A Research Article

2015313825 임정우

#NoSQL

#Neo4j

#DB

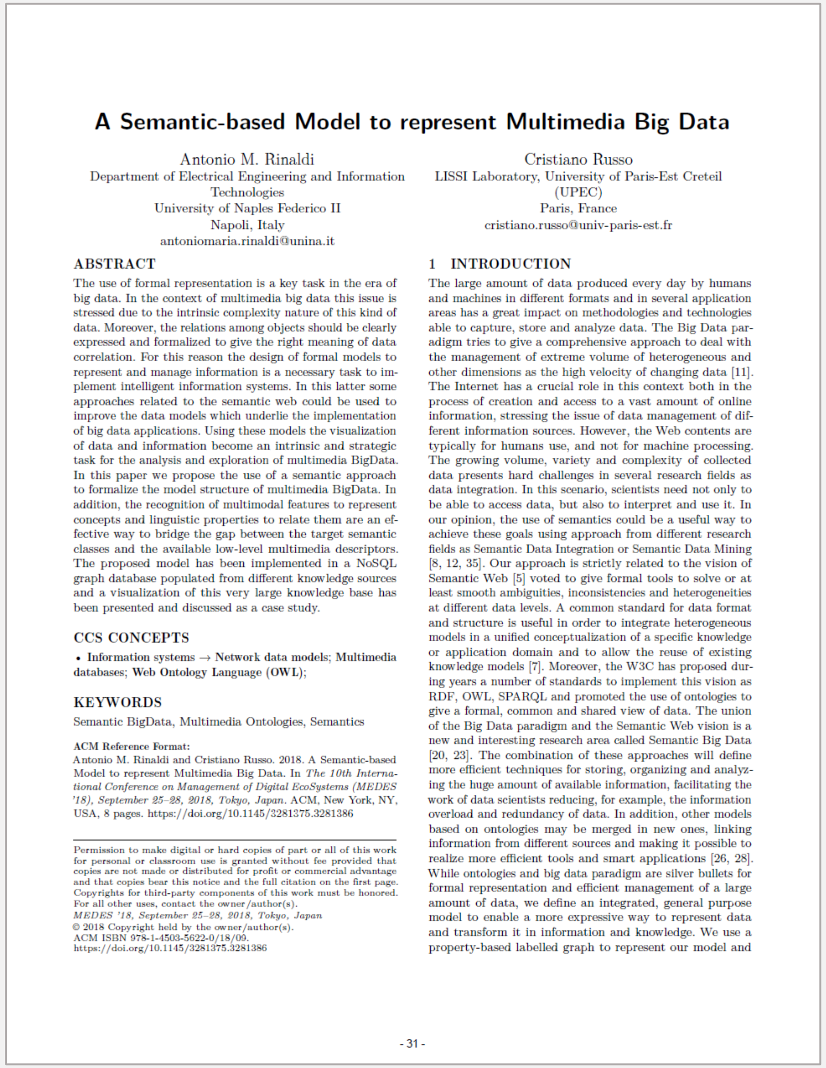
#GraphDB

#Semantic

#Ontology

Title of the Research Article

The brief exploration of this presentation



A Semantic-based Model to represent Multimedia Big Data

Antonio M. Rinaldi
Department of Electrical Engineering and Information Technologies
University of Naples Federico II
Napoli, Italy
antoniomaria.rinaldi@unina.it

Cristiano Russo
LISSI Laboratory, University of Paris-Est Creteil (UPEC)
Paris, France
cristiano.russo@univ-paris-est.fr

ABSTRACT

The use of formal representation is a key task in the era of big data. In the context of multimedia big data this issue is stressed due to the intrinsic complexity nature of this kind of data. Moreover, the relations among objects should be clearly expressed and formalized to give the right meaning of data correlation. For this reason the design of formal models to represent and manage information is a necessary task to implement intelligent information systems. In this latter some approaches related to the semantic web could be used to improve the data models which underlie the implementation of big data applications. Using these models the visualization of data and information become an intrinsic and strategic task for the analysis and exploration of multimedia BigData. In this paper we propose the use of a semantic approach to formalize the model structure of multimedia BigData. In addition, the recognition of multimodal features to represent concepts and linguistic properties to relate them are an effective way to bridge the gap between the target semantic classes and the available low-level multimedia descriptors. The proposed model has been implemented in a NoSQL graph database populated from different knowledge sources and a visualization of this very large knowledge base has been presented and discussed as a case study.

CCS CONCEPTS

- **Information systems** → **Network data models; Multimedia databases; Web Ontology Language (OWL);**

KEYWORDS

Semantic BigData, Multimedia Ontologies, Semantics

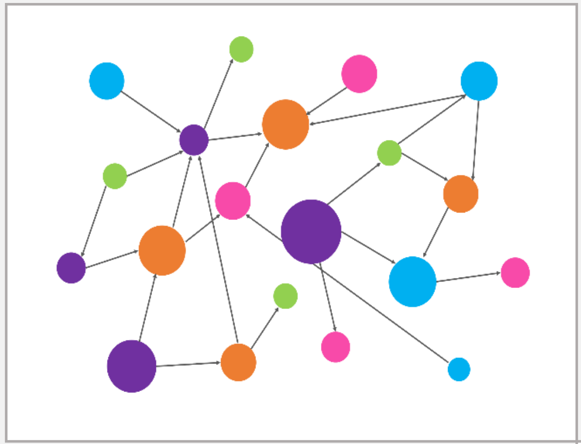
ACM Reference Format:
Antonio M. Rinaldi and Cristiano Russo. 2018. A Semantic-based Model to represent Multimedia Big Data. In *The 10th International Conference on Management of Digital EcoSystems (MEDES '18), September 25–28, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3281375.3281386

ACM ISBN 978-1-4503-5622-0/18/09.
https://doi.org/10.1145/3281375.3281386

1 INTRODUCTION

The large amount of data produced every day by humans and machines in different formats and in several application areas has a great impact on methodologies and technologies able to capture, store and analyze data. The Big Data paradigm tries to give a comprehensive approach to deal with the management of extreme volume of heterogeneous and other dimensions as the high velocity of changing data [11]. The Internet has a crucial role in this context both in the process of creation and access to a vast amount of online information, stressing the issue of data management of different information sources. However, the Web contents are typically for humans use, and not for machine processing. The growing volume, variety and complexity of collected data presents hard challenges in several research fields as data integration. In this scenario, scientists need not only to be able to access data, but also to interpret and use it. In our opinion, the use of semantics could be a useful way to achieve these goals using approach from different research fields as Semantic Data Integration or Semantic Data Mining [8, 12, 35]. Our approach is strictly related to the vision of Semantic Web [5] voted to give formal tools to solve or at least smooth ambiguities, inconsistencies and heterogeneities at different data levels. A common standard for data format and structure is useful in order to integrate heterogeneous models in a unified conceptualization of a specific knowledge or application domain and to allow the reuse of existing knowledge models [7]. Moreover, the W3C has proposed during years a number of standards to implement this vision as RDF, OWL, SPARQL and promoted the use of ontologies to give a formal, common and shared view of data. The union of the Big Data paradigm and the Semantic Web vision is a new and interesting research area called Semantic Big Data [20, 23]. The combination of these approaches will define more efficient techniques for storing, organizing and analyzing the huge amount of available information, facilitating the work of data scientists reducing, for example, the information overload and redundancy of data. In addition, other models based on ontologies may be merged in new ones, linking information from different sources and making it possible to realize more efficient tools and smart applications [26, 28]. While ontologies and big data paradigm are silver bullets for formal representation and efficient management of a large amount of data, we define an integrated, general purpose model to enable a more expressive way to represent data and transform it in information and knowledge. We use a property-based labelled graph to represent our model and

A Semantic-based Model to represent Multimedia Big Data



Semantic-based Model

Property-based graph structure with an implicit agreement about the meaning of edges, nodes, labels and properties.

Multimedia Big Data

With its description of its components and its properties.

Index

The index of this presentation

Introduction

What we need to know ?

- #0. Bigdata
- #1. Semantic
- #2. Multimedia

The Proposed Model

Property-based graph

- #3. Ontological
- #4. Property-based

Case Study

How did they implement model in NoSQL graph database?

- #D3
- #Neo4j

Introduction - #0. Big Data

'5V 'of the 'Big Data'

5V in Big Data

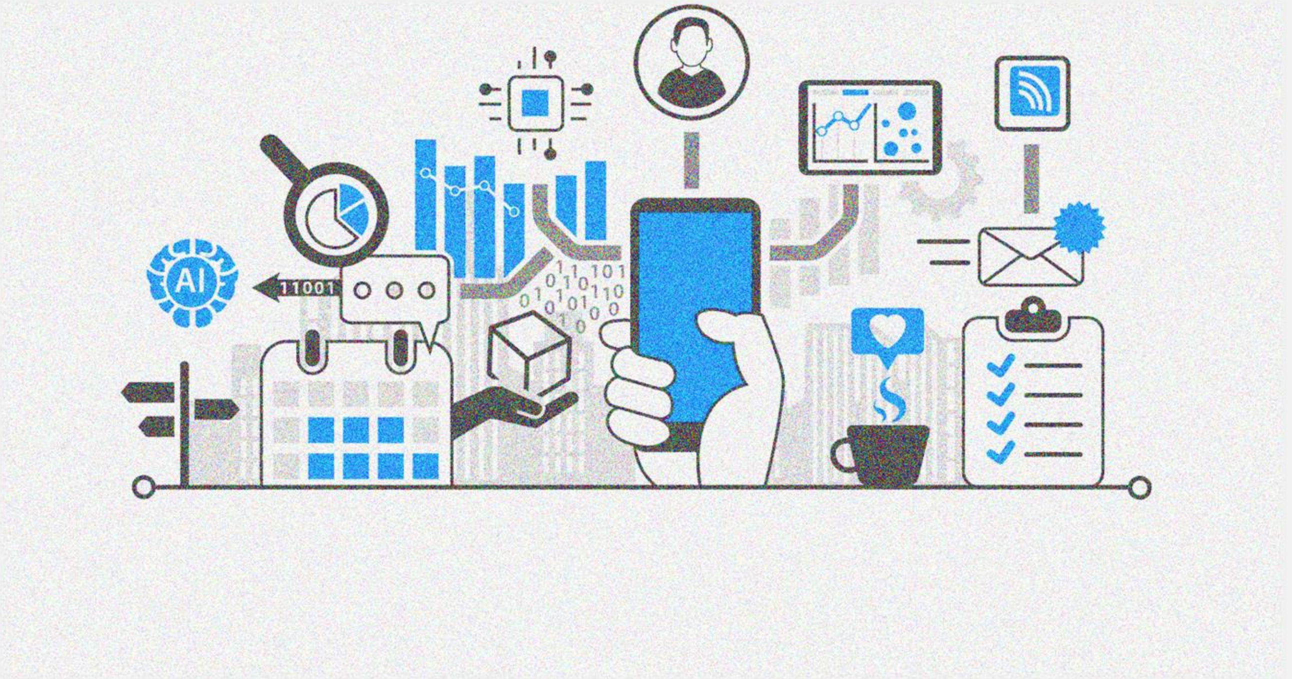
V

V

V

V

V



Introduction - #0. Big Data

'5V 'of the 'Big Data'

5V in Big Data

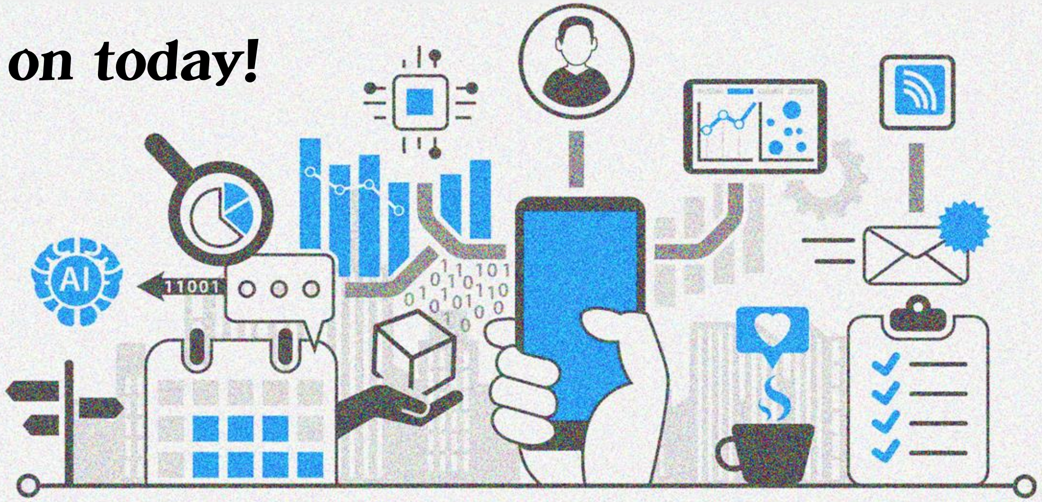
Volume

Velocity

Variety → **What we need to focus on today!**

Value

Veracity



Introduction - #1. Semantic

'5V 'of the 'Big Data'

Semantic?

: **(of words and language) connected with meaning:**

From https://dictionary.cambridge.org

" The use of semantics could be a useful way to achieve goal that is

Integration of high volume, variety and complexity of data

using approach from different research field as Semantic Data Integration or Semantic Data Mining"

Introduction - #2. Multimedia

'5V 'of the 'Big Data'

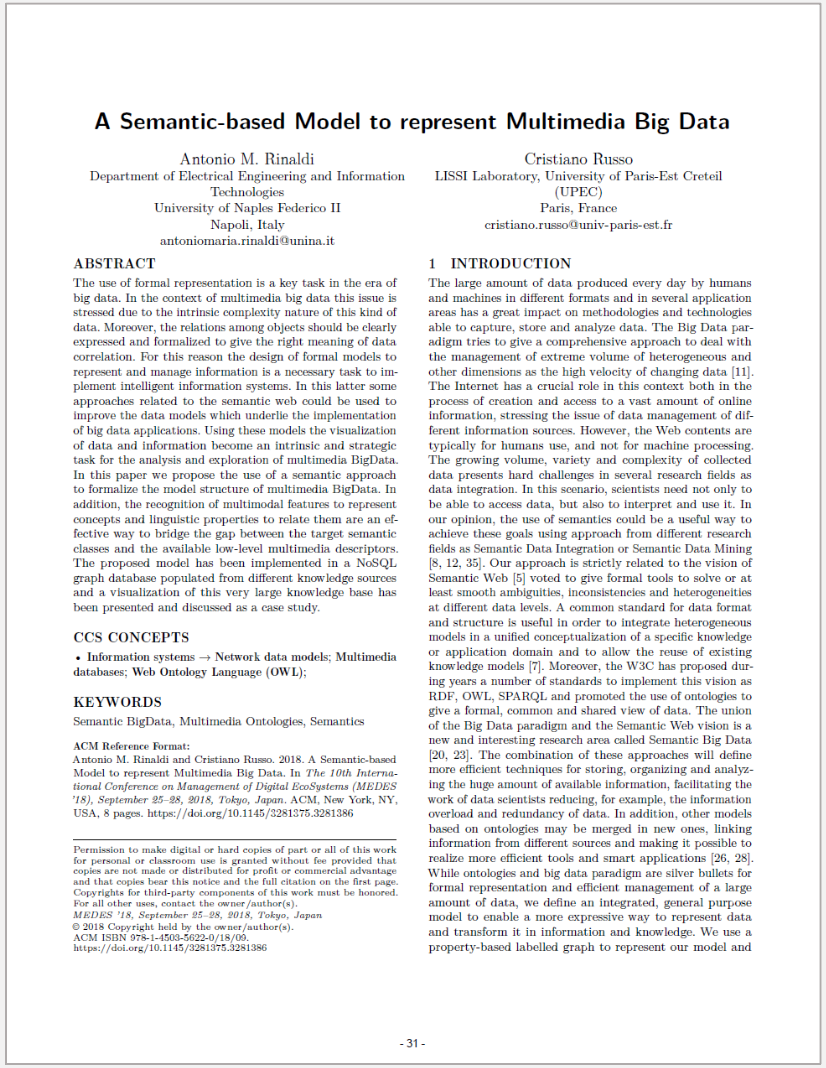
Multimedia?

: the integration of multiple forms of media. From https://techterms.com/

Ex) text, graphics, audio, video, etc.

Title of the Research Article

The brief exploration of this presentation



A Semantic-based Model to represent Multimedia Big Data

Antonio M. Rinaldi
Department of Electrical Engineering and Information Technologies
University of Naples Federico II
Napoli, Italy
antoniomaria.rinaldi@unina.it

Cristiano Russo
LISSI Laboratory, University of Paris-Est Creteil (UPEC)
Paris, France
cristiano.russo@univ-paris-est.fr

ABSTRACT

The use of formal representation is a key task in the era of big data. In the context of multimedia big data this issue is stressed due to the intrinsic complexity nature of this kind of data. Moreover, the relations among objects should be clearly expressed and formalized to give the right meaning of data correlation. For this reason the design of formal models to represent and manage information is a necessary task to implement intelligent information systems. In this latter some approaches related to the semantic web could be used to improve the data models which underlie the implementation of big data applications. Using these models the visualization of data and information become an intrinsic and strategic task for the analysis and exploration of multimedia BigData. In this paper we propose the use of a semantic approach to formalize the model structure of multimedia BigData. In addition, the recognition of multimodal features to represent concepts and linguistic properties to relate them are an effective way to bridge the gap between the target semantic classes and the available low-level multimedia descriptors. The proposed model has been implemented in a NoSQL graph database populated from different knowledge sources and a visualization of this very large knowledge base has been presented and discussed as a case study.

CCS CONCEPTS

- **Information systems** → **Network data models; Multimedia databases; Web Ontology Language (OWL);**

KEYWORDS

Semantic BigData, Multimedia Ontologies, Semantics

ACM Reference Format:
Antonio M. Rinaldi and Cristiano Russo. 2018. A Semantic-based Model to represent Multimedia Big Data. In *The 10th International Conference on Management of Digital EcoSystems (MEDES '18), September 25–28, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3281375.3281386

ACM ISBN 978-1-4503-5622-0/18/09.
https://doi.org/10.1145/3281375.3281386

1 INTRODUCTION

The large amount of data produced every day by humans and machines in different formats and in several application areas has a great impact on methodologies and technologies able to capture, store and analyze data. The Big Data paradigm tries to give a comprehensive approach to deal with the management of extreme volume of heterogeneous and other dimensions as the high velocity of changing data [11]. The Internet has a crucial role in this context both in the process of creation and access to a vast amount of online information, stressing the issue of data management of different information sources. However, the Web contents are typically for humans use, and not for machine processing. The growing volume, variety and complexity of collected data presents hard challenges in several research fields as data integration. In this scenario, scientists need not only to be able to access data, but also to interpret and use it. In our opinion, the use of semantics could be a useful way to achieve these goals using approach from different research fields as Semantic Data Integration or Semantic Data Mining [8, 12, 35]. Our approach is strictly related to the vision of Semantic Web [5] voted to give formal tools to solve or at least smooth ambiguities, inconsistencies and heterogeneities at different data levels. A common standard for data format and structure is useful in order to integrate heterogeneous models in a unified conceptualization of a specific knowledge or application domain and to allow the reuse of existing knowledge models [7]. Moreover, the W3C has proposed during years a number of standards to implement this vision as RDF, OWL, SPARQL and promoted the use of ontologies to give a formal, common and shared view of data. The union of the Big Data paradigm and the Semantic Web vision is a new and interesting research area called Semantic Big Data [20, 23]. The combination of these approaches will define more efficient techniques for storing, organizing and analyzing the huge amount of available information, facilitating the work of data scientists reducing, for example, the information overload and redundancy of data. In addition, other models based on ontologies may be merged in new ones, linking information from different sources and making it possible to realize more efficient tools and smart applications [26, 28]. While ontologies and big data paradigm are silver bullets for formal representation and efficient management of a large amount of data, we define an integrated, general purpose model to enable a more expressive way to represent data and transform it in information and knowledge. We use a property-based labelled graph to represent our model and

A **Semantic-based** Model

to represent **Multimedia** Big Data

Multimedia

Semantic ETL process →

Semantic-based model

The Proposed Model

The Proposed Model

Property-based graph

#3. Ontological

#4. Property-based

The Proposed Model

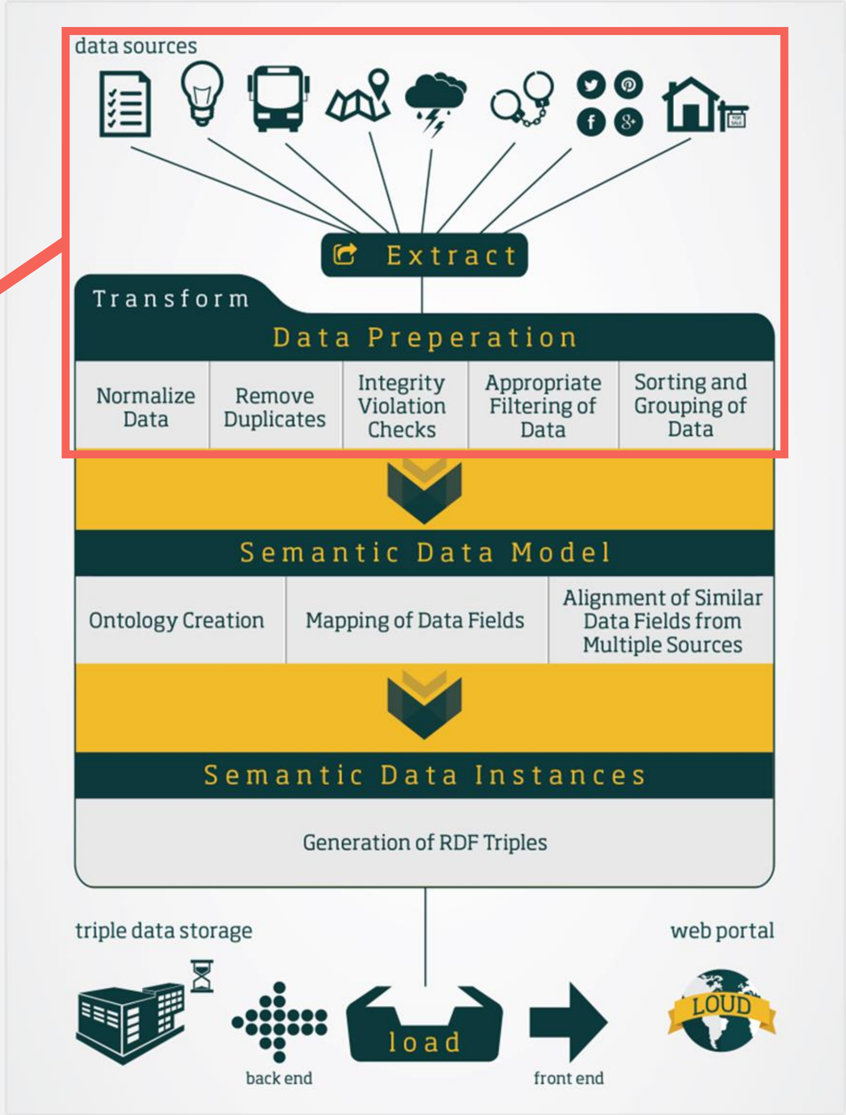
The brief exploration of this presentation

1. ETL process?

- **Extract:** this is the first phase of the process that involves data extraction from appropriate data sources.

- **Transform:** this phase involves the cleansing of data to comply with the target schema.

- **Load:** this phase involves the propagation of the data into a data mart or a data warehouse that serves Big Data.



The Proposed Model

The brief exploration of this presentation

2. Semantic ETL process?

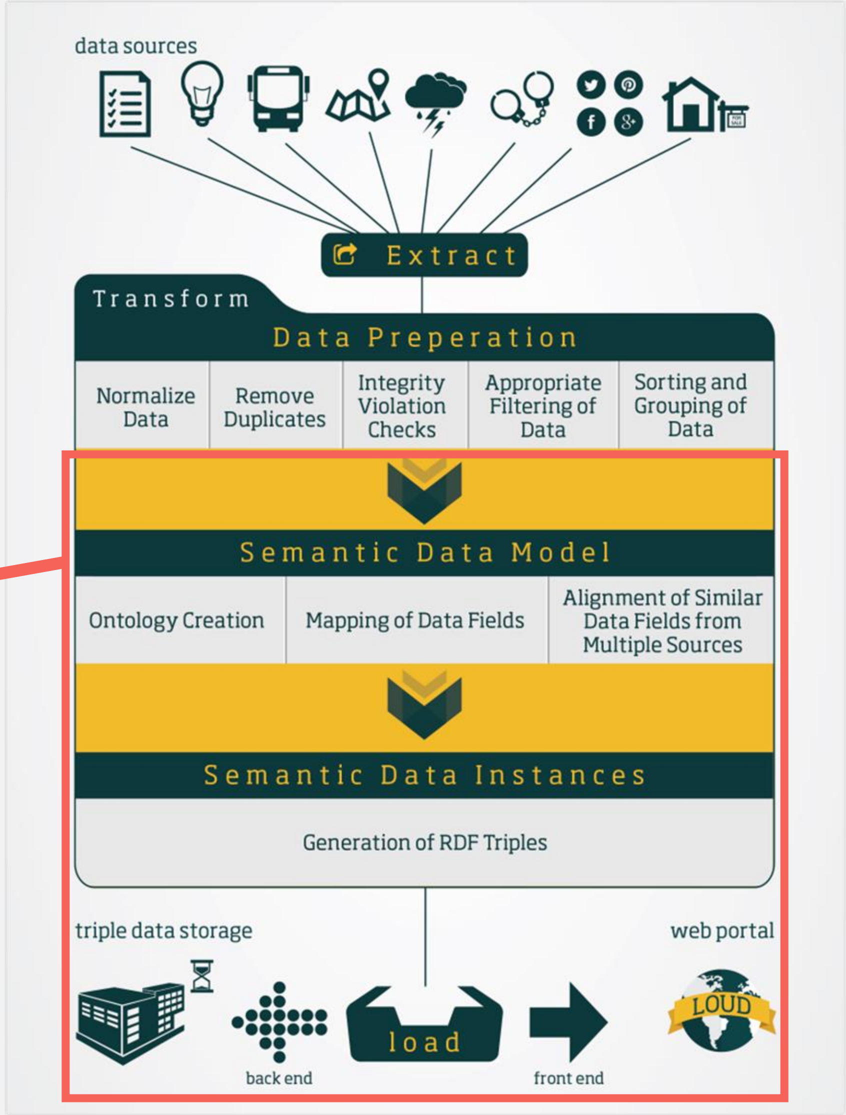
- **Extract:** this is the first phase of the process that involves data extraction from appropriate data sources.

- **Transform:** will involve a manual process of analyzing the datasets, the schema and their purpose.

- **Load:** this phase involves the propagation of the data into a data mart or a data warehouse that serves Big Data.

- **Semantic ETL framework:** generates a semantic model of the dataset(s) under integration, and then generates semantic linked data in compliance with the data model.

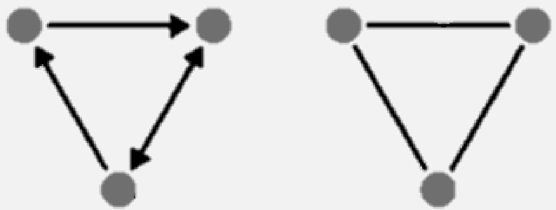
- **Semantic Data:** is made available on the web as linked data available for querying, analytics, or used in data-driven innovative apps. (RDF triples) It is to be stored in a data mart or warehouse.

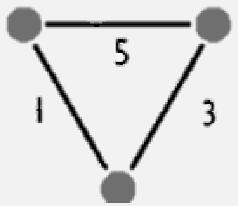


3. Proposed Model – property-based graph

- **The proposed model** to represent multimedia big data is presented with a <u>description of its components and properties</u>.

- **Property-based graph:** which allows users to represent concepts and logical relations between them through a graph-based structure with an implicit agreement about the meaning of edges, nodes, labels and properties.





Then, how do we make mapping using edges, nodes, labels and properties from multimedia?

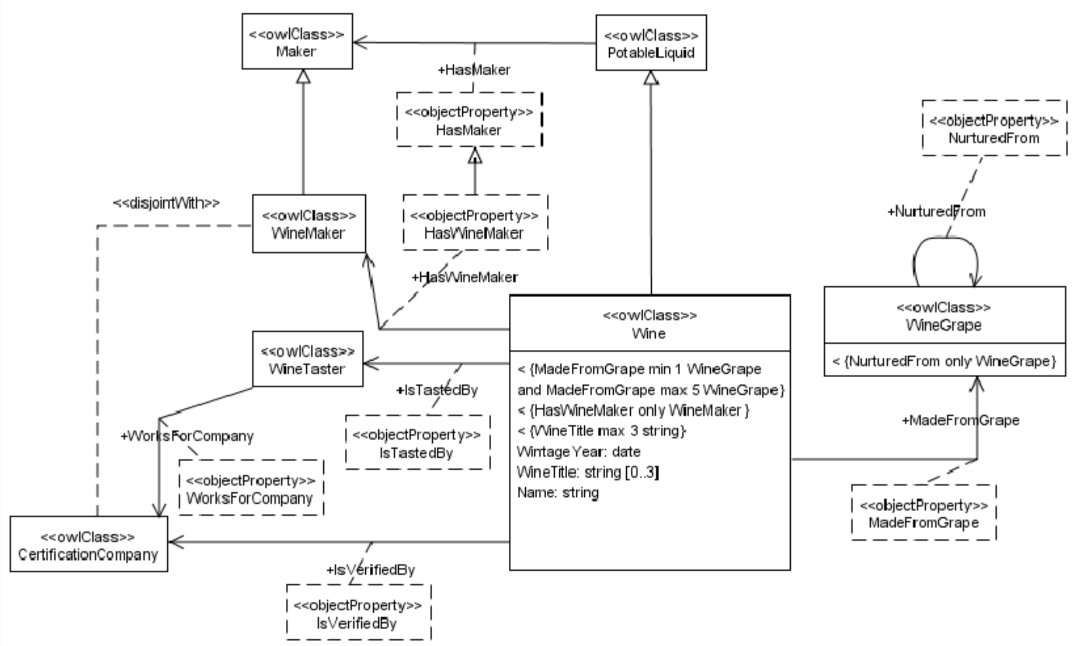
3. Proposed Model – Ontological model formalization

- **Ontology:** A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.

https://en.oxforddictionaries.com/

- **OWL(Web Ontology Language):** Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.

https://www.w3.org/OWL/



The proposed model is composed of a triple ⟨S, P, C⟩ where:

S is a set of signs;
P is a set of properties used to link the signs in S;
C is a set of constraints on P

3. Proposed Model – Ontological model formalization

- Two main classes in this meta model

1. Concept: in which all objects are defined as individuals; represents the meaning related to a multimedia form.

2. MM: which represents all the "signs" in the ontology; logical model of the multimedia forms used to express a concept.

Signs(Nodes)

- ObjectProperties in Semantic network

1. Linguisitic properties: Lexical properties and Semantic properties — **Properties(Edge)**

- Constraints

1. Model constraints: It expresses the ways that the linguistic properties are used to relate concepts and/or MM — **Constraints(Properties)**

3. Proposed Model – Ontological model formalization

– Two main classes in this meta model

1. Concept: in which all objects are defined as individuals; represents the meaning related to a multimedia form.

2. MM: which represents all the "signs" in the ontology; logical model of the multimedia forms used to express a concept.

Signs(Nodes)

Examples:

Concept: Name(represents the concept name), Description(gives a short description of concept)

Visual: Name(attribute that is the MM name), Set of features

Data Type	Features
Visual	Auto Color Correlogram(ACC), Scalable Color(SC) Fuzzy Color and Texture Histogram (FCTH), Color Layout(CL), Edge Histogram (EH), Color and Edge Directivity Descriptor (CEDD), Joint-Composite Descriptor(JCD), Pyramid Histogram of Oriented Gradients (PHOG)

Set of features

3. Proposed Model – Ontological model formalization

– ObjectProperties in Semantic network

1. Linguisitic properties: Lexical properties and Semantic properties — **Properties(Edge)**

Lexical properties	synonym, antonym, pertainym, nominalization, derived from adjective, participle of verb
Semantic properties	hypernyms,hyponyms, coordinate terms, holonym, meronym, hypernym, troponym, entailment, related nouns, similar to, coordinate terms, Participle of verb, root adjectives

– Constraints

1. Model constraints: It expresses the ways that the linguistic properties are used to relate concepts and/or MM — **Constraints(Properties)**

Costraint	Class	Property	Constraint range
AllValuesFrom	NounConcept	hyponym	NounConcept
AllValuesFrom	AdjectiveConcept	attribute	NounConcept
AllValuesFrom	NounWord	synonym	NounWord
AllValuesFrom	VerbWord	also_see	VerbWord

3. Proposed Model – Property-based graph model formalization

1. Nodes' Labelled with concept

- *id, a number used to identify each concept in the database.*
- *sid, the id used in WordNet for that concept.*
- *pos, the part of speech of the concept. It can be a noun, a verb, an adjective or an adverb.*
- *lemmas, the list of english terms used to represent the concept.*
- *glossary, a brief description of the concept, explaining its meaning.*

3. Proposed Model – Property-based graph model formalization

2. Attributes of the Nodes with the label multimedia

- *id, a number used to identify each multimedia in the database.*
- *url, the path to the multimedia file representing the concept.*
- *PHOG, the extracted Pyramid Histogram of Oriented Gradients (PHOG) feature vector.*
- *JCD, the extracted Joint-Composite Descriptor(JCD) feature vector.*
- *CEDD, the extracted Color and Edge Directivity Descriptor (CEDD) feature vector.*
- *SC, the extracted Scalable Color(SC) feature vector.*
- *EH, the extracted Edge Histogram (EH) feature vector.*
- *FCTH, the path to the Fuzzy Color and Texture Histogram (FCTH) feature vector.*
- *CL, the path to the Color Layout(CL) feature vector.*
- *ACC, the path to the Auto Color Correlogram(ACC) feature vector.*

3. Proposed Model – Property-based graph model formalization

3. Attributes of Relationships

- *id, univoque identifier of relation in the graph database*
- *type, the type of relation, i.e. hyponym, antonym, has-Concept, etc..*
- *weight, a number in the interval 0, 1 used to assign a strength level to each type of relation and consequently giving the possibility to perform similarity metrics based on weighted distances.*

3. Proposed Model – Property-based graph model formalization

Definition 1. **(Multimedia to Multimedia relationship MM)**

Let $\hat{M} \subset M \subset V$ *and* $\mathring{M} \subset M \subset V$, *with* $\hat{M} \cup \mathring{M} = \varnothing$, *two disjoint subsets of nodes of* M *in* Γ, *a multimedia to multimedia relationship is a weighted arc* $e_i \subset E$ *with a weight* $w \longrightarrow 0, 1$ *connecting the nodes in subset* $\hat{M}$ *with nodes in subset* $\mathring{M}$.

3. Proposed Model – Property-based graph model formalization

Definition 2. **(Concept to Multimedia relationship C-M)**

Let $\hat{C} \subset C \subset V$ *and* $\hat{M} \subset M \subset V$ *two subsets of* C *and* M *respectively, in* Γ*, a concept to multimedia relationship is a weighted arc* $e_i \subset E$ *with a weight* $w \longrightarrow 0, 1$ *connecting the nodes in* $\hat{C}$ *with nodes in* $\hat{M}$.

3. Proposed Model – Property-based graph model formalization

Definition 3. (Multimedia to Concept relationship M-C)

Let $\hat{M} \subset M \subset V$ and $\hat{C} \subset C \subset V$ two subsets of nodes of M and C respectively, in Γ, a multimedia to concept relationship is a weighted hyperarc $e_i \subset E$ with a weight $w \longrightarrow 0, 1$ connecting the nodes in $\hat{M}$ with nodes in $\mathring{C}$.

3. Proposed Model – Property-based graph model formalization

Definition 4. **(Concept to Concept relationship C-C)**

Let $\hat{C} \subset C \subset V$ *and* $\mathring{C} \subset C \subset V$, *with* $\hat{C} \cup \mathring{C} = \varnothing$, *two disjoint subsets of* C *in* Γ, *a concept to concept relationship is a weighted hyperarc* $e_i \subset E$ *with a weight* $w \longrightarrow 0, 1$ *connecting the nodes in* $\hat{C}$ *with nodes in* $\mathring{C}$.

Case Study

Case study

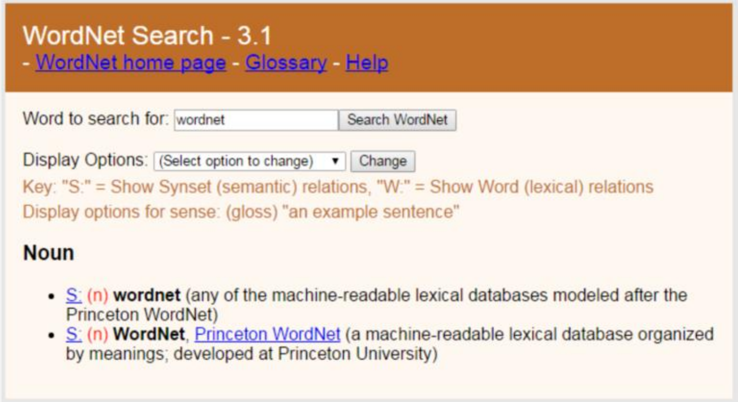
#5. Neo4j

#6. D3

4. Case Study

- **JWNL:** JWNL is a Java API for accessing the WordNet relational dictionary. WordNet is widely used for developing NLP applications, and a Java API such as JWNL will allow developers to more easily use Java for building NLP applications.

- **WordNet:** a machine-readable lexical database organized by meanings; developed at Princeton University



4. Case Study

- **ImageNet:** is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images.

- **BabelNet:** today's most far-reaching multilingual resource that covers hundreds of languages and, according to need, can be used as either an encyclopedic dictionary, or a semantic network, or a huge knowledge base.

4. Case Study

- **LIRE:** LIRE is a Java library that provides a simple way to retrieve images and photos based on color and texture characteristics

- **Neo4j:** A platform to manage graph database

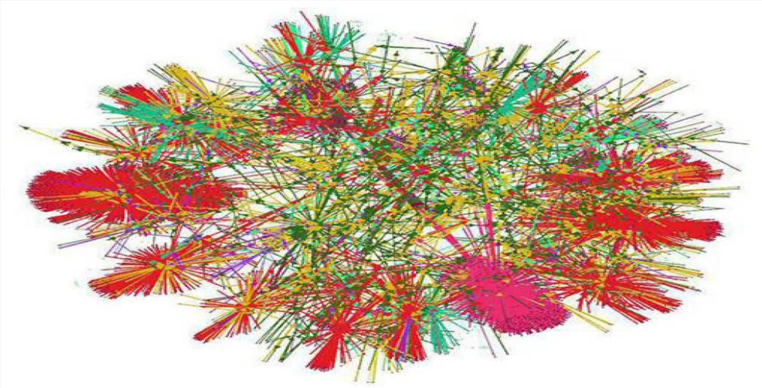
4. Case Study

- **Gephi:** an open-source software for visualization and exploration of graphs and networks.

Gephi

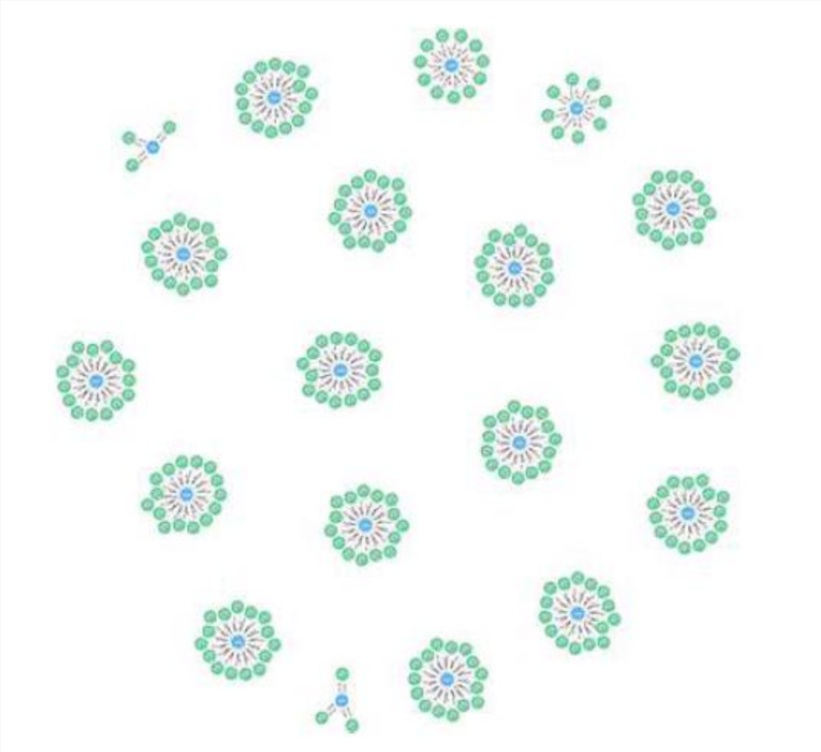
4. Case Study

- **Step 1** Create the CSV files containing nodes and relationships
- **Step 2** Then, we loaded in Neo4j nodes and relationship
- **Step 3** Using Gephi, make some visualization of the graph.



4. Case Study

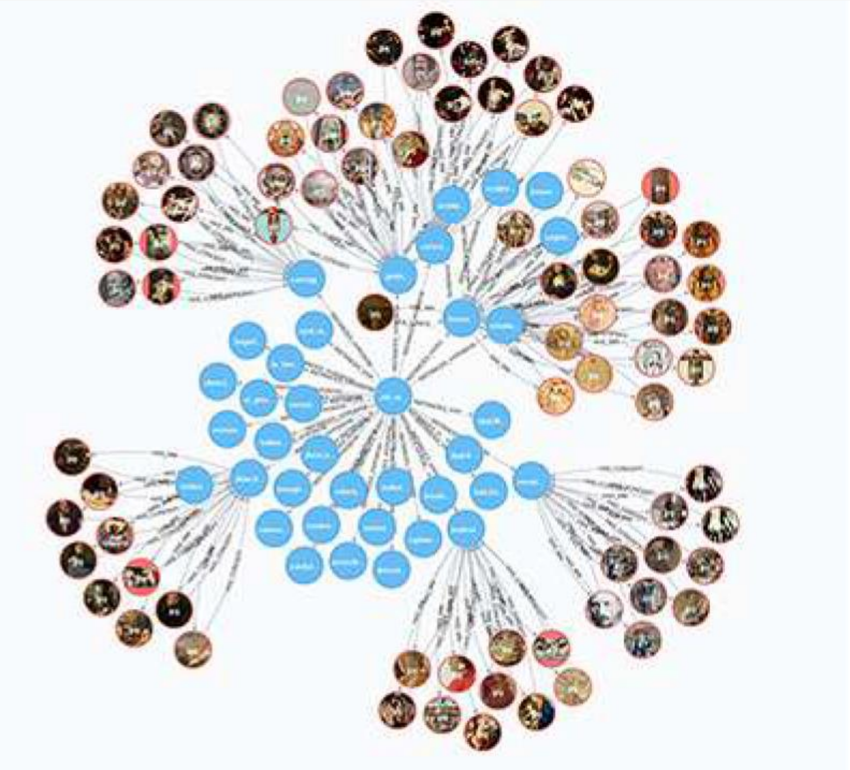
- Step 4 Using embedded Neo4J visualization tool, represent the structure of subgraphs resulting from query execution



Green: Multimedia
Blue: Concept

4. Case Study

- **Step 5** Using D3.js, make some possible improvement for the visualization of query results and interact with them



Conclusion

- **They propose a formal model combining top-level ontology models and graph models represented by a labelled, property-based structure to take into account both semantic, linguistic and multimedia aspects.**

- **The ontology has been developed using OWL in order to represent it in a machine-processable language.**

- **They have been able to generate an unambiguous machine-understandable formal machine-understandable formal representation of the semantics associated with multimedia description.**

Reference

- Srividya K Bansal and Sebastian Kagemann. 2015. Integrating big data: A semantic extract-transform-load framework. Computer 48, 3 (2015), 42–50.

- Antonio M. Rinaldi. 2018. A Semantic-based Model to represent Multimedia Big Data Computer 18, 9 (2018), 25–28.

Thank you!