Chapter 1
Introduction
Linked Data and the Semantic Web

NIKOLAOS KONSTANTINOU
DIMITRIOS-EMMANUEL SPANOS

Outline

Introduction

Preliminaries

The Linked Open Data Cloud

The Origin of the Semantic Web

Semantic Web

- Term primarily coined by Tim Berners-Lee
- For the Web to be understandable both by humans and software, it should incorporate its meaning
  - Its *semantics*

Linked Data implements the Semantic Web vision

Why a Semantic Web? (1)

Information management

- Sharing, accessing, retrieving, consuming information
  - Increasingly tedious task

Search engines

- Rely on Information Retrieval techniques
- Keyword-based searches
  - Do not release the information potential
    - Large quantity of existing data on the web is not stored in HTML
      - The Deep Web

Why a Semantic Web? (2)

Content meaning should be taken into account

- Structure to the meaningful Web page content
- Enable software agents to
  - Roam from page to page
  - Carry out sophisticated tasks for users

Semantic Web

- Complement, not replace current Web
- Transform content into an exploitable source of knowledge

Why a Semantic Web? (3)

Web of Data

- An emerging web of interconnected published datasets in the form of Linked Data
- Implements the Semantic Web vision

The Need for Adding Semantics (1)

Semantics

- From the Greek word *σημαντικός*
  - Pronounced *s̲imantikós*, means *significant*
- Term typically used to denote the study of meaning
- Using semantics
  - Capture the interpretation of a formal or natural language
  - Enables entailment, application of logical consequence
    - About relationships between the statements that are expressed in this language

The Need for Adding Semantics (2)

Syntax

- The study of the principles and processes by which sentences can be formed
- Also of Greek origin
  - συν and τάξις
  - Pronounced sin and taxis
  - Mean together and ordering, respectively

The Need for Adding Semantics (3)

Two statements can be syntactically different but semantically equivalent

- Their meaning (semantics) is the same, the syntax is different
  - More than one ways to state an assumption

The Need for Adding Semantics (4)

Search engines rely on keyword matches

- Keyword-based technologies
  - Have come a long way since the dawn of the internet
  - Will return accurate results
  - Based on keyword matches
    - Extraction and indexing of keywords contained in web pages
- But
  - More complex queries with semantics still partially covered
  - Most of the information will not be queried

Traditional search engines

Rely on keyword matches

Low precision

- Important results may not be fetched, or
- Ranked low
  - No exact keyword matches

Keyword-based searches

Do not return usually the desired results

Despite the amounts of information

Typical user behavior

- Change query rather than navigate beyond the first page of the search results

The Semantic Web (1)

Tackle these issues

- By offering ways to describe of Web resources

Enrich existing information

- Add semantics that specify the resources
- Understandable both by humans and computers

Information easier to discover, classify and sort

Enable semantic information integration

Increase serendipitous discovery of information

Allow inference

The Semantic Web (2)

Example

- A researcher would not have to visit every result page and distinguish the relevant from the irrelevant results
- Instead, retrieve a list of more related, machine-processable information
  - Ideally containing links to more relevant information

Outline

Introduction

Preliminaries

The Linked Open Data Cloud

Data-Information-Knowledge (1)

Data

> Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer

Information

- “facts about a situation, person, event, etc.”

Smallest information particle

- The bit
  - Replies with a yes or no (1 or 0)
  - Can carry data, but in the same time, it can carry the result of a process
  - For instance whether an experiment had a successful conclusion or not

Data-Information-Knowledge (2)

E.g. “temperature” in a relational database could be

- The information produced after statistical analysis over numerous measurements
  - By a researcher wishing to extract the average value in a region over the years
- A sensor measurement
  - Part of the data that will contribute in drawing conclusions regarding climate change

Data-Information-Knowledge (3)

Knowledge

Understanding of or information about a subject that you get by experience or study, either known by one person or by people generally

- Key term: “by experience or study”
  - Implies processing of the underlying information

Data-Information-Knowledge (4)

No clear lines between them

- Term that is used depends on the respective point of view

When we process collected data in order to extract meaning, we create new information

Addition of semantics

- Indispensable in generating new knowledge
- Semantic enrichment of the information
  - Makes it unambiguously understood by any interested party (by people generally)

Data ⟶ Information ⟶ Knowledge

Heterogeneity

Distributed data sources

- Interconnected, or not
- Different models and schemas
- Differences in the vocabulary
- Syntactic mismatch
- Semantic mismatch

Interoperability (1)

Interoperable

- Two systems in position to successfully exchange information

3 non-mutually exclusive approaches

- Mapping among the concepts of each source
- Intermediation in order to translate queries
- Query-based

Protocols and standards (recommendations) are crucial

- E.g. SOAP, WSDL, microformats

Interoperability (2)

Key concept: *Schema*

- Word comes from the Greek word σχήμα
  - Pronounced schíma
  - Means the shape, the outline
  - Can be regarded as a common agreement regarding the interchanged data
- Data schema
  - Defines how the data is to be structured

Information Integration (1)

Combine information from heterogeneous systems, sources of storage and processing

- Ability to process and handle it as a whole

Global-As-View

- Every element of the global schema is expressed as a query/view over the schemas of the sources
- Preferable when the source schemas are not subject to frequent changes

Information Integration (2)

Local-As-View

- Every element of the local schemas is expressed as a query/view over the global schema

P2P

- Mappings among the sources exist but no common schema

Information Integration Architecture (1)

A source Π_1 with schema S_1

A source Π_2 with a schema S_2

…

A source Π_v with source S_v

A global schema S in which the higher level queries are posed

Information Integration Architecture (2)

The goal in the information integration problem

- Submit queries to the global schema S
- Receive answers from S_1, S_2, …, S_v
- Without having to deal with or even being aware of the heterogeneity in the information sources

Data Integration (1)

A data integration system

- A triple $I = \langle G; S; M \rangle$
  - *G* is the global schema,
  - *S* is the source schema and
  - *M* is a set of mappings between *G* and *S*

Data Integration (2)

Local-As-View

- Each declaration in *M* maps an element from the source schema *S* to a query (a view) over the global schema *G*

Global-As-View

- Each declaration in *M* maps an element of the global schema *G* to a query (a view) over the source schema *S*

The global schema *G* is a unified view over the heterogeneous set of data sources

Data Integration (3)

Same goal

- Unifying the data sources under the same common schema

Semantic information integration

- Addition of its semantics in the resulting integration scheme

Mapping

Using the definition of data integration systems, we can define the concept of mapping

- A mapping *m* (member of *M*) from a schema *S* to a schema *T*
  - A declaration of the form $Q^S \rightsquigarrow Q^T$
  - Q^S is a query over *S*
  - Q^T a query over *T*

Mapping vs. Merging

(Data) mapping

- A mapping is the specification of a mechanism
  - The members of a model are transformed to members of another model
- The meta-model that can be the same, or different
- A mapping can be declared as a set of relationships, constraints, rules, templates or parameters
  - Defined during the mapping process, or through other forms that have not yet been defined

Merging

- Implies unifying the information at the implementation/storage layer

Annotation (1)

Addition of metadata to the data

Data can be encoded in any standard

Especially important in cases when data is not human-understandable in its primary form

- E.g. multimedia

Annotation (2)

(Simple) annotation

- Uses keywords or other ad hoc serialization
  - Impedes further processing of the metadata

Semantic annotation

- Describe the data using a common, established way
- Addition of the semantics in the annotation (i.e. the metadata)
- In a way that the annotation will be machine-processable
  - In order to be able to infer additional knowledge

Problems with (Semantic) Annotation

Time-consuming

- Users simply do not have enough time or do not consider it important enough in order to invest time to annotate their content

Familiarity required with both the conceptual and technical parts of the annotation

- User performing the annotation must be familiar with both the technical as well as the conceptual part

Outdated annotations

- A risk, especially in rapidly changing environments with lack of automation in the annotation process

Automated Annotation

Incomplete annotation is preferable to absent

- E.g. in video streams

Limitations of systems performing automated annotation

- Limited recall and precision (lost or inaccurate annotations)

Metadata (1)

Data about data

Efficient materialization with the use of ontologies

Ontologies can be used to describe Web resources

- Adding descriptions about their semantics and their relations
- Aims to make resources machine-understandable
- Each (semantic) annotation can correspond to a piece of information
- It is important to follow a common standard

Metadata (2)

Annotation is commonly referred to as “metadata”

Usually in a semi-structured form

- Semi-structured data sources
  - Structure accompanies the metadata
  - E.g. XML, JSON
- Structured data sources
  - Structure is stored separately
  - E.g. relational databases

Metadata (3)

Powerful languages

- Practically unlimited hierarchical structure
- Covers most of the description requirements that may occur
- Storage in separate files allows collaboration with communication protocols
- Files are independent from the environment in which they reside
  - Resilience to technological evolutions

Negatives

- Expressivity of the description model
- Limited way of structuring the information

Metadata (4)

Inclusion of semantics

- Need for more expressive capabilities and terminology

Ontologies

- Offer a richer way of describing information
  - E.g. defining relationships among concepts such as subclass/superclass, mutually disjoint concepts, inverse concepts, etc.
- RDF can be regarded as the evolution of XML
- OWL enables more comprehensive, precise and consistent description of Web Resources

Ontologies (1)

Additional description to an unclear model that aims to further clarify it

Conceptual description model of a domain

- Describe its related concepts and their relationships

Can be understood both by human and computer

- A shared conceptualization of a given specific domain

Aim at bridging and integrating multiple and heterogeneous digital content on a semantic level

Ontologies (2)

In Philosophy, a systematic recording of “Existence”

For a software system, something that “exists” is something that can be represented

The specification of a conceptualization (Gruber 1995)

- A set of definitions that associate the names of entities in the universe of discourse with human-readable text describing the meaning of the names
- A set of formal axioms that constrain the interpretation and well-formed use of these terms

Ontologies (3)

Can be used to model concepts regardless to how general or specific these concepts are

According to the degree of generalization

- Top-level ontology
- Domain ontology
- Task ontology
- Application ontology

Content is made suitable for machine consumption

- Automated increase of the system knowledge
- Logical rules to infer implicitly declared facts

Reasoners (1)

Software components

Validate consistency of an ontology

- Perform consistency checks
- Concept satisfiability and classification

Infer implicitly declared knowledge

- Apply simple rules of deductive reasoning

Reasoners (2)

Basic reasoning procedures

- Consistency checking
- Concept satisfiablility
- Concept subsumption
- Instance checking (Realization)

Properties of special interest

- Termination
- Soundness
- Completeness

Reasoners (3)

Research in Reasoning

- Tradeoff between
  - Expressiveness of ontology definition languages
  - Computational complexity of the reasoning procedure
- Discovery of efficient reasoning algorithms

Commercial

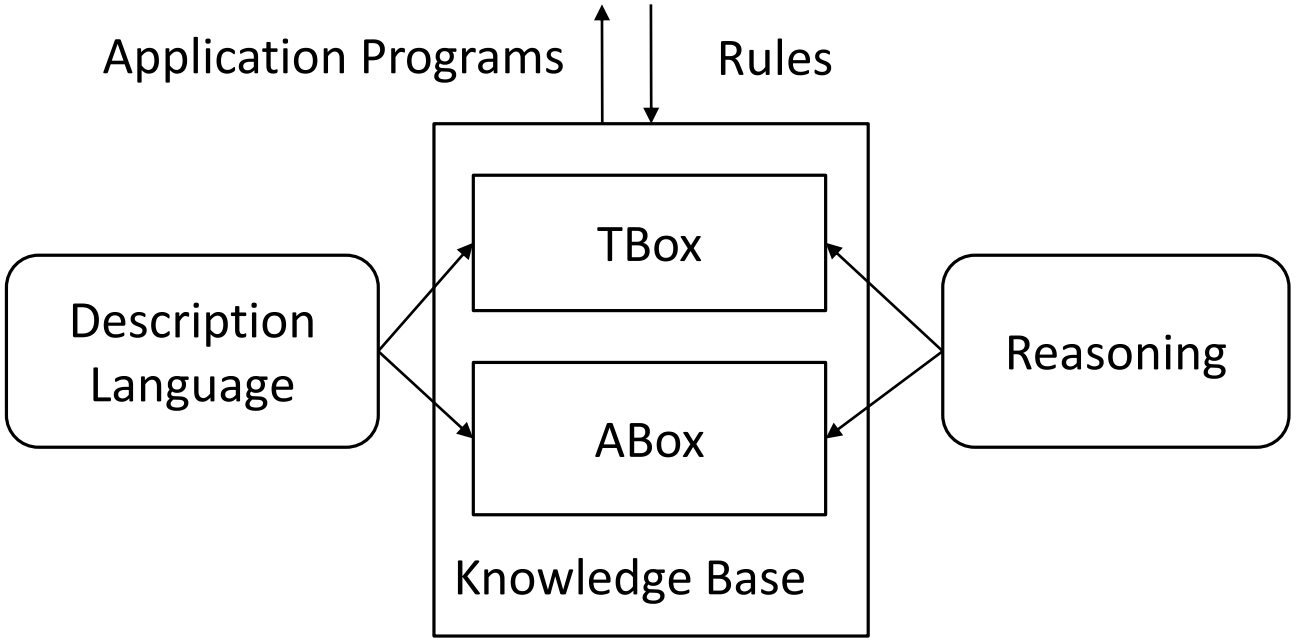
- E.g. RacerPro

Free of charge

- E.g. Pellet, FaCT++

Knowledge Bases

- Terminological Box (TBox)
  - Concept descriptions
  - Intensional knowledge
  - Terminology and vocabulary
- Assertional Box (ABox)
  - The real data
  - Extensional knowledge
  - Assertions about named individuals in terms of the TBox vocabulary



Knowledge Bases vs. Databases (1)

A naïve approach:

- TBox ≡ Schema of the relational database
- ABox ≡ Schema instance

However, things are more complex than that

Knowledge Bases vs. Databases (2)

Relational model

- Supports only untyped relationships among relations
- Does not provide enough features to assert complex relationships among data
- Used in order to manipulate large and persistent models of relatively simple data

Ontological scheme

- Allows more complex relationships
- Can provide answers about the model that have not been explicitly stated to it, with the use of a reasoner
- Contains fewer but more complex data

Closed vs. Open World Assumption (1)

Closed world assumption

- Relational databases

- Everything that has not been stated as true is false
  - What is not currently known to be true is false
- A null value about a subject's property denotes the non-existence
  - A NULL value in the isCapital field of a table Cities claims that the city is not a capital
- The database answers with certainty
  - A query "select cities that are capitals" will not return a city with a null value at a supposed boolean isCapital field

Closed vs. Open World Assumption (2)

Open world assumption

- Knowledge Bases
- A query can return three types of answers
  - True, false, cannot tell
- Information that is not explicitly declared as true is not necessarily false
  - It can also be unknown
  - Lack of knowledge does not imply falsity
- A question “Is Athens a capital city?” in an appropriate schema will return “cannot tell” if the schema is not informed
  - A database schema would return false, in the case of a null value

Monotonicity

A feature present in Knowledge Bases

A system is considered monotonic when new facts do not discard existing ones

First version of SPARQL did not include update/delete functionality

- Included in SPARQL 1.1
  - Recommendation describes that these functions should be supported
    - Does not describe the exact behavior

Outline

Introduction

Preliminaries

The Linked Open Data Cloud

The LOD Cloud (1)

Structured data

Open format

Available for everyone to use it

Published on the Web and connected using Web technologies

Related data that was not previously linked

- Or was linked using other methods

The LOD Cloud (2)

Using URIs and RDF for this is goal is very convenient

- Data can be interlinked
- Create a large pool of data
- Ability to search, combine and exploit
- Navigate between different data sources, following RDF links
  - Browse a potentially endless Web of connected data sources

The LOD Cloud (3)

Applications in many cases out of the academia

Technology maturity

Open state/government data

- data.gov (US)
- data.gov.uk (UK)
- data.gov.au (Australia)
- opengov.se (Sweden)

Open does not necessarily mean Linked

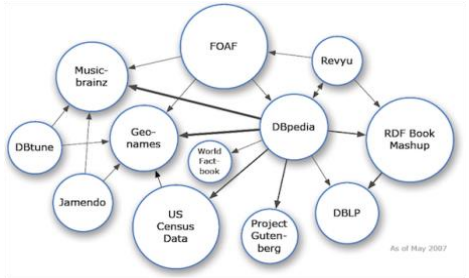
The LOD Cloud (4)

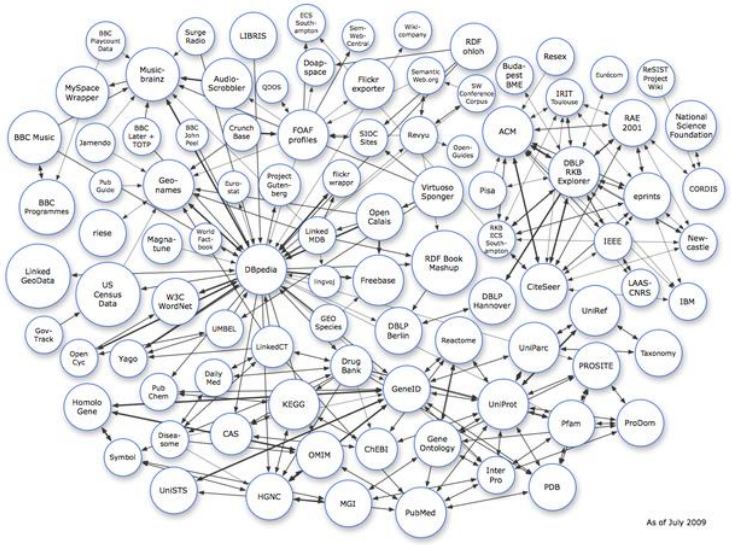
Published datasets span several domains of human activities

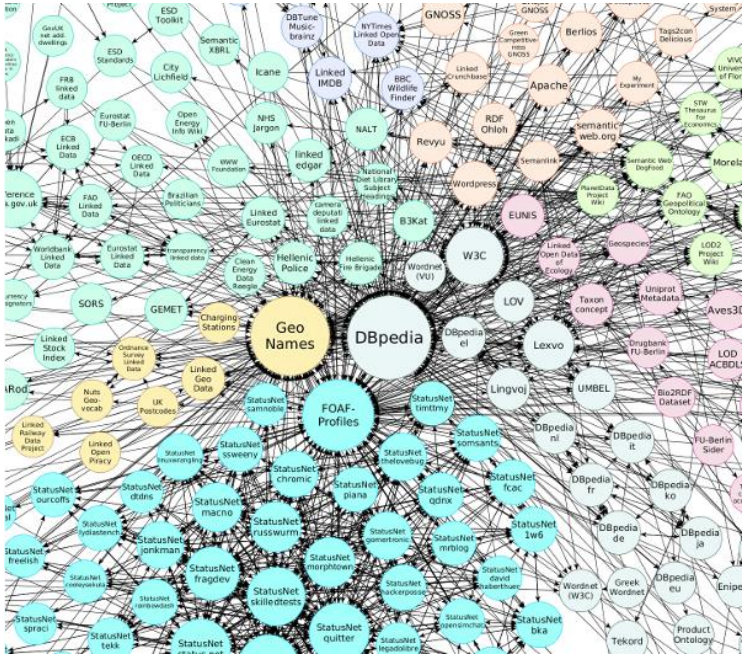
- Much more beyond government data
- Form the LOD cloud
  - Constantly increasing in terms of volume

The LOD Cloud (5)

Evolution







Media
Geographic
Publications
User-Generated Content
Government
Cross-Domain
Life Sciences
Social Networking

2007 2009 2014