DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

Part 1: Introduction

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF OXFORD

Semantic Technology Tutorial

Part 1: Introduction

Kinds of Data in Modern Applications

- Unstructured data:
  - Text organised in documents (sometimes multimedia files)
  - No schema for the data
  - No deep understanding of data → process data using statistic methods
    - Bag-of-words
    - Inverted index

- Relational data:
  - Structured according to a well-defined schema
    - Describes the kinds of entities and their relationships
    - Developed in advance → schema-first
    - Expected to be mostly stable
  - Data processing via queries
    - Often fixed ways of accessing data
    - Schema determines the meaning of query results
  - Schema used for optimising data access
  - Application closely depends on the schema

Enter Semistructured Data

- Data structure is not known (completely) in advance
- Schema is dynamic, sometimes ad hoc
- Rich structure: many different kinds of relationships
- Data processing via:
  - Querying → as in relational model; fixed access patterns
  - Exploration → ‘ask a query and see what is returned’
- Self-explanatory models → may use data without knowing all structure
- Applications are not tightly linked to a schema

Semistructured vs. Relational Data

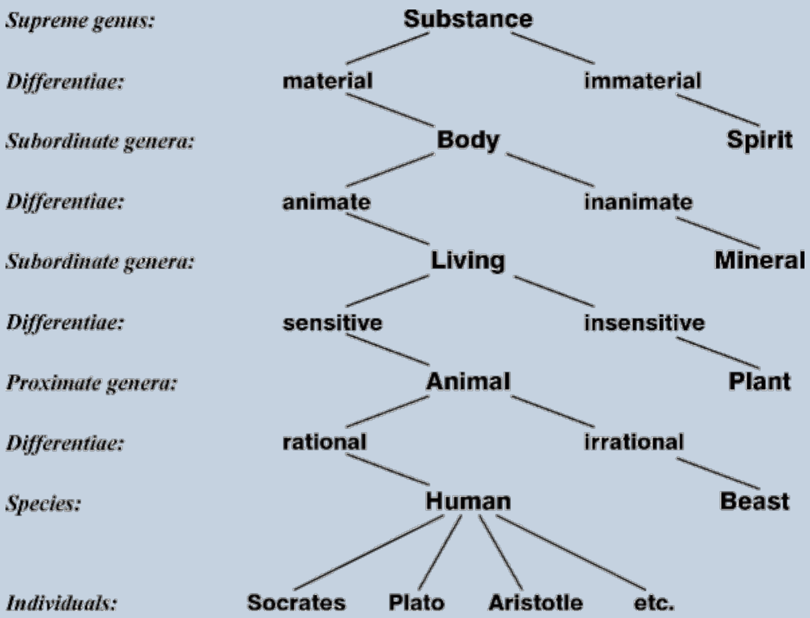
- No strict distinction→many benefits are ‘soft’
- One can often:
  - Embed semistructured into relational models
  - Use relational technology to manage semistructured data
- Key ‘soft’ aspect: flexibility
  - Extending/modifying the schema is ‘easier’
  - Ad hoc querying: systems should efficiently handle any ‘reasonable’ query
  - No distinction between querying schema and querying data
- Common use case: data integration
  - Flexibility needed to represent data from many sources

What is an Ontology?

What is an Ontology?

A fundamental branch of **metaphysics**

- Studies “being” or “existence” and their **basic categories**
- Aims to find out what **entities** and **types of entities** exist



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

DBOnto

SIRIUS

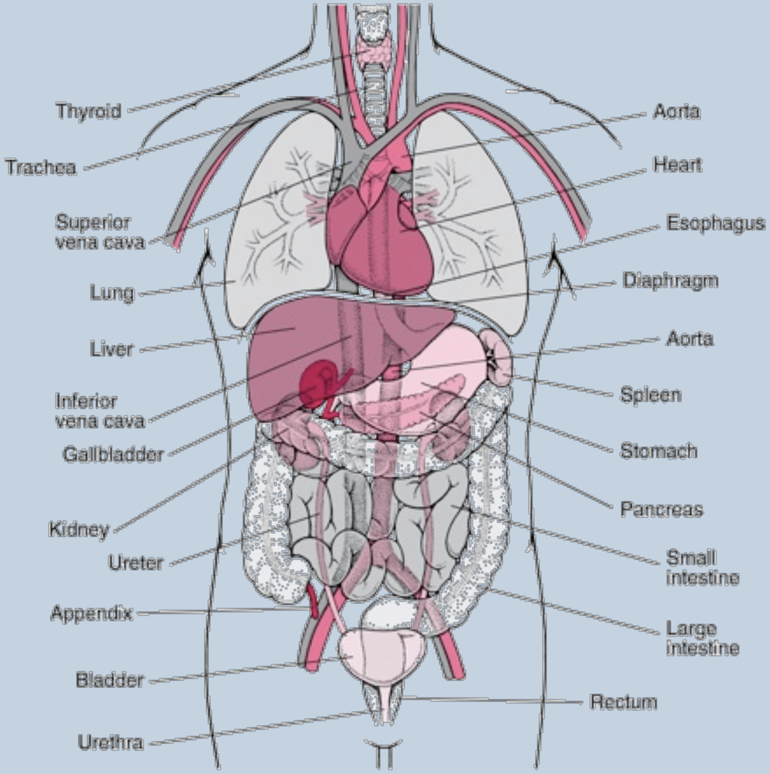
What is an Ontology?

A **model** of (some aspect of) the world

What is an Ontology?

A **model** of (some aspect of) the world

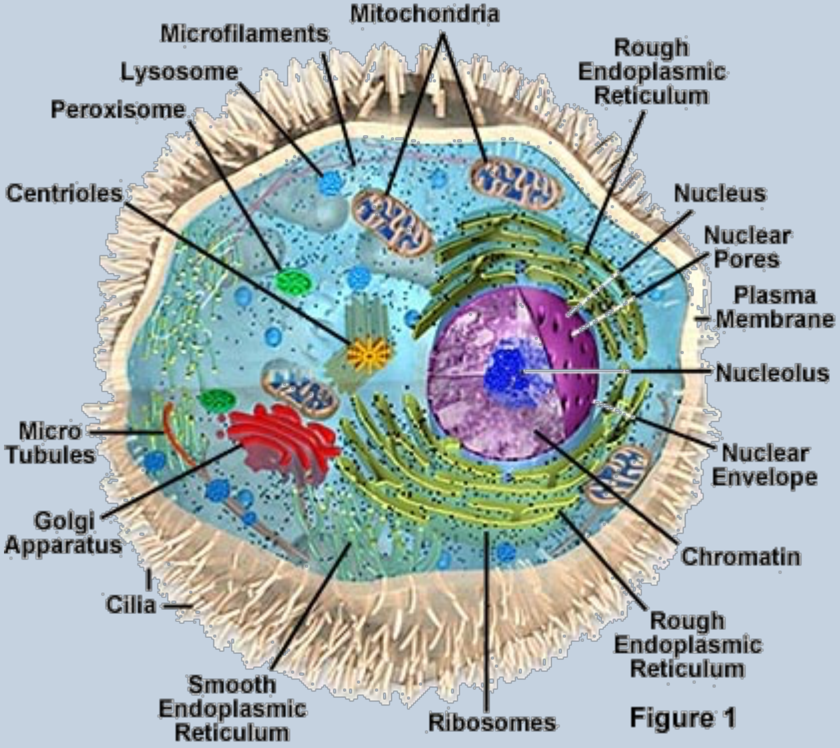
- Introduces **vocabulary** relevant to domain, e.g.:
  - Anatomy



What is an Ontology?

A **model** of (some aspect of) the world

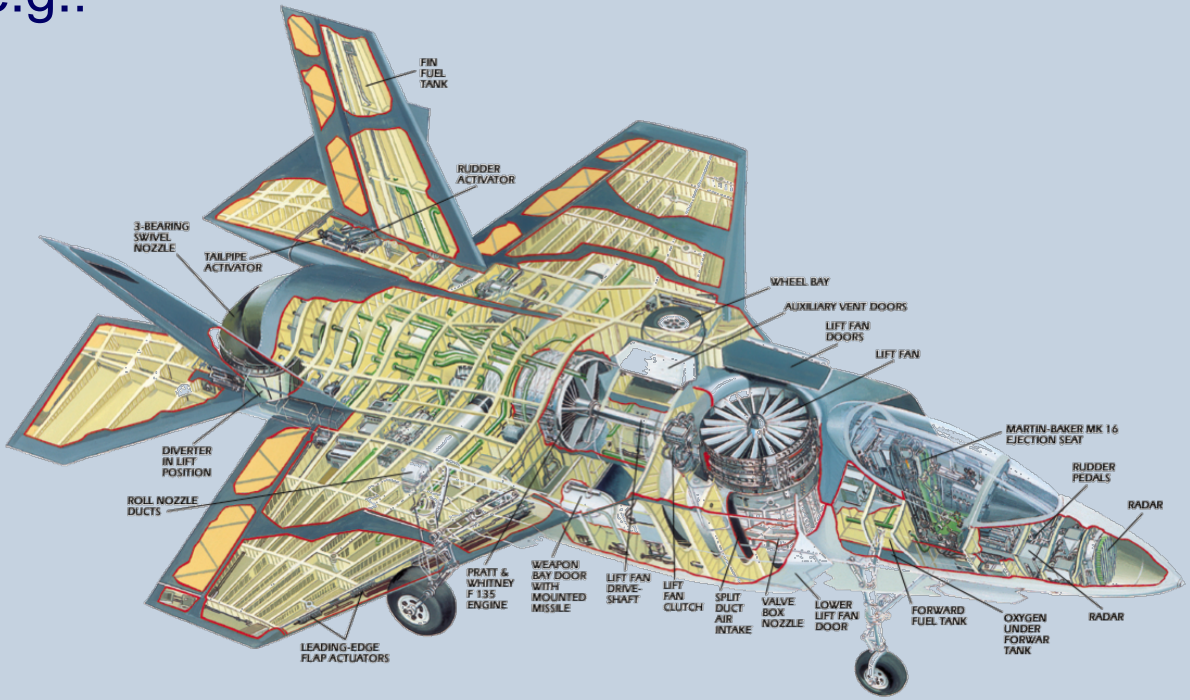
- Introduces **vocabulary** relevant to domain, e.g.:
  - Anatomy
  - Cellular biology



What is an Ontology?

A **model** of (some aspect of) the world

- Introduces **vocabulary** relevant to domain, e.g.:
  - Anatomy
  - Cellular biology
  - Aerospace



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

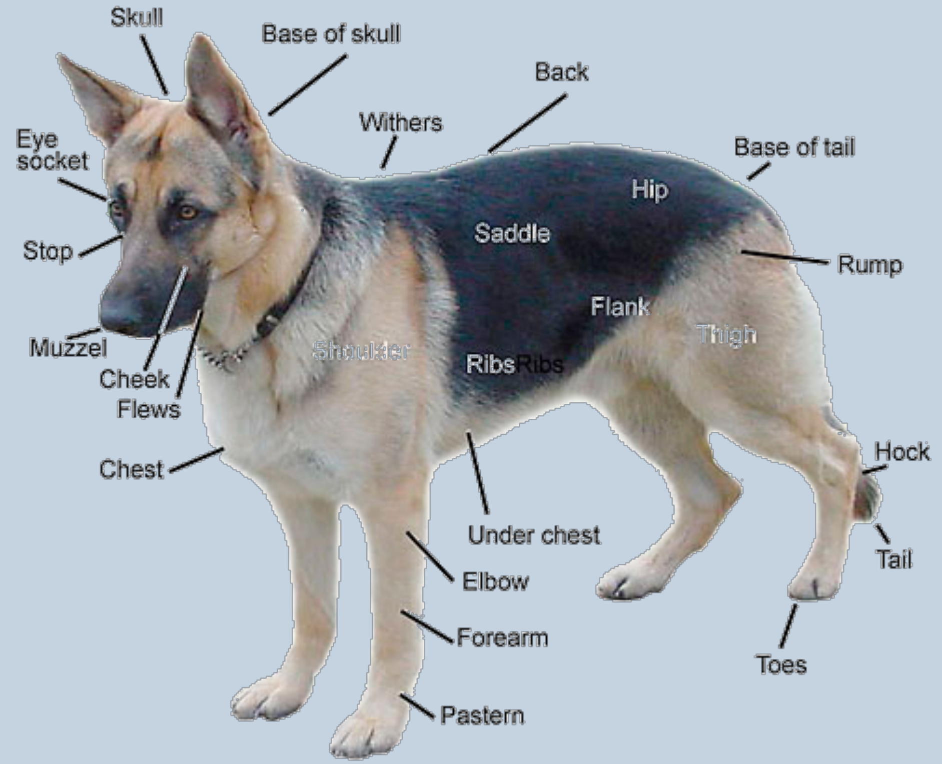
DBOnto

SIRIUS

What is an Ontology?

A **model** of (some aspect of) the world

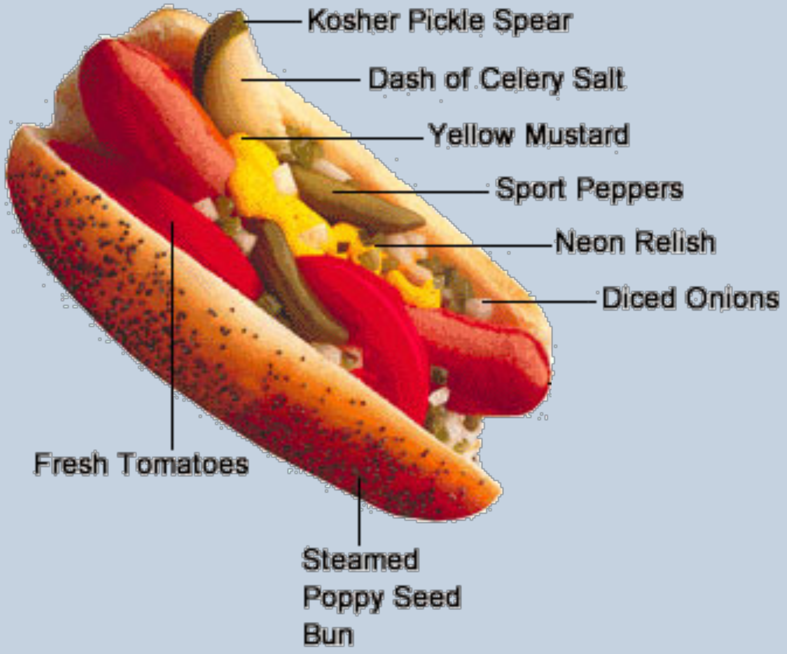
- Introduces **vocabulary** relevant to domain, e.g.:
  - Anatomy
  - Cellular biology
  - Aerospace
  - Dogs



What is an Ontology?

A **model** of (some aspect of) the world

- Introduces **vocabulary** relevant to domain, e.g.:
  - Anatomy
  - Cellular biology
  - Aerospace
  - Dogs
  - Hotdogs
  - …

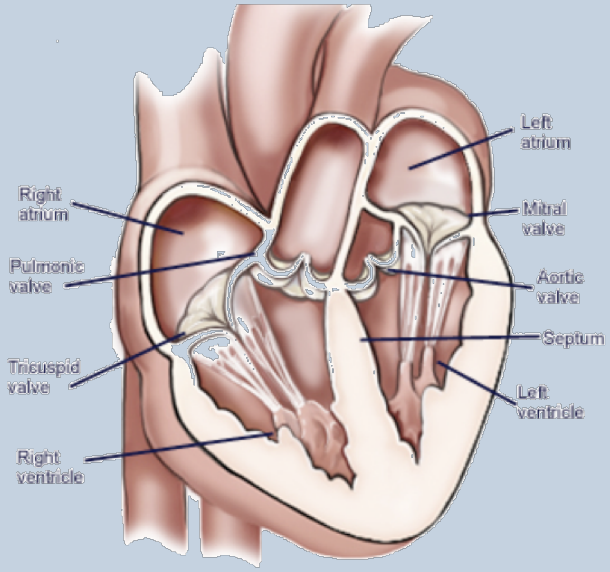


What is an Ontology?

A **model** of (some aspect of) the world

- Introduces **vocabulary** relevant to domain
- Specifies (relative) **semantics** of terms

Heart is a muscular organ that
is part of the circulatory system



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

DBOnto

SIRIUS

What is an Ontology?

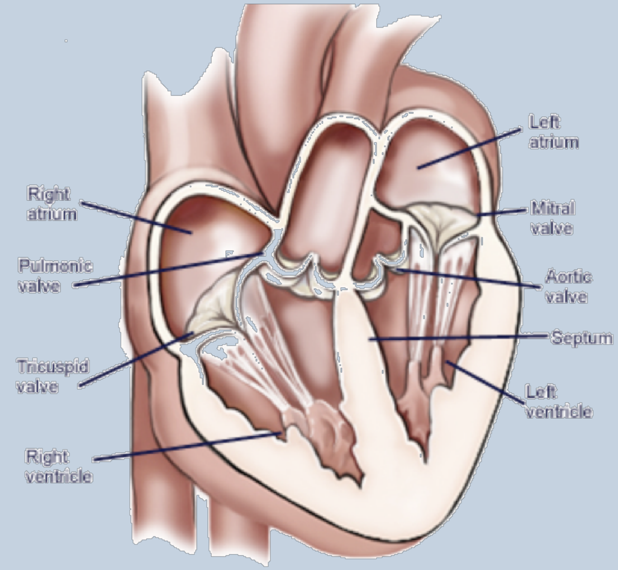
A **model** of (some aspect of) the world

- Introduces **vocabulary** relevant to domain
- Specifies (relative) **semantics** of terms

  Heart is a muscular organ that
  is part of the circulatory system

- **Formalised** using suitable logic

$$\forall x.[\mathsf{Heart}(x) \rightarrow \mathsf{MuscularOrgan}(x) \wedge \exists y.[\mathsf{isPartOf}(x,y) \wedge \mathsf{CirculatorySystem}(y)]]$$



Semantic Systems

- Semantic systems are not general purpose problem solvers
- Simple ‘litmus test’:
  - Can the domain be described as a collection of ‘truth statements’?
  - Can domain dependencies be described declaratively (e.g., using rules)?
  - Can this description be provided explicitly ‘in advance’?
  - Is structured ‘query answering’ the main use case?
  - If all ‘yes’ → use a semantic system (i.e., a ‘better database’)
- Examples of ‘non-reasoning’:
  - Information retrieval → very simple domain representation, queries are unstructured, algorithms are not declarative
  - Recommender systems → rules are implicit in users’ behaviour
  - Travel package planning → main task is not pattern matching, more related to constraint satisfaction
  - Route planning → specialised algorithms

Semantic Systems

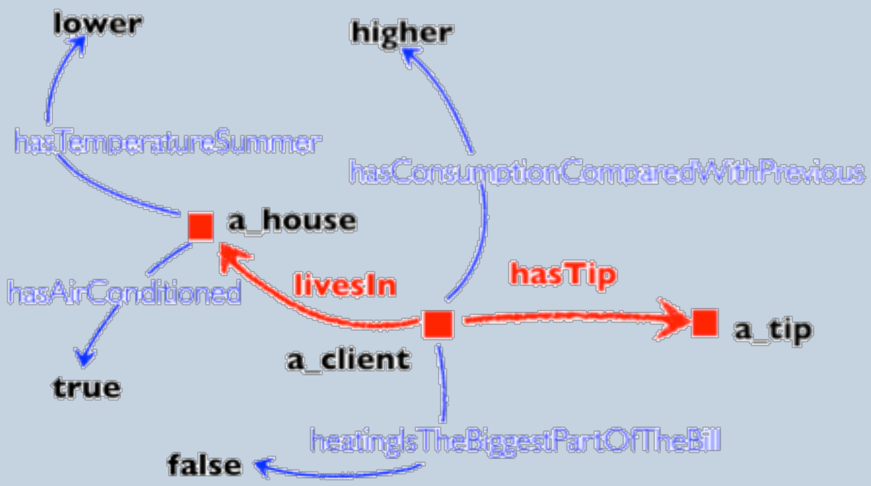
Key components:

- (Standardised) languages:
  - RDF for data
  - OWL for ontologies
  - SPARQL for queries
- Storage and reasoning systems:
  - RDF triple stores
  - Reasoning/query-answering systems (RDFox, HermiT, ELK, …)
- Other tools and infrastructure:
  - Ontology development environments (Protégé, Topbraid, …)
  - Other ontology services (bootstrapping, integrating, modularising, …)
  - APIs (RDF API, OWL API)

Motivating Applications

EDF Energy Adviser

- Produce personalised energy saving advice for EDF customers
- Describe customers' situations in RDF



- Encode advice rules using an ontology
- Interpret a situation against the rules using a reasoner
- A reasoning-intensive application!

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD
DBOnto
SIRIUS

Samsung Context-Aware Mobile Services

- Use sensors (WiFi, GPS, . . .) to identify the context
  - E.g., 'at home', 'in a shop', 'with a friend' . . .
- Adapt behaviour depending on the context
  - 'If with a friend who has birthday, remind to congratulate'
- Declaratively describe contexts and adaptations
  - Use a bunch of rules
  - E.g., 'If can see home Wifi, then context is "at home"'
- Interpret all rules in real-time using reasoning
- Main benefit: declarative, rather than procedural
- Reasoning plays a central role
- Challenges:
  - Handle rapid changes in sensor readings
  - Limited computational resources

HCLS Applications

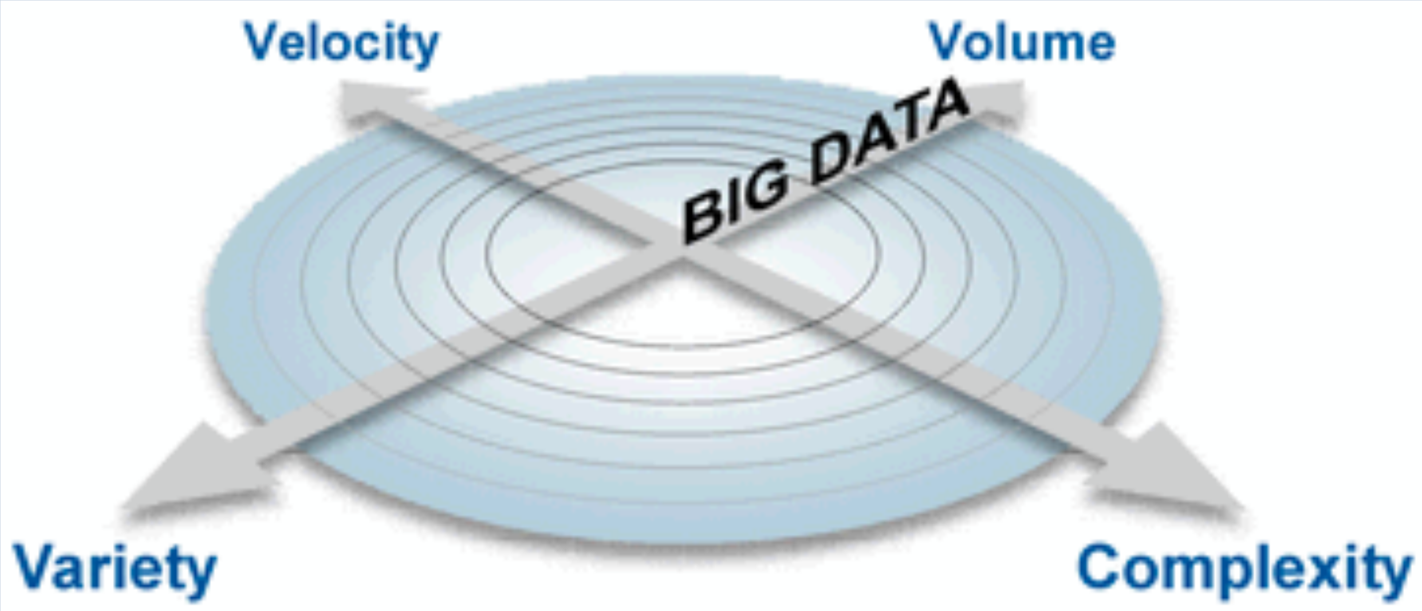
- **OBO foundry** includes more than 100 biological and biomedical ontologies
- **BioPax** “actively building OWL based clinical solutions”
  - Represents biological pathways data
  - Used in numerous databases, for visualisation, for data analysis
- **SNOMED-CT** (Clinical Terms) ontology
  - used in healthcare systems of more than 25 countries, including Australia, Canada, Denmark, Spain, Sweden and the UK
  - also used by major US providers, e.g., Kaiser Permanente
  - ontology provides common vocabulary for recording clinical data

Accessing (Big) Data

"a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications" (wikipedia)

Accessing (Big) Data

“a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications” (wikipedia)



Case Study: SIEMENS Energy Services

- Service centres responsible for remote monitoring and diagnostics of 1,000s of gas/steam turbines
- Engineers use a variety of data for visualization, diagnostics and trend detection:
  - several TB of time-stamped sensor data
  - several GB of event data
  - data grows at 30GB per day

Case Study: SIEMENS Energy Services

- Service centres responsible for remote monitoring and diagnostics of 1,000s of gas/steam turbines
- Engineers use a variety of data for visualization, diagnostics and trend detection:
  - several TB of time-stamped sensor data
  - several GB of event data
  - data grows at 30GB per day

Service Requests

- 1,000 requests per center per year
- 80% of time used on data gathering
- Potential saving: €50,000,000/year

Case Study: SIEMENS Energy Services

- Service centres responsible for remote monitoring and diagnostics of 1,000s of gas/steam turbines
- Engineers use a variety of data for visualization, diagnostics and trend detection:
  - several TB of time-stamped sensor data
  - several GB of event data
  - data grows at 30GB per day

Service Requests

- 1,000 requests per center per year
- 80% of time used on data gathering
- Potential saving: €50,000,000/year

Diagnostic Functionality

- 2–6 p/m to add new function
- New diagnostics → better exploitation of data
- Potential saving: incalculable

Case Study: Statoil Exploration

- Develop stratigraphic models of unexplored areas
- Geologists & geophysicists use data from previous operations in nearby locations
    - 1,000 TB of relational data
    - using diverse schemata
    - spread over 1,000s of tables
    - and multiple data bases

Case Study: Statoil Exploration

- Develop stratigraphic models of unexplored areas
- Geologists & geophysicists use data from previous operations in nearby locations
    - 1,000 TB of relational data
    - using diverse schemata
    - spread over 1,000s of tables
    - and multiple data bases

Data Access

- 900 geologists & geophysicists
- 30-70% of time on data gathering
- 4 day turnaround for new queries
- Potential saving: €70,000,000/year

UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

DBOnto

SIRIUS

Case Study: Statoil Exploration

- Develop stratigraphic models of unexplored areas
- Geologists & geophysicists use data from previous operations in nearby locations
  - 1,000 TB of relational data
  - using diverse schemata
  - spread over 1,000s of tables
  - and multiple data bases

Data Access

- 900 geologists & geophysicists
- 30-70% of time on data gathering
- 4 day turnaround for new queries
- Potential saving: €70,000,000/year

Data Exploitation

- Better use of experts time
- Data analysis “most important factor” for drilling success
- Potential value: > €10bn/project

UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

DBOnto

SIRIUS

Semantic Web

What is the Semantic Web?

- According to **TBL** circa 1998:

  "... a **consistent logical web of data** ..." in which
  "... information is given **well-defined meaning** …"

What is the Semantic Web?

- According to **TBL** circa 1998:

  “... a **consistent logical web of data** ...” in which
  “... information is given **well-defined meaning** …”

- By now has evolved into:

  “a platform for distributed applications and sharing (linking) data”

What is the Semantic Web?

- According to **TBL** circa 1998:

  "... a **consistent logical web of data** ..." in which
  "... information is given **well-defined meaning** …"

- By now has evolved into:

  "a platform for distributed applications and sharing (linking) data"

  - **RDF** provides uniform syntactic structure for data
  - **OWL** provides machine readable schemas (**ontologies**)
  - **SPARQL** provides standard query language

What is the Semantic Web?

- According to **TBL** circa 1998:

  "... a **consistent logical web of data** ..." in which
  "... information is given **well-defined meaning** …"

- By now has evolved into:

  "a platform for distributed applications and sharing (linking) data"

  - **RDF** provides uniform syntactic structure for data
  - **OWL** provides machine readable schemas (**ontologies**)
  - **SPARQL** provides standard query language

i.e., a large distributed ontology based information system

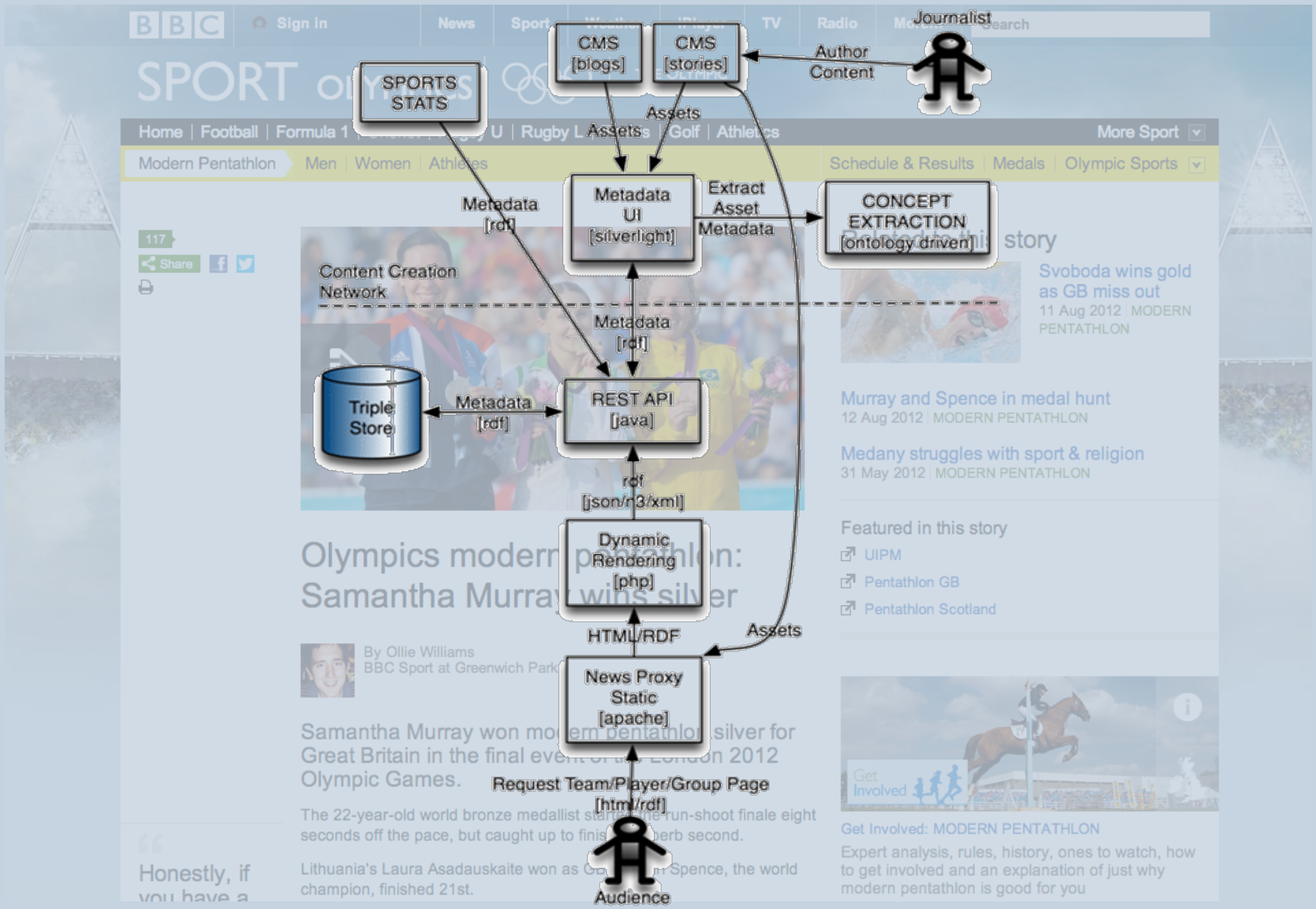
Semantic Web Applications

UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

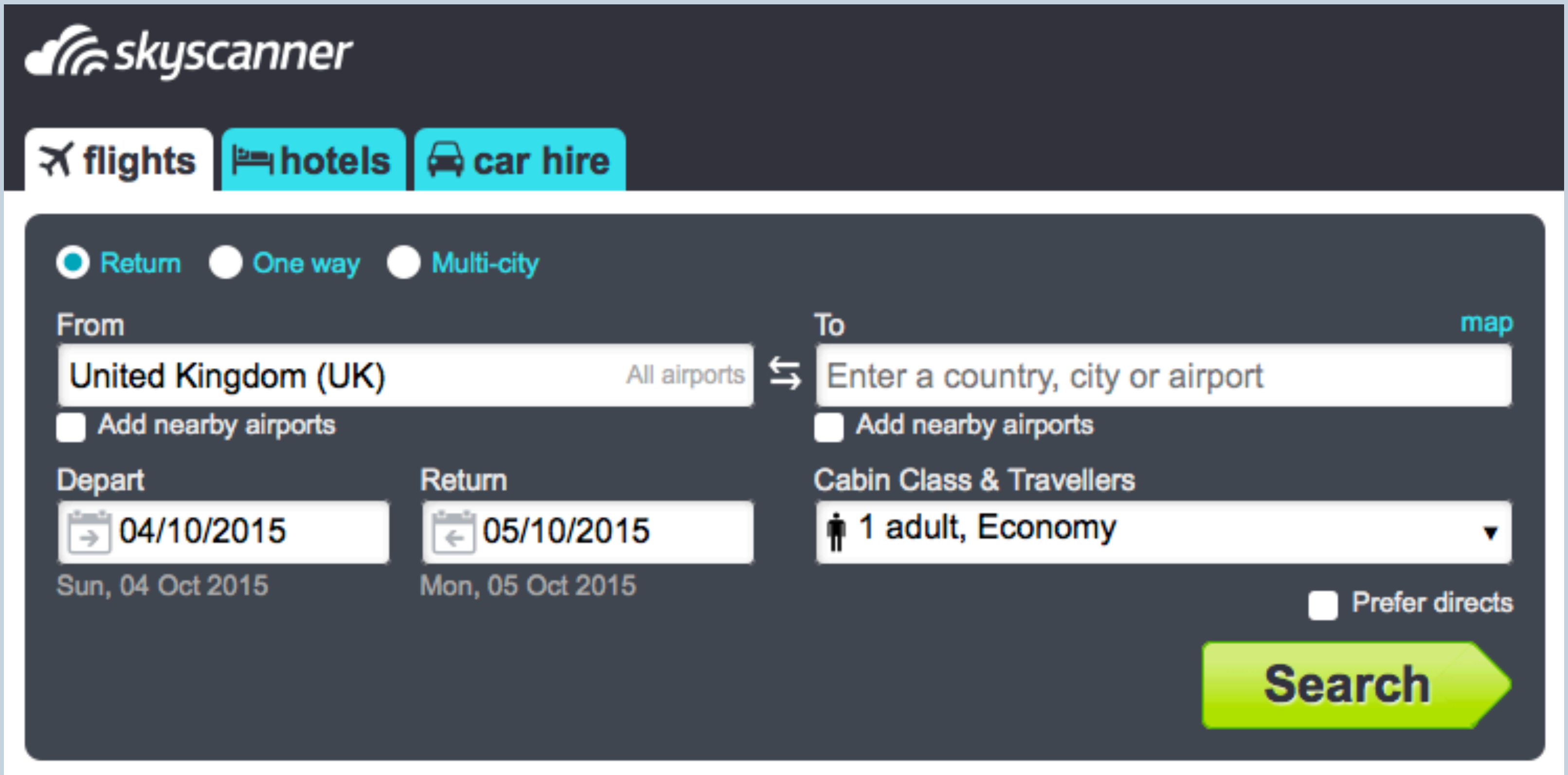
DBOnto

SIRIUS

Semantic Web Applications



Semantic Web Applications



UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

DBOnto

SIRIUS

Semantic Web Applications

Google

Semantic Web Applications

- Explicit KR sometimes needed, e.g., **Knowledge Graph**
  - Less rigorous treatment of semantics
  - Not using Semantic Web standards

Semantic Web Applications

- Explicit KR sometimes needed, e.g., **Knowledge Graph**
  - Less rigorous treatment of semantics
  - Not using Semantic Web standards
- **Hiring** Semantic Web people