

F2xBD: Big Data Management

Exam Revision Notes

Lectures 1-3

Introduction to Big Data
Semantic Web & Knowledge Graphs
Semantic Web Technologies (RDF)

Assessment Breakdown:

Class Test (Week 7): 20%
Final Exam Part 1: 40% (Ontologies, OWL, SPARQL)
Final Exam Part 2: 40% (NoSQL)

What This Document Covers:

- Data → Information → Knowledge
- Big Data Characteristics (5 V's)
- Semantic Web & Knowledge Graphs
- RDF Triples and Serialization Formats
- Linked Data Principles

Contents

1 LECTURE 1: Introduction to Big Data Management	2
1.1 The Big Picture: Why This Course Matters	2
1.2 Data vs Information vs Knowledge	2
1.2.1 Step-by-Step Explanation	2
1.2.2 The DIKW Pyramid (Exam Favorite!)	3
1.3 Big Data: The 5 V's	4
1.4 Types of Data in Big Data Context	5
1.5 What is Semantics?	6
2 LECTURE 2: Semantic Web & Knowledge Graphs	7
2.1 The Problem: Web of Documents	7
2.2 What is the Semantic Web?	8
2.3 What is a Knowledge Graph?	8
2.4 Who Uses Knowledge Graphs?	9
2.5 Linked Data Principles	10
2.6 5-Star Linked Open Data	10
2.7 Knowledge Graph Construction (Exam Topic)	11
3 LECTURE 3: Semantic Web Technologies (RDF)	12
3.1 The Semantic Web Technology Stack	12
3.2 IRIs and Namespaces	12
3.3 RDF: Resource Description Framework	13
3.3.1 Understanding RDF Triples	13
3.3.2 Types of Objects in Triples	14
3.4 RDF Serialization Formats (EXAM IMPORTANT!)	15
3.4.1 Turtle Format (MOST IMPORTANT!)	15
3.4.2 N-Triples Format	16
3.4.3 Other Formats (Brief Overview)	17
3.5 Open World vs Closed World Assumption	17
3.6 Inference (Basic Introduction)	18
4 EXAM PREPARATION	19
4.1 Key Terms Glossary	19
4.2 Common Exam Question Types	19
4.3 Practice Questions	20
4.4 Quick Revision Checklist	21

1 LECTURE 1: Introduction to Big Data Management

1.1 The Big Picture: Why This Course Matters

► Key Concept

What is Big Data Management?

Big Data Management is about **storing, organizing, and preparing** massive amounts of data so we can analyze it and make decisions.

Think of it like this:

- You have millions of puzzle pieces (data)
- You need to sort them (management)
- So you can see the picture (analytics/insights)

1.2 Data vs Information vs Knowledge

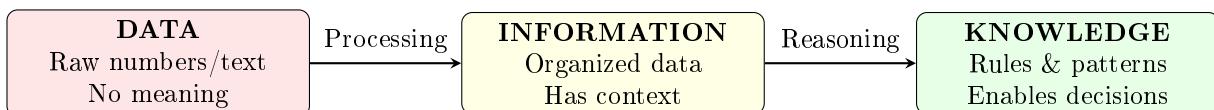
This is one of the **most frequently tested concepts**. You MUST understand the difference.

► Definition

The Three Levels:

1. **DATA** = Raw facts with no meaning
2. **INFORMATION** = Data + Context (organized data)
3. **KNOWLEDGE** = Information + Understanding (rules for decisions)

1.2.1 Step-by-Step Explanation



- Real-World Example: Coffee Shop

Scenario: You own a coffee shop.

DATA (raw facts):

- 42
- Monday
- Latte
- £3.50

These numbers mean nothing on their own!

INFORMATION (data + context):

- “On Monday, we sold 42 lattes at £3.50 each”
- Total = $42 \times £3.50 = £147$

Now we understand what the data means!

KNOWLEDGE (patterns + decisions):

- “Monday mornings are our busiest time for lattes”
- “We should have extra staff on Monday mornings”
- “We should order more milk before weekends”

Now we can make smart business decisions!

1.2.2 The DIKW Pyramid (Exam Favorite!)

★ Exam Tip

The DIKW Pyramid (Ackoff, 1989) appears frequently in exams. Remember the order from bottom to top: **D**ata → **I**nformation → **K**nowledge → **W**isdom

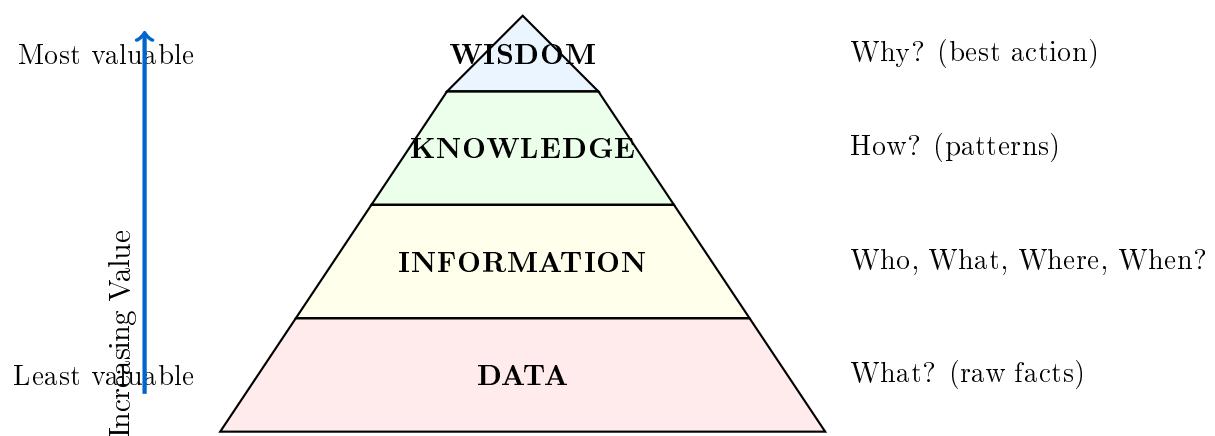


Table 1: DIKW Pyramid - Complete Reference Table

examblue!20 Level	Description	Example	Question Answered
Data	Raw, unprocessed facts. No context or meaning.	25, “Dubai”, 15:30	What are the facts?
Information	Data organized with context. Describes a situation.	“Temperature in Dubai at 15:30 was 25°C”	What is happening?
Knowledge	Patterns and rules derived from information.	“Dubai is hot in summer, mild in winter”	How does it work?
Wisdom	Applying knowledge to make the best decision.	“Visit Dubai in winter for comfortable weather”	What should we do?

1.3 Big Data: The 5 V's

► Key Concept

Big Data refers to data that is so large, fast, or complex that traditional methods cannot handle it.

Remember: 5 V's (likely exam question!)

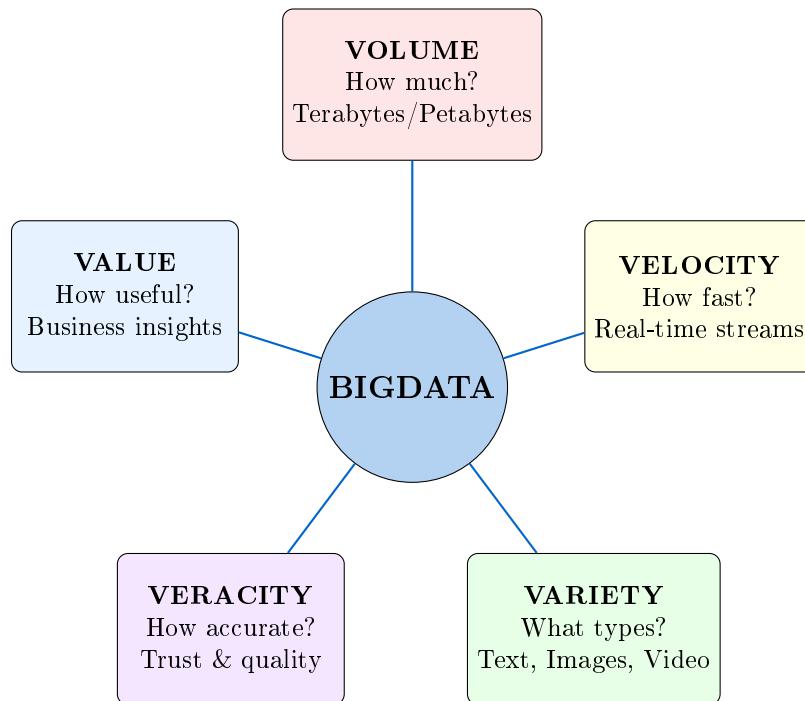


Table 2: The 5 V's of Big Data - Detailed Breakdown

examblue!20 V	Meaning	Real Example	Challenge
Volume	Amount of data	Facebook: 500+ TB new data daily	Need special storage systems
Velocity	Speed of data creation	Stock market: millions of trades/second	Need real-time processing
Variety	Different data types	Hospital: X-rays + notes + sensors	Need flexible formats
Veracity	Accuracy & trust	Twitter: real news vs fake news?	Need data validation
Value	Usefulness	Can we make money from this data?	Need good analytics

- **Exam Practice: Identify the V's**

Question: A hospital collects patient data. Match each scenario to its V:

1. The hospital has 10 years of patient records (5 million records)
2. Data includes text notes, X-ray images, and heart monitor readings
3. Heart monitors send 1000 readings per second
4. Some handwritten notes are hard to read accurately
5. The data helps predict which patients might have complications

Answers:

1. **Volume** - Large amount of data
2. **Variety** - Different types (text, images, sensor data)
3. **Velocity** - Fast, real-time data
4. **Veracity** - Accuracy/quality concern
5. **Value** - Useful insights for decisions

1.4 Types of Data in Big Data Context

Table 3: Types of Data - Exam Reference

examblue!20 Type	Definition	Example
Big Data	High volume, velocity, variety data requiring special processing	All tweets posted globally
Open Data	Data available free to use and share	UK government statistics
Linked Data	Data connected to other data using web standards	Wikipedia linking to maps
Linked Open Data	Open data that is also linked	DBpedia, Wikidata
Smart Data	Big data processed into actionable insights	“Customer likely to buy X”

1.5 What is Semantics?

▷ Definition

Semantics = The **meaning** of data.

Why it matters: A computer sees “apple” as just 5 letters. It doesn’t know if you mean:

- The fruit
- The company
- A person’s name

Semantic technologies add **meaning** so computers can understand context.

2 LECTURE 2: Semantic Web & Knowledge Graphs

2.1 The Problem: Web of Documents

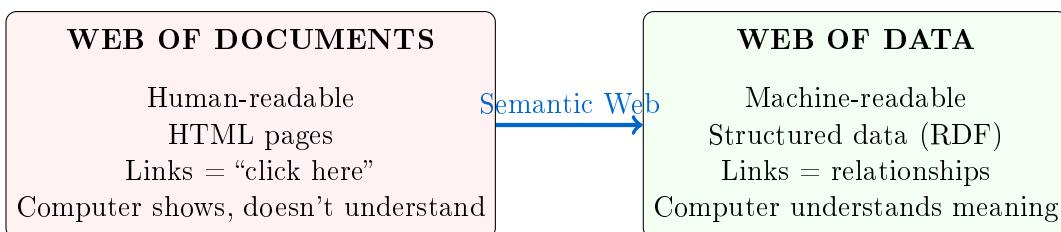
► Key Concept

The Current Web (Web of Documents):

- Made for **humans** to read
- Pages contain text, images, links
- Computers can **display** pages but can't **understand** them

The Goal (Semantic Web / Web of Data):

- Made for **machines** to understand
- Data has meaning attached
- Computers can **reason** and **connect** information



Invented: 1989 by Tim Berners-Lee

Goal of Semantic Web

• Why Machines Can't Understand Web Pages

You search Google for “Radu Mihailescu”:

What a human sees and understands:

- This is a person
- He’s a male professor
- He works at Heriot-Watt University
- His research areas include Machine Learning

What a computer sees:

- Text: “Radu” “Mihailescu” “Professor” “Heriot” “Watt”...
- Just strings of characters - no understanding!

Solution: Add metadata (data about data):

```

1 <Person>
2   <name>Radu Mihailescu</name>
3   <jobTitle>Associate Professor</jobTitle>
4   <worksAt>Heriot-Watt University</worksAt>
5 </Person>
```

Now the computer knows the structure and meaning!

2.2 What is the Semantic Web?

► Definition

Semantic Web = An extension of the World Wide Web where:

- Data has **meaning** (semantics) attached
- Machines can **understand** and **process** data
- Data from different sources can be **connected**

Created by: Tim Berners-Lee (inventor of the Web)

Managed by: W3C (World Wide Web Consortium)

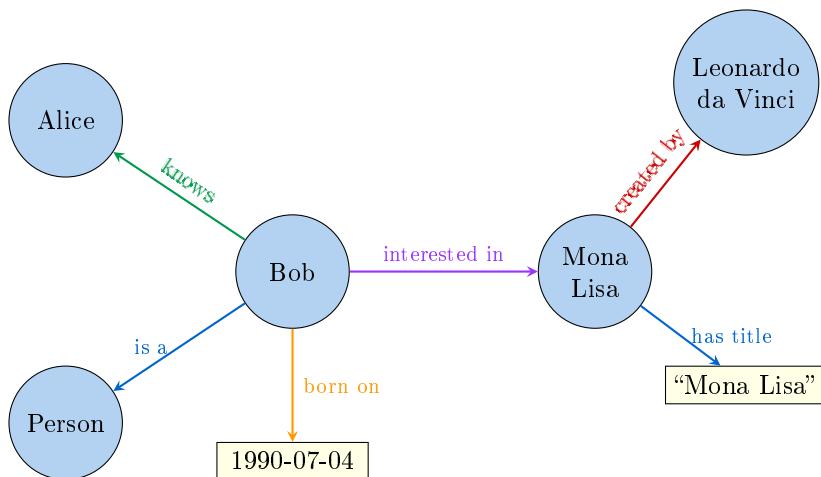
2.3 What is a Knowledge Graph?

► Key Concept

Knowledge Graph = A way to store knowledge using:

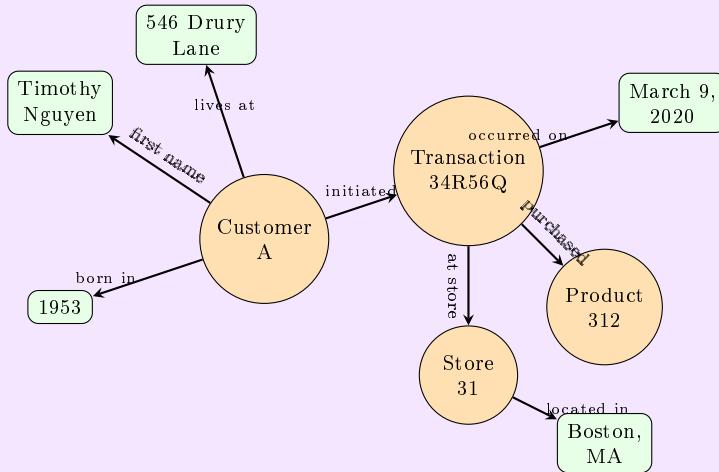
- **Nodes** = Things (people, places, concepts)
- **Edges** = Relationships between things

It's like a mind map, but for computers!



- Knowledge Graph: Customer Example

A retail company might store customer data like this:



What questions can we answer?

- Where does Customer A live? → 546 Drury Lane
- What did they buy? → Product 312
- Where did they buy it? → Store 31 in Boston

2.4 Who Uses Knowledge Graphs?

Table 4: Major Knowledge Graph Applications

examblue!20 Company	Knowledge Graph	Use Case
Google	Google Knowledge Graph	Search results (info boxes)
Facebook	Social Graph	Friend recommendations
Amazon	Product Graph	“Customers also bought...”
Microsoft	Microsoft Graph	Office 365 integration

★ Exam Tip

When you search for a famous person on Google and see a box with their photo, birth date, and facts - that comes from Google's Knowledge Graph! It has over **500 billion facts about 5 billion entities**.

2.5 Linked Data Principles

► Key Concept

Linked Data = Best practices for publishing and connecting data on the Web.

Tim Berners-Lee's 4 Rules:

1. Use **IRIs** (web addresses) to name things
2. Use **HTTP IRIs** so people can look up those names
3. Provide **useful information** when someone looks up an IRI
4. Include **links to other IRIs** so people can discover more

► Definition

IRI = International Resource Identifier

Think of it as a **unique web address for any thing** - not just web pages, but also people, places, concepts.

Examples:

- <http://dbpedia.org/resource/Edinburgh> - The city Edinburgh
- <http://example.org/person/Bob> - A person called Bob

2.6 5-Star Linked Open Data

★ Exam Tip

The 5-star scheme is a common exam topic! More stars = better data quality for the Semantic Web.

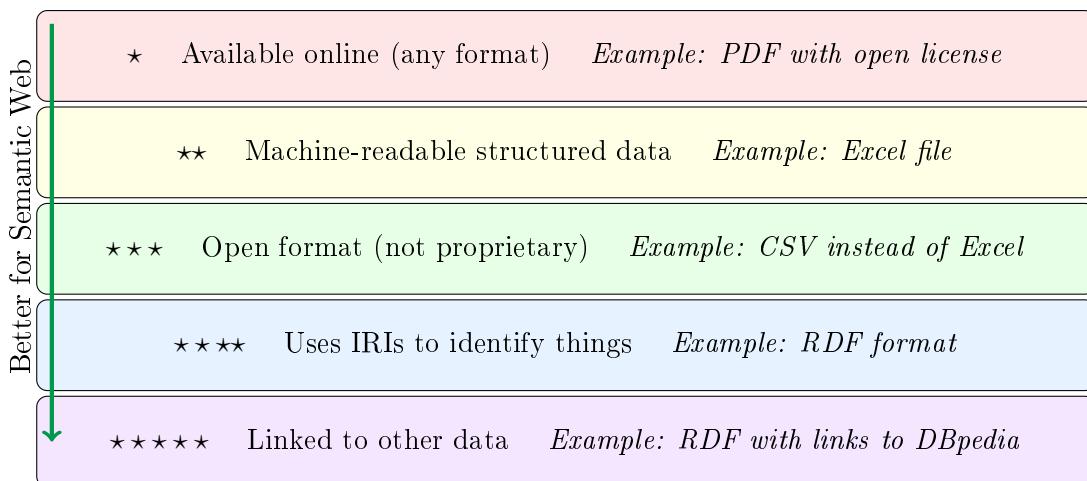


Table 5: 5-Star Linked Open Data - Quick Reference

examblue!20 Stars	Requirement	Format Example	Can machines use it?
★	Online + Open license	PDF, JPG	No - just view
★★	+ Structured data	Excel (XLS)	Partially
★★★	+ Non-proprietary format	CSV, XML	Yes - open tools
★★★★	+ Uses IRIs (RDF)	RDF/Turtle	Yes - identified entities
★★★★★	+ Links to other data	Linked RDF	Yes - connected web of data

2.7 Knowledge Graph Construction (Exam Topic)

- Exercise: Build a Knowledge Graph

Given these statements:

1. John is a lecturer
2. John teaches Big Data Management
3. Big Data Management is a course at Heriot-Watt University
4. Big Data Management is taught in Edinburgh and Dubai

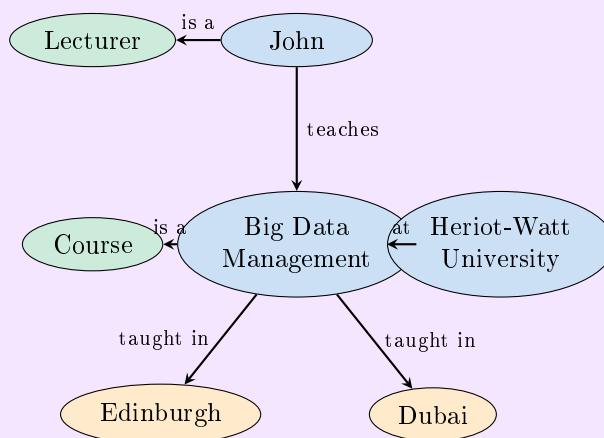
Step 1: Identify Entities (Nodes)

- John (Person)
- Lecturer (Class/Type)
- Big Data Management (Course)
- Course (Class/Type)
- Heriot-Watt University (Organization)
- Edinburgh (Location)
- Dubai (Location)

Step 2: Identify Relationships (Edges)

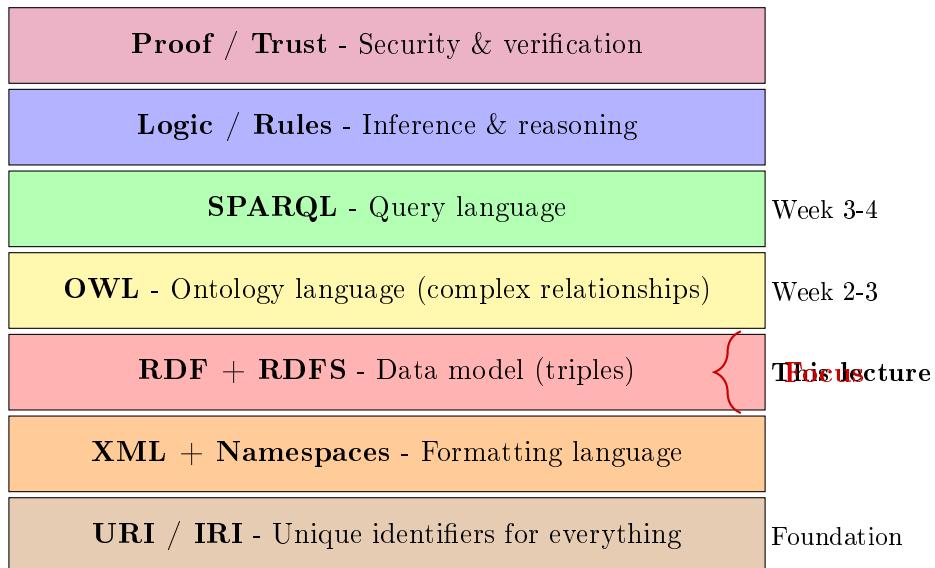
- is a (type relationship)
- teaches
- at (location)
- taught in

Step 3: Draw the Graph



3 LECTURE 3: Semantic Web Technologies (RDF)

3.1 The Semantic Web Technology Stack



3.2 IRIs and Namespaces

► Definition

IRI (International Resource Identifier) = A unique web address that identifies ANY thing.
Structure:

`http://example.org/resource/Thing`

Why IRIs matter:

- Two different databases might both have a “Person” - IRIs make them unique
- `http://companyA.com/Person` ≠ `http://companyB.com/Person`

► Key Concept

Namespace = A shortcut for long IRIs.

Instead of writing:

```

1 <http://xmlns.com/foaf/0.1/Person>
2 <http://xmlns.com/foaf/0.1/knows>
3 <http://xmlns.com/foaf/0.1/name>
```

We declare a prefix:

```
1 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

Then write:

```

1 foaf:Person
2 foaf:knows
3 foaf:name
```

Much shorter and cleaner!

Table 6: Common Namespaces (Memorize for Exam!)

examblue!20 Prefix	Full IRI	Used For
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF basics
rdfs:	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
owl:	http://www.w3.org/2002/07/owl#	Ontologies
xsd:	http://www.w3.org/2001/XMLSchema#	Data types
foaf:	http://xmlns.com/foaf/0.1/	People/social
dc / dcterms:	http://purl.org/dc/terms/	Document metadata

3.3 RDF: Resource Description Framework

► Key Concept

RDF is the foundation of the Semantic Web. It describes data using **TRIPLES**.

A Triple = Subject + Predicate + Object

Think of it as: “Something has a relationship with something else”

3.3.1 Understanding RDF Triples

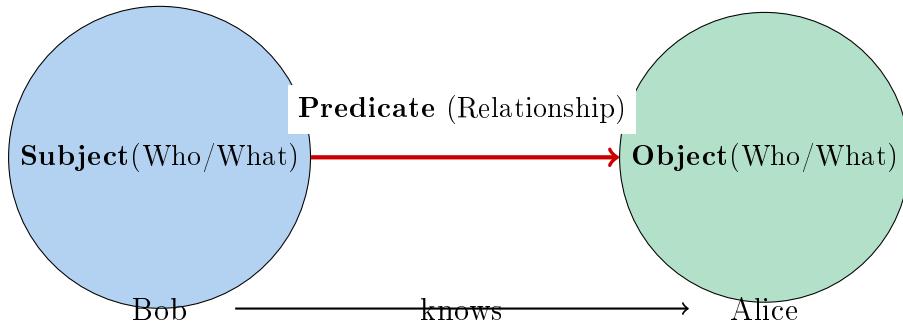


Table 7: RDF Triple Components

examblue!20 Component	What it is	Must be IRI?	Example
Subject	The thing being described	Yes (always IRI)	<code>ex:Bob</code>
Predicate	The relationship/property	Yes (always IRI)	<code>foaf:knows</code>
Object	The value or related thing	IRI or Literal	<code>ex:Alice</code> or “Bob”

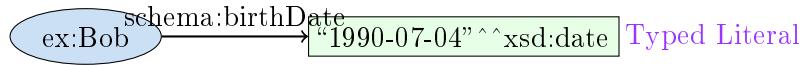
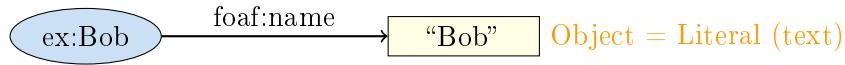
△ Warning - Common Mistake

Subject and Predicate must **ALWAYS** be IRIs!

Object can be:

- An **IRI** (another resource): `ex:Alice`
- A **Literal** (plain value): “Bob” or “1990-07-04”

3.3.2 Types of Objects in Triples

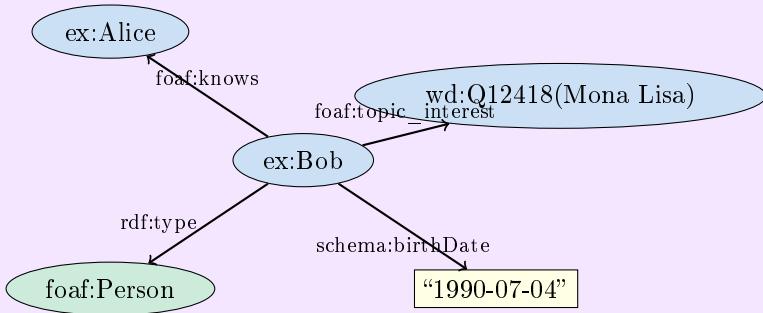


- Complete RDF Example

Describe Bob with these facts:

- Bob is a Person
- Bob knows Alice
- Bob was born on July 4, 1990
- Bob is interested in the Mona Lisa

As a graph:



As triples (table format):

Subject	Predicate	Object
ex:Bob	rdf:type	foaf:Person
ex:Bob	foaf:knows	ex:Alice
ex:Bob	schema:birthDate	“1990-07-04” ^{^^} xsd:date
ex:Bob	foaf:topic_interest	wd:Q12418(Mona Lisa)

3.4 RDF Serialization Formats (EXAM IMPORTANT!)

► Key Concept

RDF is an **abstract model** (the concept of triples). To store/share RDF, we need a **concrete format** (serialization).

Main formats you need to know:

1. **Turtle** - Human-readable, most common
2. **N-Triples** - Simple, one triple per line
3. **RDF/XML** - XML-based (original format)
4. **JSON-LD** - JSON-based (for web)
5. **RDFa** - Embedded in HTML

3.4.1 Turtle Format (MOST IMPORTANT!)

★ Exam Tip

Turtle is the format most likely to appear in exams. Learn to read AND write it!

```

1 # Step 1: Declare prefixes (namespaces)
2 @prefix ex: <http://example.org/> .
3 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
4 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
5
6 # Step 2: Write triples
7 # Full form: subject predicate object .
8 ex:Bob foaf:knows ex:Alice .
9
10 # Same subject, different predicates: use ;
11 ex:Bob foaf:name "Bob" ;
12     foaf:age "34"^^xsd:integer ;
13     foaf:knows ex:Alice .
14
15 # Same subject AND predicate, different objects: use ,
16 ex:Bob foaf:knows ex:Alice , ex:Carol , ex:Dave .
17
18 # rdf:type can be shortened to 'a'
19 ex:Bob a foaf:Person .
20 # This means: ex:Bob rdf:type foaf:Person .

```

Listing 1: Turtle Syntax Basics

Table 8: Turtle Syntax Quick Reference

examblue!20 Symbol	Meaning	Example
.	End of statement	ex:Bob foaf:knows ex:Alice .
;	Same subject, new predicate	ex:Bob foaf:name "Bob" ; foaf:age 25 .
,	Same subject+predicate, new object	ex:Bob foaf:knows ex:Alice , ex:Carol .
a	Shortcut for rdf:type	ex:Bob a foaf:Person .
@prefix	Declare namespace	@prefix ex: <http://example.org/> .
“text”	String literal	foaf:name "Bob"
“value”^^type	Typed literal	"25"^^xsd:integer

- **Turtle Practice:** Write RDF for a CD

Facts to encode:

- “Empire Burlesque” is a CD
- Artist: Bob Dylan
- Company: Columbia
- Year: 1985

Solution:

```

1 @prefix cd: <http://example.org/cd#> .
2 @prefix dct: <http://purl.org/dc/terms/> .
3 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
4
5 cd:Empire_Burlesque a cd:CD ;
6   dct:title "Empire Burlesque" ;
7   cd:artist "Bob Dylan" ;
8   cd:company "Columbia" ;
9   cd:year "1985"^^xsd:gYear .

```

This creates 5 triples:

1. cd:Empire_Burlesque rdf:type cd:CD
2. cd:Empire_Burlesque dct:title “Empire Burlesque”
3. cd:Empire_Burlesque cd:artist “Bob Dylan”
4. cd:Empire_Burlesque cd:company “Columbia”
5. cd:Empire_Burlesque cd:year “1985”

3.4.2 N-Triples Format

```

1 # Every triple is complete - no shortcuts!
2 # Full IRIs in angle brackets < >
3 <http://example.org/Bob> <http://xmlns.com/foaf/0.1/knows> <http://example.org/Alice> .
4 <http://example.org/Bob> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
5 <http://example.org/Bob> <http://xmlns.com/foaf/0.1/name> "Bob" .

```

Listing 2: N-Triples Example

Table 9: Comparison: Turtle vs N-Triples

examblue!20 Feature	Turtle	N-Triples
Prefixes	Yes (makes code shorter)	No (full IRIs always)
Shortcuts	; , a	None
Readability	High (human-friendly)	Low (verbose)
Best for	Writing by hand, teaching	Machine processing, bulk loading

3.4.3 Other Formats (Brief Overview)

Table 10: RDF Serialization Formats Summary

examblue!20 Format	Description	Best For	Extension
Turtle	Human-readable, compact	Writing, learning	.ttl
N-Triples	One triple per line	Bulk data loading	.nt
N-Quads	N-Triples + graph name	Multiple graphs	.nq
RDF/XML	XML syntax for RDF	Legacy systems	.rdf
JSON-LD	JSON with linked data	Web APIs	.jsonld
RDFa	RDF embedded in HTML	Web pages	.html
TriG	Turtle + named graphs	Multiple graphs	.trig

3.5 Open World vs Closed World Assumption

► Key Concept

This affects how we answer questions when data is missing!

Closed World Assumption (CWA) - Used by databases

- If something is not stated, it is **FALSE**
- “What I don’t know isn’t true”

Open World Assumption (OWA) - Used by Semantic Web

- If something is not stated, it is **UNKNOWN**
- “What I don’t know might still be true”

• Open World vs Closed World

Knowledge Base:

```
1 ex:Socrates a ex:Human .
2 ex:Human rdfs:subClassOf ex:Mortal .
```

Question: Is Plato mortal?

Closed World Answer:

- Plato is not in the database
- Therefore: **NO**, Plato is not mortal (false)

Open World Answer:

- Plato is not in the database
- But he might exist and be mortal
- Therefore: **UNKNOWN**

★ Exam Tip

Remember: **RDF/Semantic Web uses Open World Assumption!**

This means:

- Missing data \neq False
- We can always add more facts later
- Useful for distributed data (web)

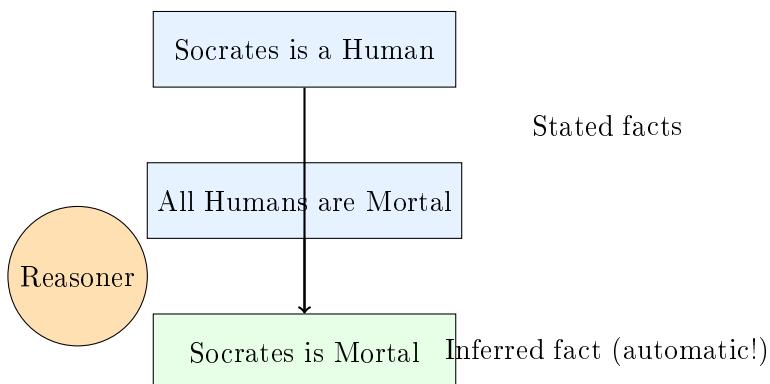
3.6 Inference (Basic Introduction)

▷ Definition

Inference = Deriving new facts from existing facts using rules.

Example:

- Fact 1: Socrates is a Human
- Fact 2: All Humans are Mortal
- **Inference:** Socrates is Mortal (derived automatically!)



4 EXAM PREPARATION

4.1 Key Terms Glossary

examblue!20 Term	Definition (Exam-Ready)
Big Data	Data characterized by high Volume, Velocity, Variety, Veracity, and Value
Semantic Web	Extension of the web where data has machine-readable meaning
Knowledge Graph	A graph-based knowledge representation with nodes (entities) and edges (relationships)
RDF	Resource Description Framework - W3C standard for describing data as triples
Triple	Subject-Predicate-Object statement in RDF
IRI	International Resource Identifier - unique web address for any resource
Namespace	Prefix shortcut for IRIs (e.g., foaf: for FOAF vocabulary)
Literal	A data value (string, number, date) rather than a resource
Turtle	Human-readable RDF serialization format
OWA	Open World Assumption - missing info means unknown, not false
CWA	Closed World Assumption - missing info means false
Linked Data	Best practices for publishing connected data on the web
LOD	Linked Open Data - linked data with open license
Inference	Deriving new facts from existing facts using rules
DIKW	Data-Information-Knowledge-Wisdom pyramid

4.2 Common Exam Question Types

Q1. Define and Compare: “What is the difference between Data and Information?”

Q2. 5 V’s: “Identify which V applies to this scenario...”

Q3. Draw Knowledge Graph: “Given these facts, draw the knowledge graph...”

Q4. Write Turtle: “Express the following in Turtle format...”

Q5. Read Turtle: “How many triples are in this Turtle code?”

Q6. 5-Star LOD: “What star rating does this data have?”

Q7. OWA vs CWA: “What would a reasoner conclude under OWA?”

4.3 Practice Questions

Practice Question 1: DIKW

Classify each as Data, Information, or Knowledge:

- a) 37.5°C
- b) "Patient has a fever of 37.5°C"
- c) "Fever above 38°C requires medication"

Answer:

- a) **Data** - raw number
- b) **Information** - data with context
- c) **Knowledge** - rule for decision

Practice Question 2: Write Turtle

Express in Turtle: "John is a Student who studies at Heriot-Watt University"

Answer:

```

1 @prefix ex: <http://example.org/> .
2 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3
4 ex:John a ex:Student ;
5     ex:studiesAt ex:HeriotWattUniversity .
```

Practice Question 3: Count Triples

How many triples in this Turtle code?

```

1 ex:Alice a foaf:Person ;
2     foaf:name "Alice" ;
3     foaf:knows ex:Bob , ex:Carol .
```

Answer: 4 triples

1. ex:Alice rdf:type foaf:Person
2. ex:Alice foaf:name "Alice"
3. ex:Alice foaf:knows ex:Bob
4. ex:Alice foaf:knows ex:Carol

4.4 Quick Revision Checklist

examblue!20 Topic	Check
Can explain DIKW pyramid with examples	<input type="checkbox"/>
Can name and explain all 5 V's of Big Data	<input type="checkbox"/>
Can define Semantic Web and its purpose	<input type="checkbox"/>
Can explain what a Knowledge Graph is	<input type="checkbox"/>
Know the 4 Linked Data principles	<input type="checkbox"/>
Can explain 5-star Linked Open Data	<input type="checkbox"/>
Understand RDF triple structure (S-P-O)	<input type="checkbox"/>
Can read and write Turtle syntax	<input type="checkbox"/>
Know difference between IRI and Literal	<input type="checkbox"/>
Can use Turtle shortcuts (; , a)	<input type="checkbox"/>
Understand OWA vs CWA	<input type="checkbox"/>
Know common namespace prefixes	<input type="checkbox"/>

Good luck with your exam!

Remember: Understanding > Memorization
Practice writing Turtle by hand!