# Semantic Web and Knowledge Graphs

`Integrating in one place heterogeneous information sources`.

## 1. Introduction

Tim Berners-Lee came out in 2007 saying: the whole internet structure is a **huge knowledge base**, we need to exploit this knowledge base in order to perform **reasoning**.

A web page is a set of informations that tell us how the data should be represented, according to the needs of humans being. Therefore, the content is published with the principal aim of being **human-readable**.

Usually, it's used the *HTML* standard to show informations. HTML is focused on **how** to represent content instead of **what** is represented. This means: looking to a news paper web page we recognize instantly where is located the title, even though there isn't any notion about what is a title.

In addition to the content, web pages contains also **links** to other pages. Links are problematic, they create connections between pages but do not tell us anything about these connections or any clue about the subsequent resource selected.

The problem is: it is not possible to automatically reason about the data. We need a different representation of this huge amount of informations.

## 2. Semantic Web

The **semantic web** is the integration and combination of data from diverse resources, where the original Web mainly concentrated on the interchange of documents.

It's based on the idea of extend the current web within the knowledge given by the content.

Semantic web should preserve:

- Globality.
- Information **distribution**. Since the web is a complex structure, it's great having different resources around the world.
- Information **inconsistency**. Different incoherent and inconsistent opinions about the same information source.
- Information **incompleteness**.

Adding information is not enough. First of all, we should describe a structure of the informations and, after that, we need an inference mechanism to be able to infer new knowledge.

For these main reasons, Tim Berners-Lee introduced some tools:

- **URI**, Uniform Resource Identifier.
  Every resource available is identified by an unique tag. Each URI corresponds to one and only one concept, but more URI can refer to the same concept.

- **XML**, eXtensible Markup Language.
  *HTML* is an extension of XML, created for supporting data exchange between heterogeneous systems.

- **RDF**, Resource Description Framework.
  Any concept can be described by the **triple**: $< \text{subject}, \text{predicate}, \text{object} >$. If we look deeper, this representation is the same seen so far in First Order Logic.

  ### Example

  Given the triple:

  $$< \text{John}, \text{has}, \text{" 30 "} >$$

  is equal to:

  ```
  has(John, "30").
  ```

---

## 3. Knowledge Graphs

The notion of **knowledge graph** is born above two main reasons, which are:

- Overcoming search approaches based on *stastical data retrieval* with something able to *represent and reason* upon the knowledge.
- *Semantic web* seems to be too complex; based on description logics, they returned out to be too expensive from the time complexity point of view.

*Google* answers the previous problems by:

- Creating a common and simple **vocabulary**.
- Creating a simple and **robust** corpus of types.
- Pushing the Web to **adopt** these standards.

## Example

From *description logics*, the **T-Box** approach is very expensive (T-Box describes generic properties about a category).

Given the concept:

$$\forall x \in \text{Human also } x \in \text{Mortal}$$

where `Human` and `Mortal` are categories, they have been represented by Google as simple records stored inside a database.

As we already know, the enterprise is very strong in database management tasks; instead of storing the axiom `Every human is mortal`, for each entity of `Human` category it has been decided to store in their databases the notion of **mortality** per individual.

This approach guarantees fast data retrieval.

Since reasoning is expensive by the computational point of view, we can just store the informations in terms of **nodes** and **arcs**. In this way, billions of factual knowledge are represented in form of **triples**, mixing together data coming from heterogeneous sources.

The construction of a knowledge graph is similar to semantic networks: boxes used to reprensent concepts and links used to describe relationships between boxes.

The quality of a knowledge graph is given by:

- **Coverage**. *Does the graph have all the required informations?*
- **Correctness**. *Is the information correct?*
- **Freshness**. *Is the content up-to-date?*

(The trustability of the informations is not done by knowledge graph, but instead it's on people judge their correctness).