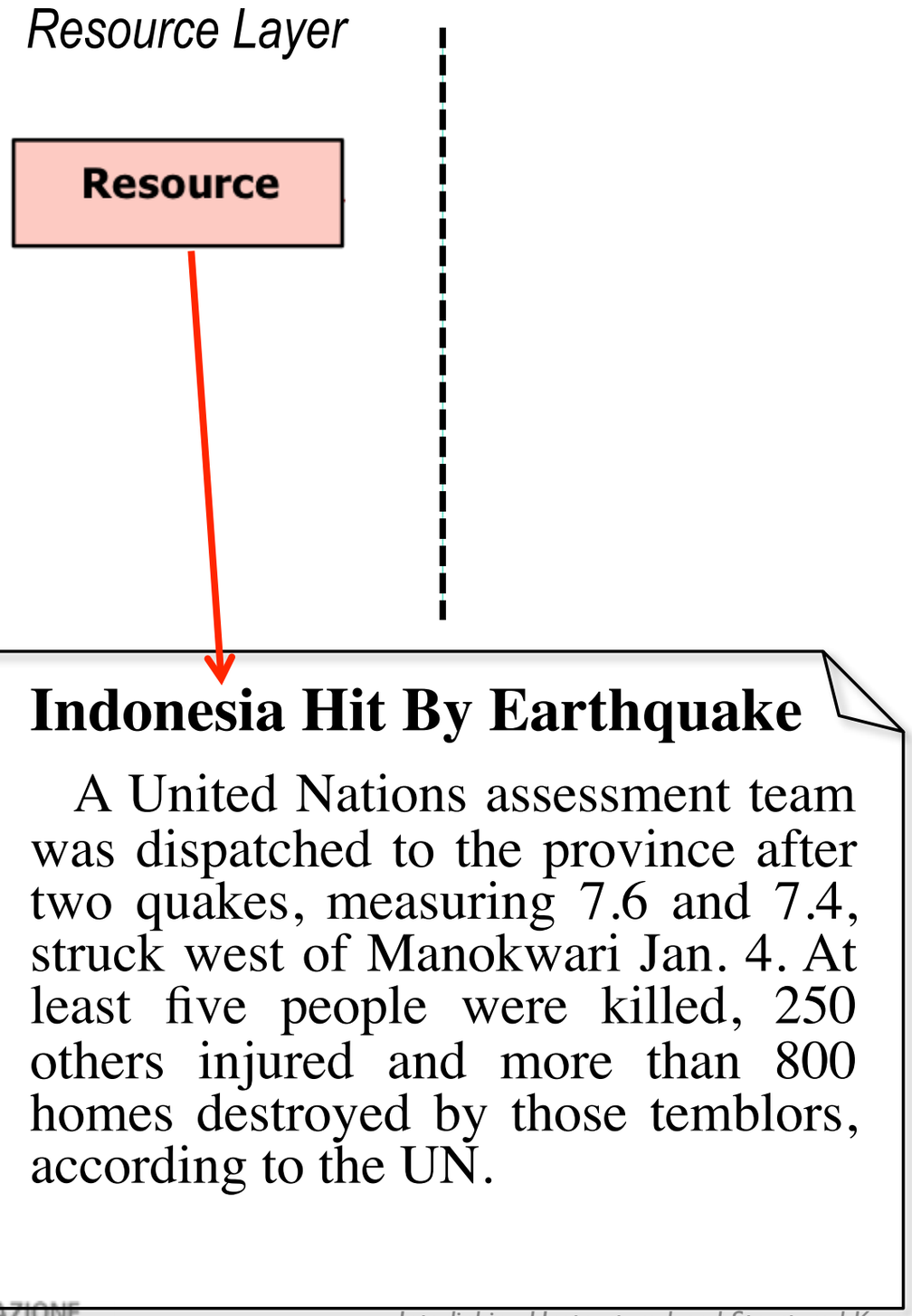
- A bridge between Natural Language Processing and Semantic Web

Our Contribution

The Knowledge Store

Our Contribution

The Knowledge Store



Our Contribution

The Knowledge Store

Resource Layer

Entity Layer

Resource

Entity

Statement

described by

Indonesia Hit By Earthquake

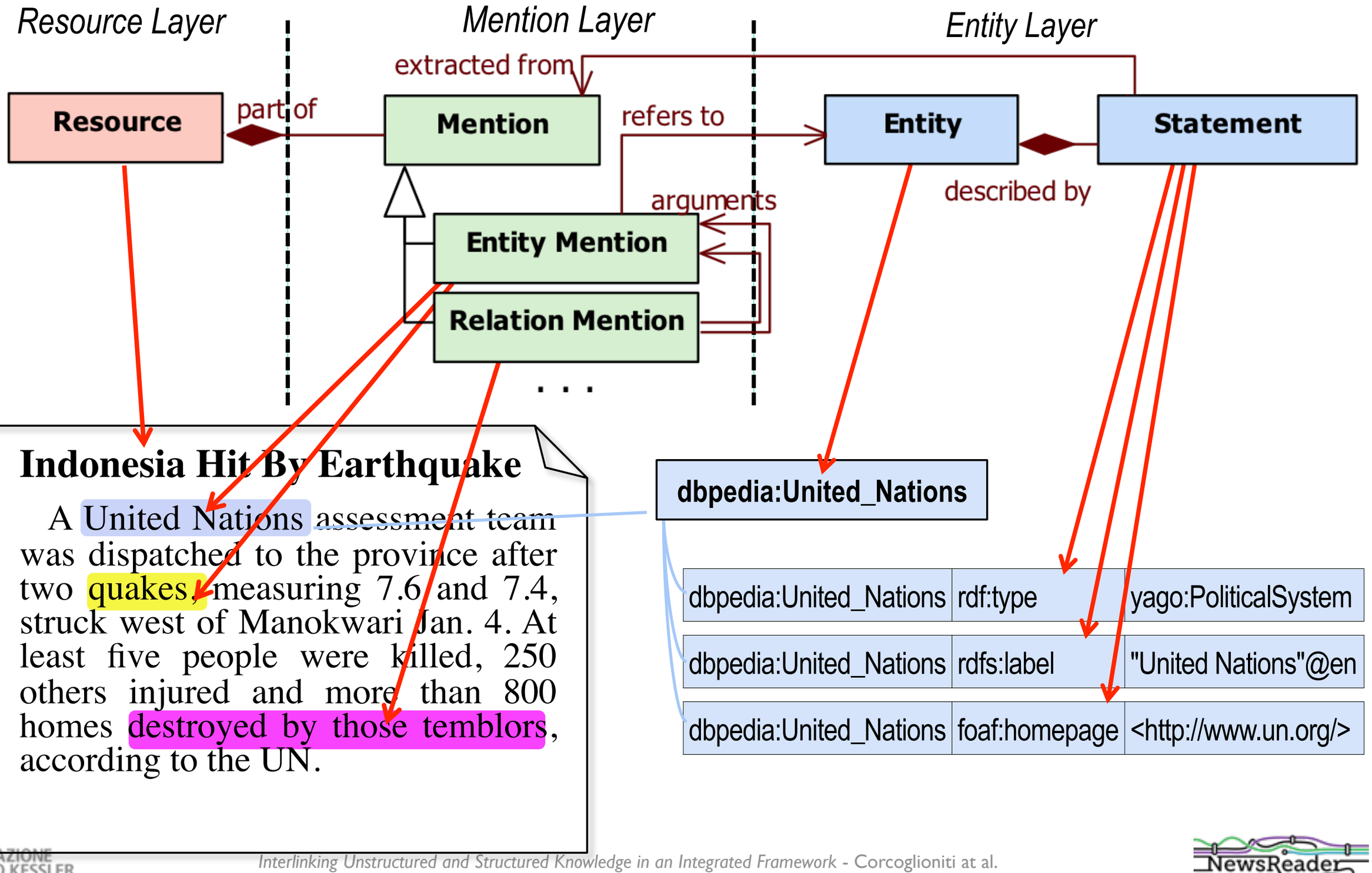
A United Nations assessment team was dispatched to the province after two quakes, measuring 7.6 and 7.4, struck west of Manokwari Jan. 4. At least five people were killed, 250 others injured and more than 800 homes destroyed by those temblors, according to the UN.

dbpedia:United_Nations

dbpedia:United_Nations	rdf:type	yago:PoliticalSystem
dbpedia:United_Nations	rdfs:label	"United Nations"@en
dbpedia:United_Nations	foaf:homepage	<http://www.un.org/>

Our Contribution

The Knowledge Store



Our Contribution

The Knowledge Store

- It can play a central role in applications/tasks that deals with both structured and structured knowledge
  - it enables effective decision making support: possibility to perform mixed queries
    - "retrieve all the documents mentioning that person Barack Obama participated to a sport event"
  - it favors the implementation and evaluation of tools improving the performance of coreference resolution tasks
  - it provides an ideal scenario for developing, training, and evaluating ontology population techniques
    - e.g. knowledge fusion, knowledge crystallization

Outline

- A concrete scenario: NewsReader
- The Knowledge Store
  - Data Model
  - Interfaces
  - Internal Architecture
- Preliminary Version
- Conclusions

A concrete scenario

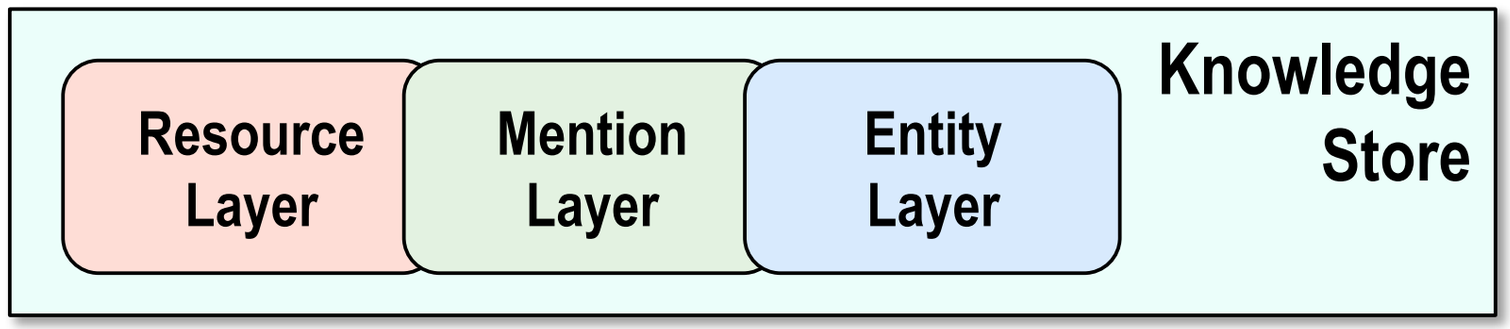
- EU ICT FP7 project [Jan 2013 - Dec 2015]
  - Partners: Netherlands (VU, LexisNexis, Synerscope), Spain (EHU), UK (ScraperWiki) and Italy (FBK)
  - http://www.newsreader-project.eu
- Automatically process massive streams of daily news from thousands of sources in 4 different languages to:
  - extract events (what happened, where, when and who is involved), and relations among them
  - organise and visualise events as narrative stories, combining new events with past events and background information, to provide more efficient access / decision support

Challenging Requirements

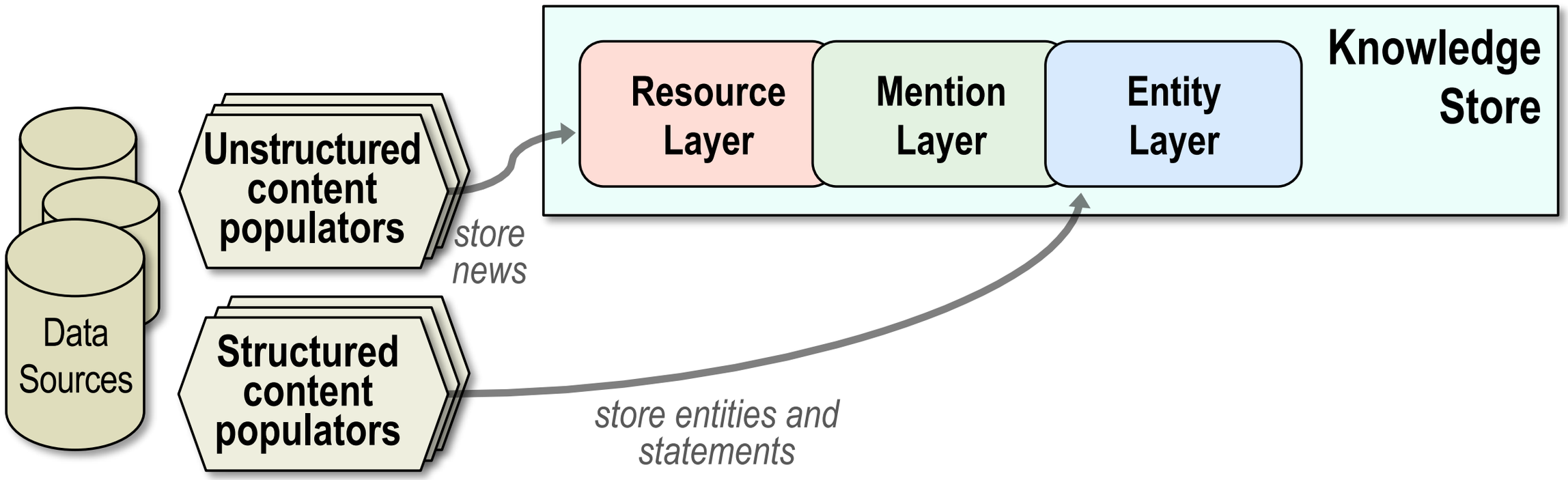
- To process document resources detecting mentions of events, event participants (e.g., PER, ORG), locations, time expressions, and so on
- To link mentions with entities, co-referring mentions of the same entity
- To complete entity descriptions, complementing extracted mention information with available structured knowledge (e.g., DBPedia)
- To interrelate entities (e.g., events and their participants) to support the construction of narrative stories
- To reason over events to check consistency, completeness, factuality and relevance
- To store all this huge quantity of information in a scalable way enabling efficient retrieval and intelligent queries
- To effectively offer narrative stories to decision makers

Role of the KS

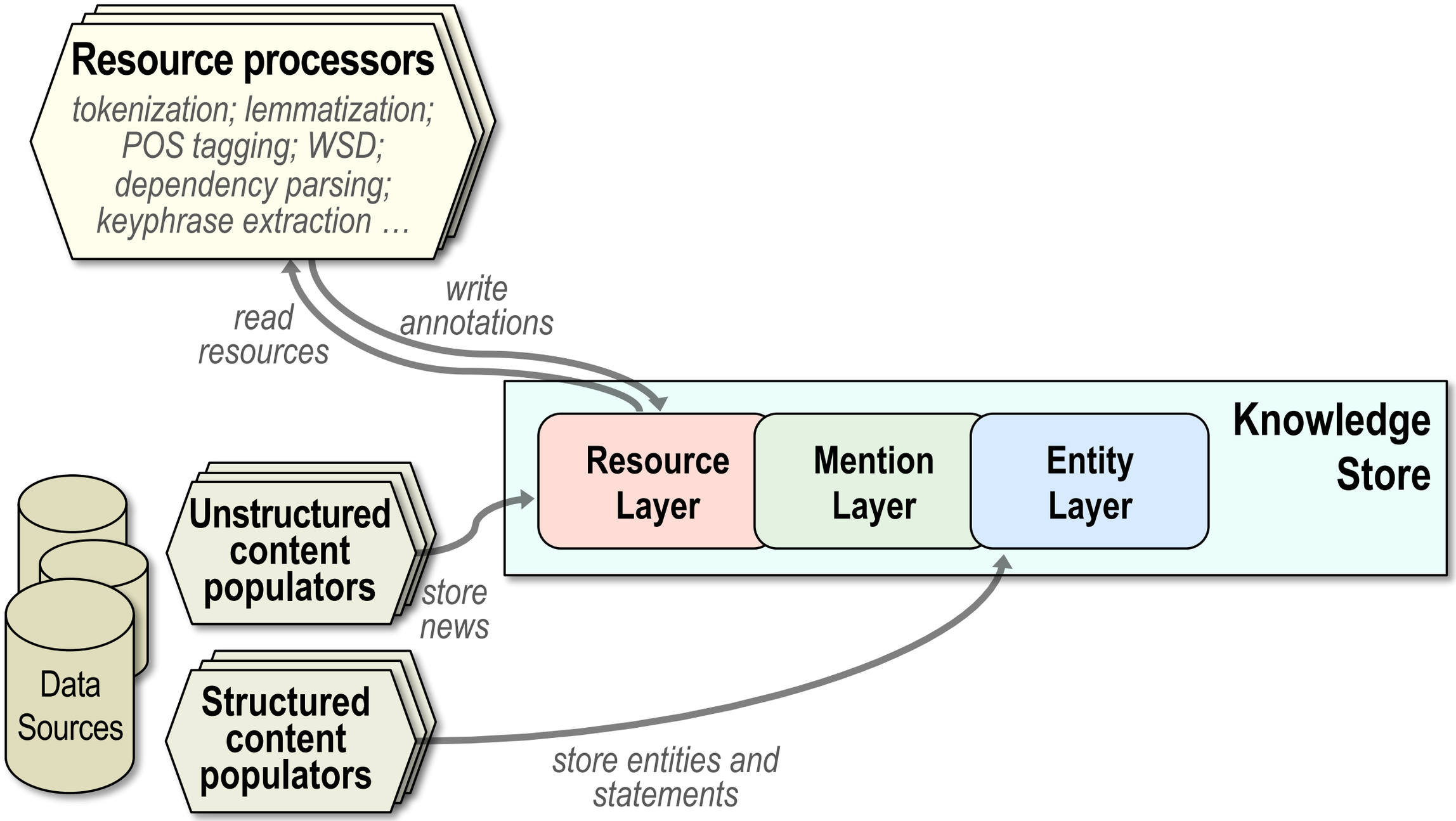
Role of the KS



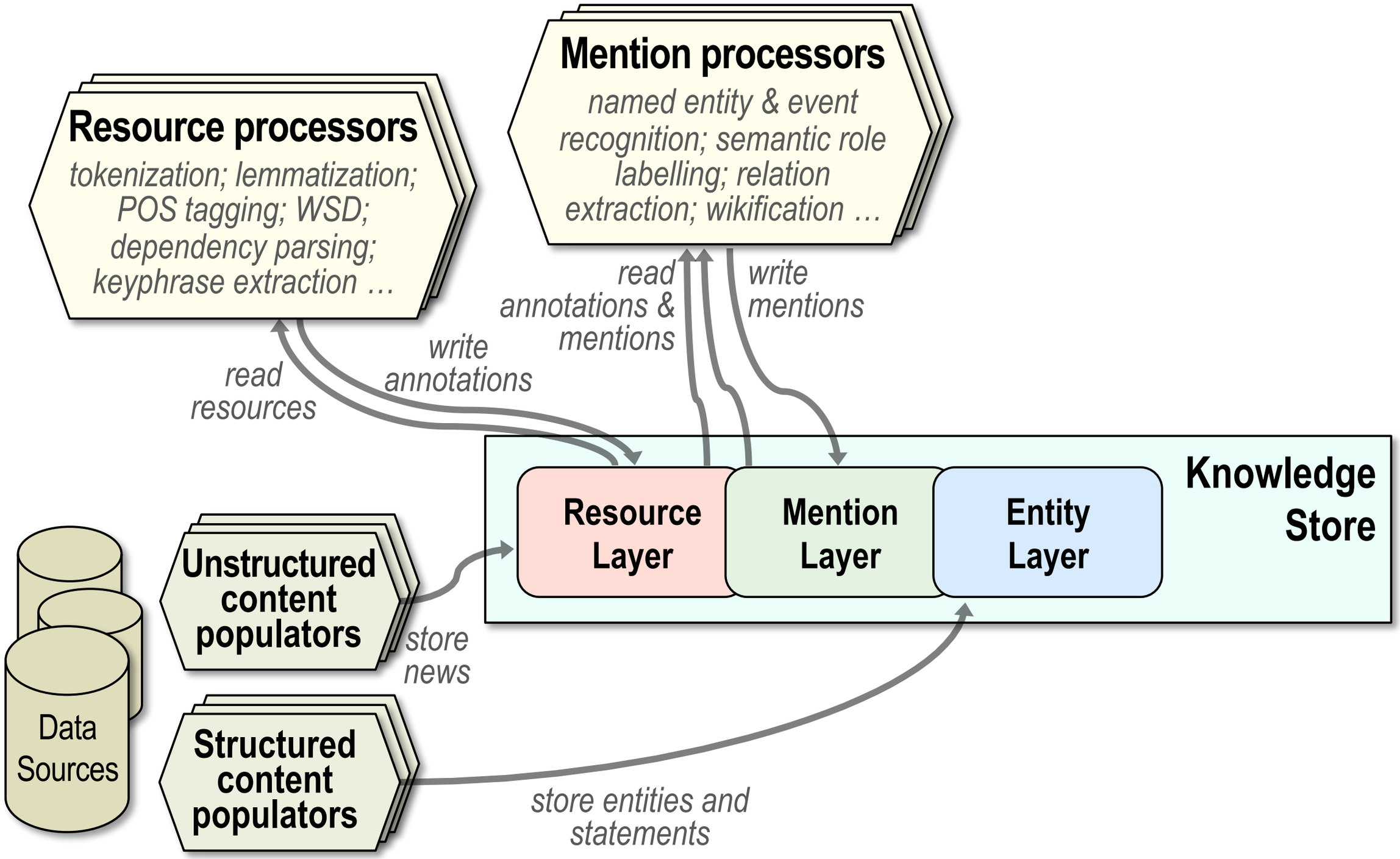
Role of the KS



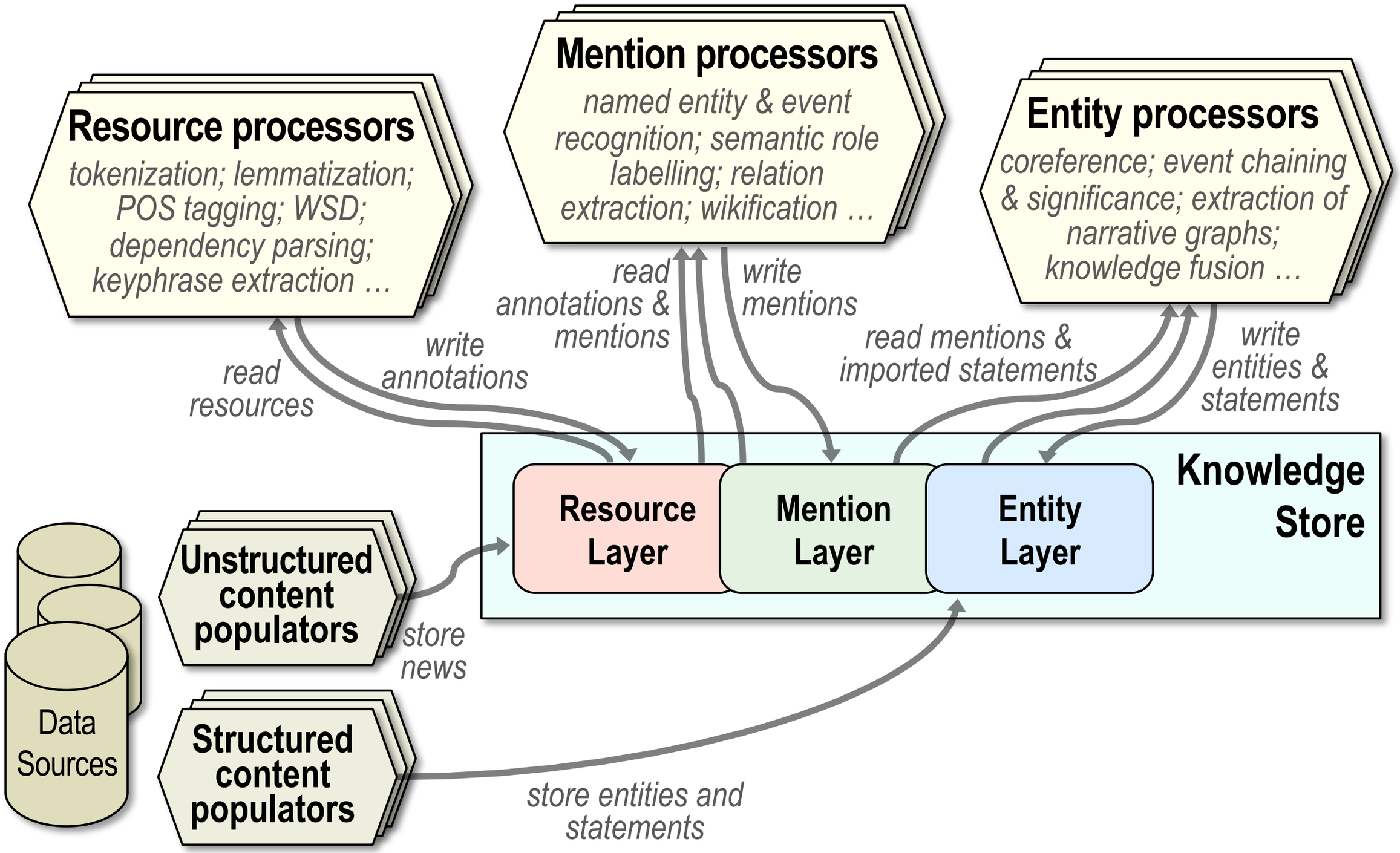
Role of the KS



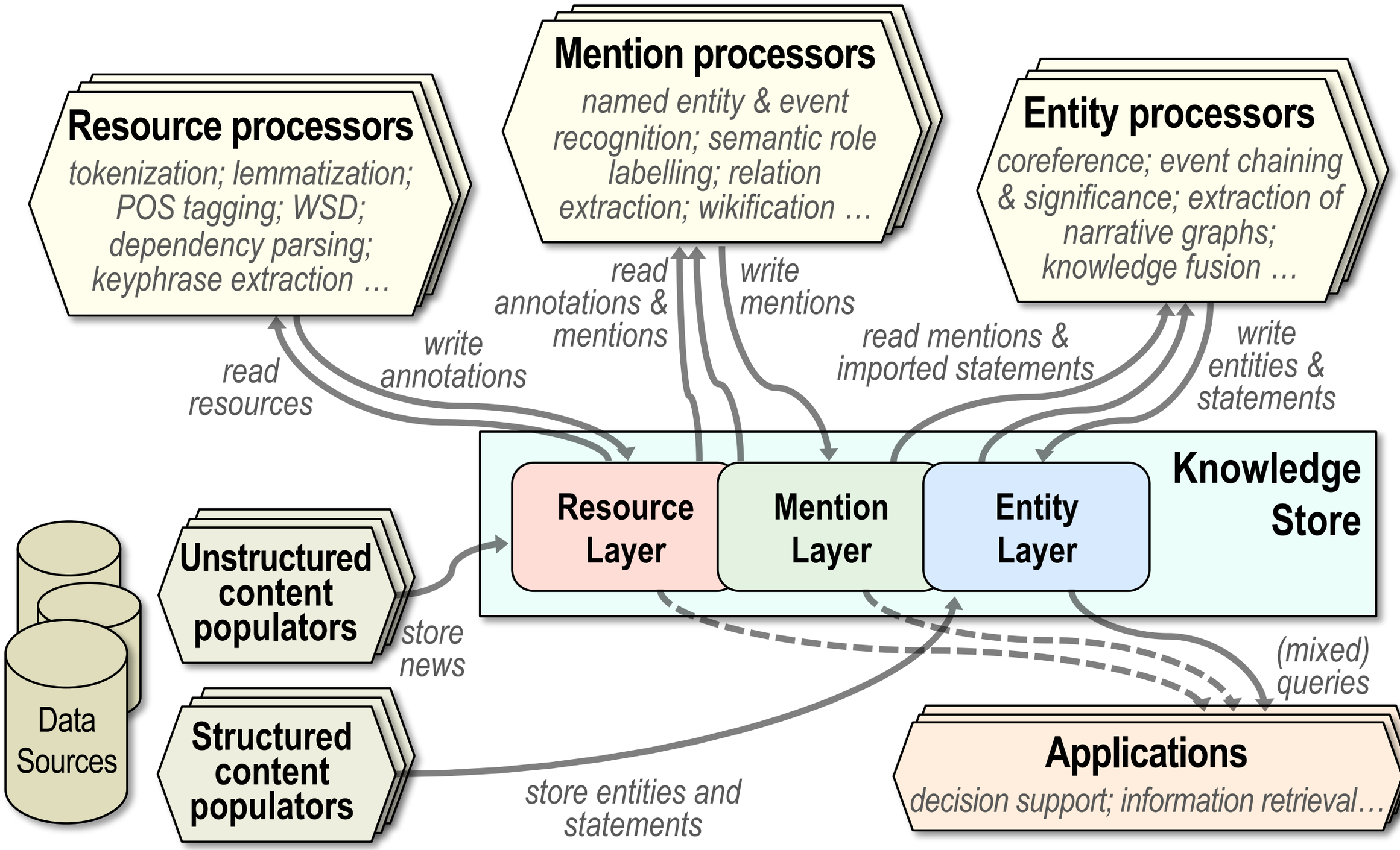
Role of the KS



Role of the KS

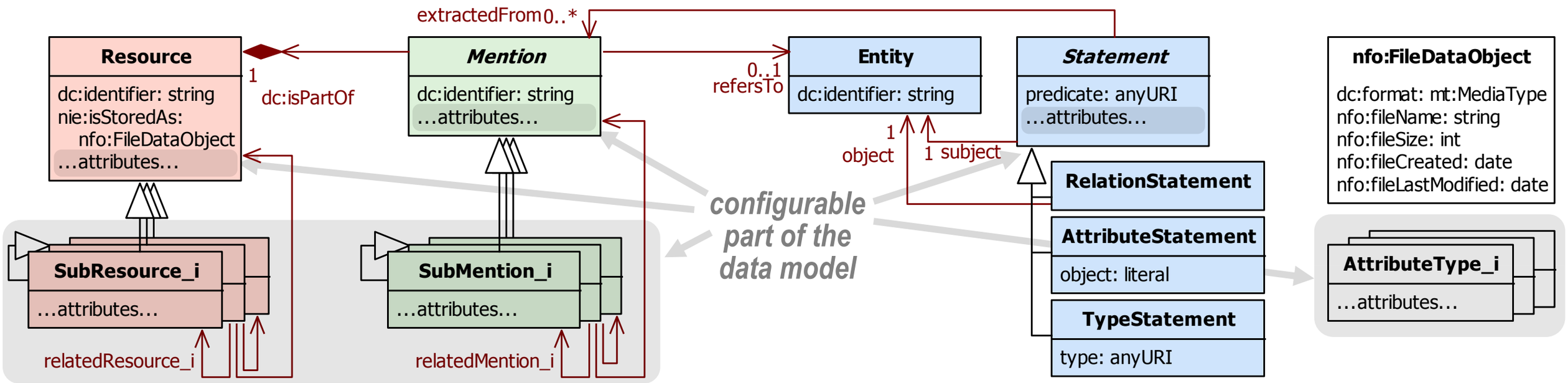


Role of the KS



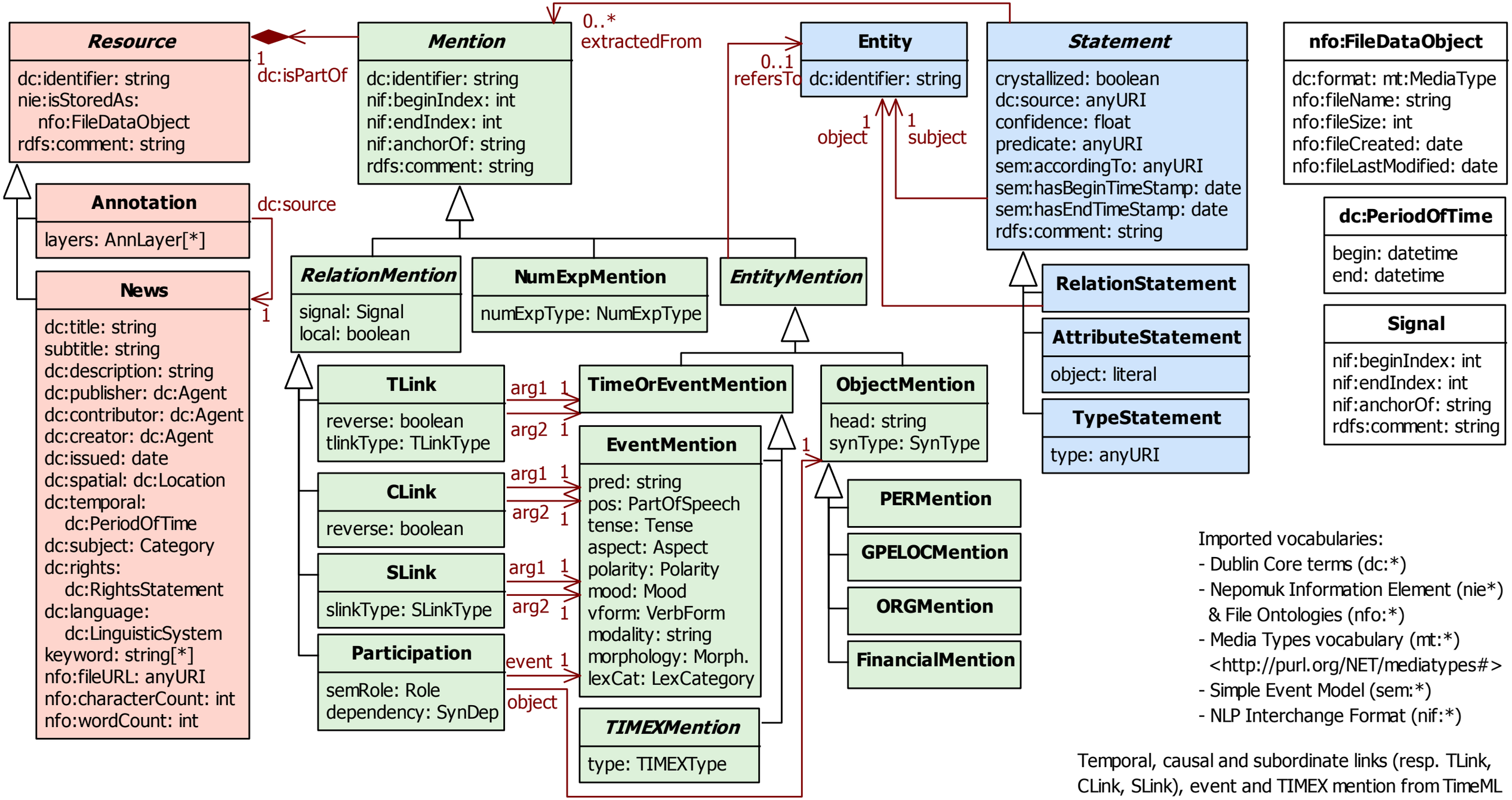
The Knowledge Store Data Model

Achieving Flexibility



The Knowledge Store Data Model

Specialization for NewsReader



Imported vocabularies:
- Dublin Core terms (dc:*)
- Nepomuk Information Element (nie*) & File Ontologies (nfo:*)
- Media Types vocabulary (mt:*) <http://purl.org/NET/mediatypes#>
- Simple Event Model (sem:*)
- NLP Interchange Format (nif:*)

Temporal, causal and subordinate links (resp. TLink, CLink, SLink), event and TIMEX mention from TimeML

The Knowledge Store Data Model

Remarks

- Data model grounded in OWL 2
  - it allows sharing stored data on the Semantic Web, e.g., by publishing it as Linked Open Data
  - inference and data validation may be performed using an OWL 2 reasoner
    - OWA and UNA to be considered!

- Available @
  - https://dkm.fbk.eu/ontologies/knowledgestore.html
  - https://dkm.fbk.eu/ontologies/newsreader.html

The Knowledge Store Interfaces

- Definition of interfaces by involving potential users (NLP, KR, DS)
  - fill in template describing possible operations
- Post-processing of collected operations to find commonalities and to further generalize them
- Organized in three main categories
  - CRUD operations
  - Intra-layer operations
  - Inter-layer operations

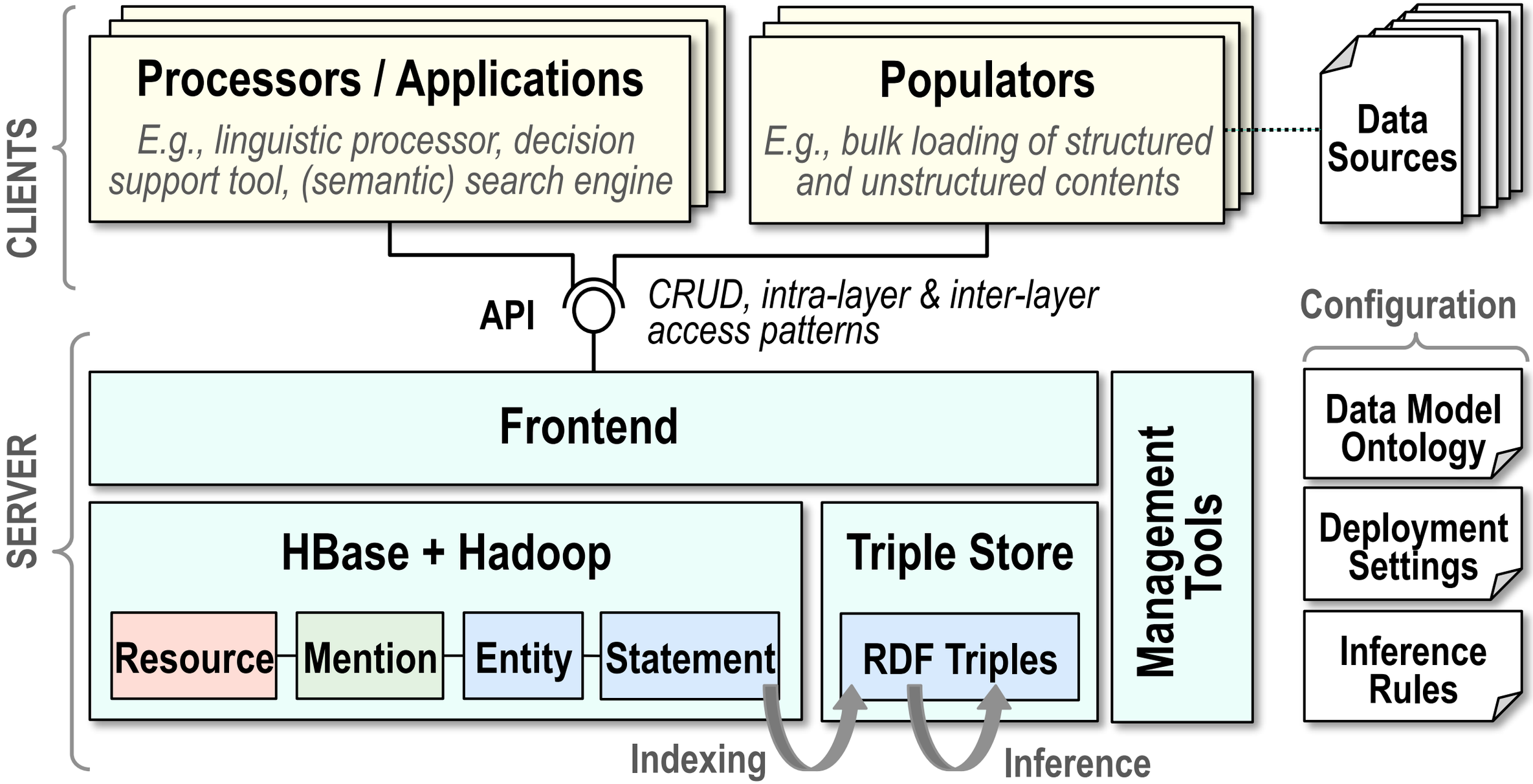
The Knowledge Store Interfaces

Example of inter-layer operation

`name`	getResourcesFromEntity()
`description`	given an entity, return a list of news in which it is mentioned
`input`	the entity URI
`output`	a list with essential info for each matching resource (e.g., id, title and date)
`notes`	an optional parameter may determine the amount of info to be returned
`example`	get resources mentioning entity `nwr:E105`

The Knowledge Store Architecture

Overview



The Knowledge Store Architecture

HBase & Hadoop

- Primary storage for the Knowledge Store
- Hadoop
  - distributed (partitioned and replicated) file system, used to store the unstructured content (e.g., news texts).
- HBase
  - column oriented NoSQL database, used to store the structured content
  - three main tables: resources, mentions and statements
  - Redundant tables and schema de-normalization are employed to avoid expensive join operations

The Knowledge Store Architecture

Triple Store

- Statements are partially indexed in a triple store in order to enable efficient, inference-aware query answering
  - exclude from inference statements whose extraction confidence level is below a given threshold
- Stored as a ⟨subject, predicate, object⟩ triple within a named graph
  - context where the statement holds
- Accessible via SPARQL queries
- Reasoning based on closure materialization and custom rule-based inference
- Abstracting the actual triple store implementation by means of the OpenRDF Sesame Java API
- Current choice: Open Source Edition of the Virtuoso
  - excellent performances in recent benchmarks (April 2013)

The Knowledge Store Architecture

Front End

- Implements the external API of the KnowledgeStore by dispatching client requests to other components
  - majority of API operations is forwarded to a single component

- Mixed queries decomposed into:
  - one or more semantic queries, targeted at the triple store,
  - one or more retrieval operation for structured and unstructured data in Hadoop/Hbase
  - Example: all news mentioning that "Barack Obama" participated to a sport event

- Replicated to avoid single points of failure

Preliminary Version

LiveMemories
Active Digital Memories of Collective Life

- Tested in the scope of the LiveMemories project
  - http://www.livememories.org
- Limitations
  - no storing of and reasoning on events and related information
  - no triple store / semantic queries mechanism
- Some stats:
  - Resources: ~800K (~56 GB) of textual news, images and videos in Italian language
  - Mentions: ~12M
  - Entities: ~420K

Conclusions

- We presented the KnowledgeStore:
  - a framework enabling to jointly store, manage, retrieve, and semantically query, both unstructured and structured content
  - enables the development of enhanced applications, and favors the design and empirical investigation of several information processing task
- Implementation is on-going
  - first complete prototype planned for Dec 2013
- We plan to validate the KnowledgeStore idea in NewsReader
  - functional evaluation
    - store an overwhelming daily stream of economical and financial contents
    - support a complex NLP pipeline in extracting knowledge
    - provide suitable online and offline query capabilities
  - performance evaluation
    - scalability with respect to data size, query load
    - tolerance to nodes and network failures

Thank you! Questions?

Marco Rospocher

Fondazione Bruno Kessler,
Data and Knowledge Management Unit
Trento, Italy
rospocher@fbk.eu :: https://dkm.fbk.eu/rospocher