7BDIN006W
Big Data Theory and Practice

Lecture 8

Big Data and Metadata

UNIVERSITY OF WESTMINSTER

Big Data and Metadata

Metadata, often described as "data about data," encompasses information that provides context and attributes to datasets. This includes data source, creation date, format, and authorship.

Example:

Consider a dataset of *customer reviews for an e-commerce platform*.

Metadata could include the date *when each review was posted*, *the product reviewed*, and *the user who wrote it*.

This metadata helps in organising and searching for reviews based on various criteria.

Semantics in Big Data

Semantics in the context of Big Data refers to the meaning and interpretation of data beyond its structure or syntax.

Example:

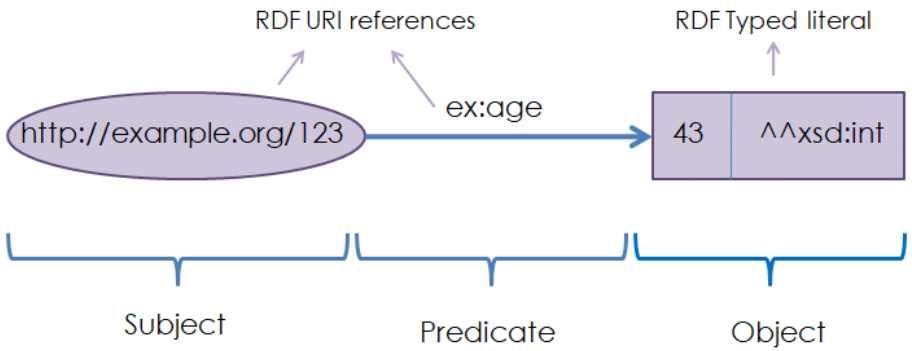
When you search for "*apple*," the engine understands whether you're looking for information about the *fruit* or the *technology company*.

This understanding is achieved through semantic analysis, allowing the system to provide relevant search results.

Triples and RDF

RDF (Resource Description Framework) is a framework for describing resources on the web using triples.

A triple consists of three parts: subject, predicate, and object.



Example:

Consider the triple: "Albert Einstein (subject) was born in (predicate) Ulm (object)."

Here, RDF triples facilitate the structured representation of information on the web, allowing machines to understand and link data efficiently.

RDF Syntaxes

RDF data can be expressed in different syntaxes, including

- RDF/XML
- Turtle
- JSON-LD
- N-Triples

These syntaxes provide various ways to write RDF data while adhering to the RDF data model.

Ontologies

Ontology is a formal, explicit specification of concepts, entities, and their relationships within a specific domain or knowledge area. It defines a shared vocabulary to represent and structure knowledge, making it understandable by *both humans and machines*.

In the context of Big Data, ontologies serve several purposes:

- **Data Integration**: Ontologies help integrate heterogeneous data sources by providing a common framework for data representation. They bridge the gap between different data formats and structures, making data interoperable.
- **Semantic Enrichment**: Ontologies add semantic meaning to data. They allow Big Data systems to understand the context and relationships within the data, enabling more advanced queries and analytics.
- **Data Quality and Consistency**: Ontologies help maintain data quality by standardising terminology and ensuring consistent data representation. This reduces errors and enhances data reliability.
- **Knowledge Discovery**: Ontologies facilitate knowledge discovery by enabling automated reasoning and inference. They assist in uncovering hidden patterns and insights within large datasets.

Examples of Existing Ontologies

1. **Financial Industry Business Ontology (FIBO)**:
    1. You can visit the official website for the EDM Council, which is the organization that manages FIBO. Search for "EDM Council FIBO" in your preferred search engine.
2. **Gene Ontology (GO)**:
    1. The official website for the Gene Ontology Consortium can be found by searching for "Gene Ontology Consortium" or "GO Consortium" in your preferred search engine.
3. **SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms)**:
    1. To access information about SNOMED CT, visit the official website of SNOMED International. You can search for "SNOMED International" to find their website.
4. **Friend of a Friend (FOAF)**:
    1. Information about the FOAF ontology can be found in various Semantic Web and RDF-related resources. You can search for "FOAF ontology" or "Friend of a Friend RDF" to find relevant information.
5. **GeoNames Ontology**:
    1. The GeoNames Ontology is associated with the GeoNames geographical database. Visit the GeoNames website at geonames.org to explore the ontology and related resources.

Protégé

Protege is an open-source platform and tool for developing and managing ontologies. It provides a user-friendly interface and a suite of features for creating, editing, and visualising ontologies.

Protege supports various ontology languages, including RDF, JSON and OWL (Web Ontology Language), and is widely used in academia and industry for ontology development.

Protege is valuable in the Big Data context as it allows data scientists, researchers, and domain experts to create ontologies that can be applied to large and complex datasets. Its purposes in the Big Data domain include:

1. **Ontology Development**: Protege simplifies the process of creating ontologies, making them accessible to a broader audience.
2. **Ontology Management**: It provides tools for organising and maintaining ontologies, which is crucial when dealing with evolving Big Data environments.
3. **Data Integration**: Protégé assists in creating ontologies that can integrate and harmonise diverse data sources.
4. **Semantic Analysis**: It enables semantic analysis of Big Data by allowing users to define and apply ontological reasoning rules to extract meaningful information.

Protégé Demonstration

Let’s consider
a simplified Ontology of Musical Instruments…