CSCE60413 - Semantic Web

# Web versus Semantic Web

Adila Krisnadhi

UNIVERSITAS INDONESIA
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS INDONESIA
tokopedia AI Center of Excellence

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

*Hello!*

# I am Adila Krisnadhi

Faculty member at the Faculty of Computer Science, Universitas Indonesia.
Co-director of Tokopedia-UI AI Center of Excellence
Ontology engineer and Semantic Web enthusiast

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# Credits

- Presentation template by SlidesCarnival
- Photographs by Unsplash
- Backgrounds by SubtlePatterns
- Content by Eero Hyvonen from the CS-E410 lecture slides “WWW Today” and “Semantic Web” (with partial adaptation by me).

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# 1. The World Wide Web

# The Web in a nutshell

## Users

7.676 billion world population.

4.4 billion Internet users.

5.1 billion unique mobile users.

## Content

Approx. 55 billion pages indexed by Google (size of Surface Web).

Deep Web (not just Dark Web) is est. up to 400 times larger

## Effectiveness of publishing

All information readable by everyone.

New content easy to publish to billions of audiences.

Almost “free” usage.

# Basic ingredients of the Web

`<http://www.blablabla.com/myfolder/mydoc.html>`

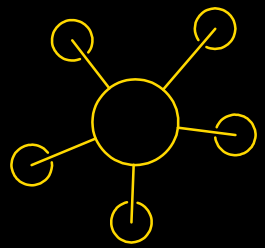
## URL addresses

Represent resources such as websites, documents, pictures

## HTML language

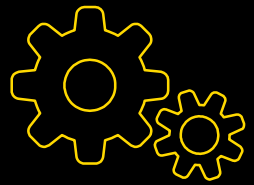
Expresses resource content (Web pages)

Enables hyperlinks



## HTTP protocol

Specifies mechanism for transferring resources between server and client.



UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# 2. Services on the web

# The Web provides the following services ...

**Functional**

Banking, shopping, government, entertainment, etc.

**Information retrieval**

Search engines and browsing.

Portals, directories.

Databases in different applications.

# Information retrieval challenges: end-user perspective

- **Search query formulation**
  - Creating queries that work as intended is not always easy.
- **Search result quality**
  - Recall: proportion of the relevant information is found.
  - Precision: proportion of the found information is relevant.
  - Relevance: how well the results correspond to the user needs
- **Search result presentation**
  - Ease of understanding
  - Ranking and structuring

# Limitation of basic text search: Examples (1)

## Search term may appear in an irrelevant document

- "Halaman ini *tidak berbicara tentang* **kucing.**"

## Identifying (some) synonyms

- "hewan" vs. "satwa", "pranala" vs. "tautan"
  - Bad recall (relevant pages not found), query formulation difficult

## Identifying homonyms

- "Palu": a city, a tool
- "Depok": a city south of Jakarta, a district in Yogyakarta
  - Bad precision, garbage results; results hard to understand; query formulation hard

# Limitation of basic text search: Examples (2)

**Computer does not understand relations between concepts**

- Narrower-broader concept, part-whole (Google is smarter now)
  - Search “surface depression in solar system” should return all craters on Earth.
- Missing background or common sense knowledge
  - Searching ‘smoke’ may not return pages about ‘fire’

**Information to be searched is fragmented, but results cannot be aggregated**

- “Search publications of the members of the research group X”

# Limitation of basic text search: Examples (3)

**Finding relations between information resources is hard**

- "How is Dieng related to Borobudur?"

**Search does not go beyond "memorizing"**

- How much does a kilogram of feathers weigh on the moon?
- With lots of memorizing, problem solving resembles remembering.

**Context/personalization is under-utilized**

- What could I do today in Depok?

**Computer unable to "understand" meaning/semantics**

- Which city has Makassar as one of its district?

# Browsing challenges in the Web: end-user perspective

- **Understanding the ”big picture” in a large fragmented information space**
  - ”Lost in the hyperspace”
- **Links get out of date and destroyed**
  - The linked target pages expire or are removed entirely
  - New pages do not get linked to old ones
  - Old pages do not get linked to new ones
- **Reliability of information and their providers**
  - ”Web of trust”
  - ”Flat Earth” organization’s page vs. our university’s scientific page
  - Wikipedia vs. Encyclopedia Britannica

# Knowledge management challenges: Information provider perspective

- **Structuring contents with links is manual work**
  - Information does not get linked at content level without human effort
- **Different organizations create overlapping information**
  - The same work is done multiple times
- **The contents and their structures are not interoperable**
  - Aggregation of collections of different memory organizations is difficult
  - Lack of interoperability prevents combining of contents
  - Lack of interoperability prevents the management of contents
- **Information about the contents and their changes is not communicated between organizations**

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# 3.
# Machine Processability (and Understandability?)

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# Web content is mainly created for ...

- Humans
- Machines

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# What can machines do w.r.t. information on the Web?

- Mediate
- Display
- Understand

# Machine processability problem

- Web content → for humans
- Machines only mediate and display content, and do not understand Web content.
- A web service → machine to help human
  - But requires machine to understand Web content.

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# How can we build a more intelligent Web?

- Make applications to be more intelligent?
- Represent content in a more intelligent way?

# Smart application approach

Machines need to be able to understand/interpret content, but:

- Automated interpretation of natural language is difficult
- Non-textual content is very hard to interpret
- Interpretation often needs context and common sense.

# Smart content approach

- Information is provided in a way that allows machines to understand it.
  - → **Key idea of Semantic Web**
- Who provides the information in that form?
  - → Human (and machines)
- History:
  - W3C Semantic Web activity in 2001
  - W3C Web Services activity in 2002

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# “Semantic” in Semantic Web means “understandable” to machines


# 4.
# Data (Knowledge?) Representation

# Markup Languages: Main Ideas

| | |
|---|---|
| Domain- and environment-independent standard for documents | Creation |
| | Management |
| | Transferring |
| Documents are text files | Open, simple format |
| | Usable on all platforms |
| | Easy modification, storing, reading, transfer |
| | Future-proof |
| Separate structure, content, and presentation | Describing document structure (for programmer) → HTML: <h1> Heading </h1> |
| | Describing information content (for programmer) → XML: <address>Jl. Margonda</address> |
| | Presentation is decided by the reader (browser) |

# The case for XML

- **Utilization of content structure**
  - Better recall/precision by search engines
- **Syntax validation**
- **Widely used on the Web**
  - Encoding of knowledge in open format
  - Open APIs for programming languages allow programmatic processing of the pages' content
- **Vendor-independent**
- **File format is unlikely to change**
  - Pages are text files
- **Enables domain-specific standard languages to be built on top**

# Syntactic interoperability through standardization

- World Wide Web Consortium (W3C)
  - Cooperation body of vendors, operators, universities, etc.
  - Creates and maintains Web standards/recommendations (HTML, XML, etc.)

- Domain-specific organizations
  - ISO: various domains, except electrical/electronical
  - IEC: electronical
  - CEN: european body, various domains
  - UN/CEFACT: electronic
  - OASIS: IT-related
  - ....

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# Case against XML

- Difficult to read and process by humans
  - Notation is not human-friendly
- Documents contain lots of repetition
  - Lots of redundancy, e.g., start and end tags, blowing up the size of the markup files.
  - Laborius to write.
  - Needs more bandwith for transfer

# Alternative (better) notations

- JSON (JavaScript Object Notation)
  - Representation as nested key-value pairs
  - Integrated into JavaScript: easy, efficient to use
  - Widely adopted
  - Standardized by IETF (RFC) and ISO/IEC
  - Adopted and extended into Semantic Web as JSON-LD

- Simple Semantic Web notations
  - Turtle and various OWL notations
  - Widely used
  - Standardized by W3C

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# 5. Semantic Web: Why Syntax only is not Enough

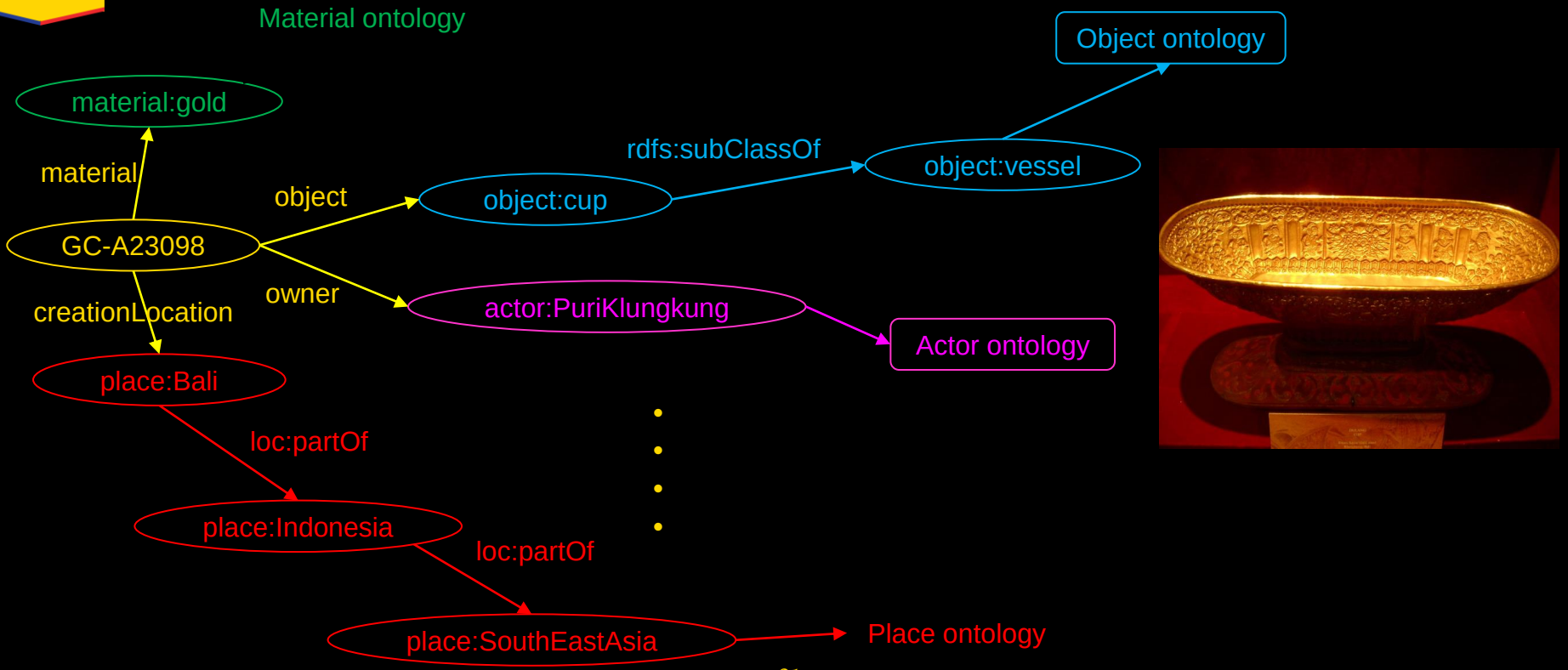
# Why syntax-only content is limited

```
<artifact>
    <id>GC:A23098</id>
    <target>cup</target>
    <material>gold</material>
    <creationLocation>Bali</creationLocation>
    <originalOwner>Puri Klungkung</originalOwner>
</artifact>
```

Metadata cannot answer the following:

- Find all vessels?
- Find all metal product?
- Find all artifacts manufactured in South East Asia?
- Does Puri Klungkung own metal artifacts?

# Key approach in Semantic Web: Use (Knowledge) **Graph**



# Key approach in Semantic Web: Use (Knowledge) **Graph**

```
GC-A23098
  rdfs:label "Cup" ;
  :object object:cup ;
  :material material:gold ;
  :creationLocation place:Bali ;
  :originalOwner actor:PuriKlungkung .
```

```
object:cup rdfs:subClassOf object:vessel .

place:Bali loc:partof place:Indonesia .
place:Indonesia loc:partOf place:SouthEastAsia .
```

- 
- 
- 
- 

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# Why XML (and JSON) only is not enough

- Interpretation of XML languages need to be defined in domain-specific way.
- Integration of XML documents is cumbersome.
- Semantic of XML tags is only in human brain
  - Machines only see arbitrary symbols.

# So, we need ...

- A markup language whose interpretation is:
  - commonly agreed
  - shared across different application domains
  - machine-understandable

Semantic Web solution → Resource Description Framework (RDF)

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# *Semantic Web original vision*

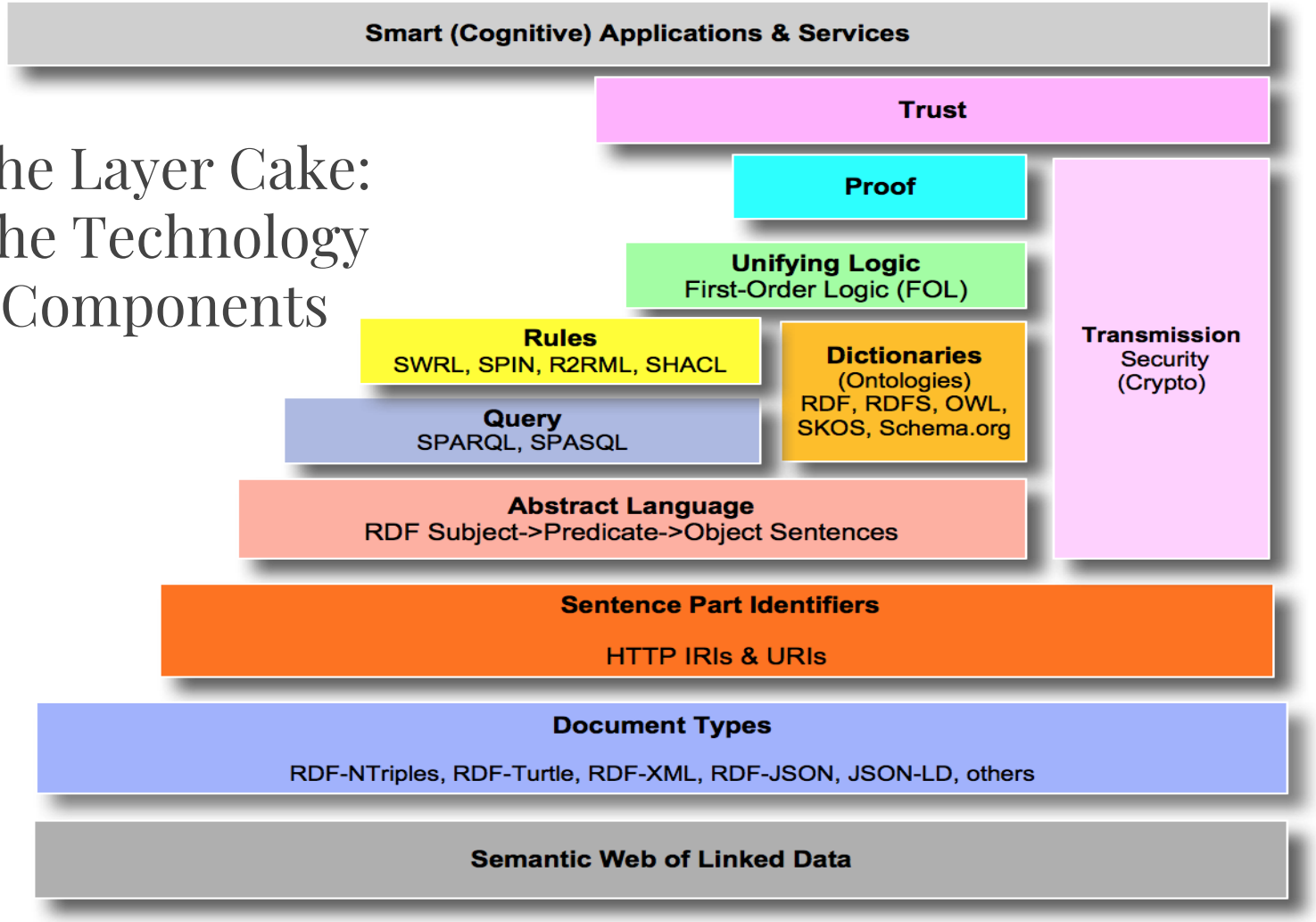
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities …

Tim Berners-Lee, James Hendler and Ora Lassila (Scientific American, May 2001)

https://www.scientificamerican.com/article/the-semantic-web/

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

# The Layer Cake: The Technology Components



# So, what is Semantic Web?

- **Content-perspective**: a new metadata layer on the web describing its contents in terms of shared vocabularies
  - Web as a global database system
  - Web of data (instead of Web of documents)
- **Application perspective**: machine-understandable web
  - Meaning/semantics of content accessible by machines.
  - Intelligent web services
  - Semantic interoperability
- **Technology perspective**: the layer(s) above XML.


UNIVERSITAS
INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF
COMPUTER
SCIENCE

# Fin.