

Linked Data Visualization TECHNIQUES, TOOLS, AND BIG DATA

Laura Po Nikos Bikakis Federico Desimoni George Papastefanatos

Synthesis Lectures on Data, Semantics, and Knowledge

Ying Ding and Paul Groth, Series Editors

Linked Data Visualization

Techniques, Tools, and Big Data

Synthesis Lectures on Data, Semantics, and Knowledge

Editors Ying Ding, University of Texas at Austin Paul Groth, University of Amsterdam

Founding Editor Emeritus James Hendler, *Rensselaer Polytechnic Institute*

Synthesis Lectures on Data, Semantics, and Knowledge is edited by Ying Ding of the University of Texas at Austin and Paul Groth of the University of Amsterdam. The series focuses on the pivotal role that data on the web and the emergent technologies that surround it play both in the evolution of the World Wide Web as well as applications in domains requiring data integration and semantic analysis. The large-scale availability of both structured and unstructured data on the Web has enabled radically new technologies to develop. It has impacted developments in a variety of areas including machine learning, deep learning, semantic search, and natural language processing. Knowledge and semantics are a critical foundation for the sharing, utilization, and organization of this data. The series aims both to provide pathways into the field of research and an understanding of the principles underlying these technologies for an audience of scientists, engineers, and practitioners.

Topics to be included:

- Knowledge graphs, both public and private
- Linked Data
- Knowledge graph and automated knowledge base construction
- Knowledge engineering for large-scale data
- Machine reading
- Uses of Semantic Web technologies
- Information and knowledge integration, data fusion
- Various forms of semantics on the web (e.g., ontologies, language models, and distributional semantics)
- Terminology, Thesaurus, & Ontology Management
- Query languages

Linked Data Visualization: Techniques, Tools, and Big Data

Laura Po, Nikos Bikakis, Federico Desimoni, and George Papastefanatos 2019

Ontology Engineering Elisa F. Kendall and Deborah L. McGuinness 2019

Demistifying OWL for the Enterprise Michael Uschold

2018

Validating RDF Data

Jose Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, and Dimitris Kontokostas 2017

Natural Language Processing for the Semantic Web

Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein 2016

The Epistemology of Intelligent Semantic Web Systems Mathieu d'Aquin and Enrico Motta 2016

Entity Resolution in the Web of Data Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis 2015

Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description

Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter 2015

Semantic Mining of Social Networks Jie Tang and Juanzi Li 2015

Social Semantic Web Mining Tope Omitola, Sebastián A. Ríos, and John G. Breslin 2015

Semantic Breakthrough in Drug Discovery Bin Chen, Huijun Wang, Ying Ding, and David Wild 2014

iv

Semantics in Mobile Sensing

Zhixian Yan and Dipanjan Chakraborty 2014

Provenance: An Introduction to PROV

Luc Moreau and Paul Groth 2013

Resource-Oriented Architecture Patterns for Webs of Data Brian Sletten

2013

Aaron Swartz's A Programmable Web: An Unfinished Work Aaron Swartz 2013

Incentive-Centric Semantic Web Application Engineering Elena Simperl, Roberta Cuel, and Martin Stein 2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web Eero Hyvönen 2012 v

VIVO: A Semantic Approach to Scholarly Networking and Discovery Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding 2012

Linked Data: Evolving the Web into a Global Data Space Tom Heath and Christian Bizer 2011 Copyright © 2020 by Morgan & Claypool

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Linked Data Visualization: Techniques, Tools, and Big Data Laura Po, Nikos Bikakis, Federico Desimoni, and George Papastefanatos

www.morganclaypool.com

ISBN: 9781681737256	paperback
ISBN: 9781681737263	ebook
ISBN: 9781681738345	epub
ISBN: 9781681737270	hardcover

DOI 10.2200/S00967ED1V01Y201911WBE019

A Publication in the Morgan & Claypool Publishers series SYNTHESIS LECTURES ON DATA, SEMANTICS, AND KNOWLEDGE

Lecture #19 Series Editors: Ying Ding, University of Texas at Austin Paul Groth, University of Amsterdam Founding Editor Emeritus: James Hendler, Rensselaer Polytechnic Institute Series ISSN Print 2160-4711 Electronic 2160-472X

Linked Data Visualization Techniques, Tools, and Big Data

Laura Po University of Modena and Reggio Emilia, Italy

Nikos Bikakis University of Ioannina, Greece

Federico Desimoni University of Modena and Reggio Emilia, Italy

George Papastefanatos ATHENA Research Center, Greece

SYNTHESIS LECTURES ON DATA, SEMANTICS, AND KNOWLEDGE #19



MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Linked Data (LD) is a well-established standard for publishing and managing structured information on the Web, gathering and bridging together knowledge from different scientific and commercial domains. The development of Linked Data Visualization techniques and tools has been adopted as the established practice for the analysis of this vast amount of information by data scientists, domain experts, business users, and citizens.

This book covers a wide spectrum of visualization topics, providing an overview of the recent advances in this area, focusing on techniques, tools, and use cases of visualization and visual analysis of LD. It presents core concepts related to data visualization and LD technologies, techniques employed for data visualization based on the characteristics of data, techniques for Big Data visualization, tools and use cases in the LD context, and, finally, a thorough assessment of the usability of these tools under different scenarios.

The purpose of this book is to offer a complete guide to the evolution of LD visualization for interested readers from any background and to empower them to get started with the visual analysis of such data. This book can serve as a course textbook or as a primer for all those interested in LD and data visualization.

KEYWORDS

linked data, data visualization, visual analytics, big data, visualization tools, web of data, semantic web, data exploration, information visualization, usability evaluation, human-computer interaction

Contents

	Preface				
	Ack	nowledgments xiii			
1	Intro	Introduction			
	1.1	The Power of Visualization on Linked Data 1			
	1.2	The Web of Linked, Open, and Semantic Data 4			
	1.3	Principles of Linked Data			
	1.4	The Linked Open Data Cloud			
	1.5	Web of Data in Numbers 11			
	1.6	1.6 The Value and Impact of Linked and Open Data 1			
	1.7	Semantic Web Technologies 13			
	1.8	Conclusions			
2	Prin	Principles of Data Visualization			
	2.1	Data Visualization Design Process			
	2.2	Data Visualization Types			
		2.2.1 Visualizing Patterns over Time			
		2.2.2 Visualizing Proportions			
		2.2.3 Visualizing Graph Relationships			
		2.2.4 Visualizing Data on Maps 34			
	2.3	Interactive Visualization			
	2.4	Visualization in Big Data Era 38			
		2.4.1 How Does the Visualization of Big Data Differ from Traditional			
		Ones?			
	25	2.4.2 Visualization Systems and Techniques			
	2.5	Conclusions			
3	Link	ced Data Visualization Tools			
3.1 Evolution Over Time		Evolution Over Time			
	Browsers and Exploratory Tools				

х		
	3.3	Tools Using Multiple Visualization Types
	3.4	Graph-Based Visualization Tools 56
	3.5	Domain, Vocabulary-Specific, and Device-Oriented Visualization Tools 64
	3.6	Ontology Visualization Tools
	3.7	Conclusions
4	Visu	alization Use Cases
	4.1	User Needs on LD Visual Exploration
	4.2	Use Cases
	4.3	Modeling Use Cases
		4.3.1 T-Box Related Use Cases
		4.3.2 A-Box Related Use Cases
		4.3.3 T-Box and A-Box Related Use Cases
	4.4	Conclusions
5	Emp	pirical Evaluation of Linked Data Visualization Tools
	5.1	Basic Characteristics of the Tools
	5.2	Evaluation
		5.2.1 Evaluation of T-Box Use Cases 100
		5.2.2 Evaluation of A-Box Use Cases 102
		5.2.3 Evaluation of A-Box and T-Box Uses Cases 102
		5.2.4 Evaluation Summary 106
	5.3	Different Tools for Different Tasks 106
	5.4	Conclusions
6	Con	clusions and Future Challenges
	6.1	Future Challenges 110
	Bibl	iography
	Autl	hors' Biographies

Preface

The Linked Data Principles defined by Tim Berners-Lee promise that a large portion of Web Data will be usable as one big interlinked RDF database. Today, we are assisting the staggering growth in both the production and consumption of Linked Data (LD) coming from diverse domains such as health and biology, humanities and social sciences, or open government. In the early phases of LD adoption, most efforts focused on the representation and publication of large volumes of privately held data in the form of Linked Open Data (LOD), contributing to the generation of the Linked Open Data Cloud.

Nowadays, given the wide adoption and availability of a very large number of LD sources, it is crucial to provide intuitive tools for researchers, data scientists, and domain experts as well as business users and citizens to visualize and interact with increasingly large datasets. Visual analytics integrates the analytic capabilities of the computer and the abilities of the human analyst, allowing novel discoveries and empowering individuals to take control of the analytical process. LD visualization aims to provide graphical representations of datasets or of some information of interest selected by a user, with the aim of facilitating their analysis and generating insights into complex interconnected information. Visualization techniques can vary according to the domain, the type of data, the task that the user is trying to perform, as well as the characteristics of the user (e.g., skills).

This book presents the principles of LD visualization, as well as demonstrates and evaluates state-of-the-art LD visualization tools. Moreover, future challenges and opportunities in the field of Big (Linked) Data visualization are presented.

The book is written for everyone who wants to explore and exploit LD, whether undergraduate and post-graduate students, data scientists, semantic technology developers, or UI & UX designers who wish to gain some practical experience with LD tools. Previous knowledge of Semantic Web technologies such as RDF, OWL, SPARQL, or programming skills is not required. The purpose of this book is to empower readers of any background to get started with their own experiments on the LOD Cloud, select the most appropriate LD tool for each scenario, and be aware of the challenges and techniques related to Big Linked Data exploration.

Since readers are likely to have a wide variety of different backgrounds, each chapter presents an overview of its content at the beginning. A reader who wishes to have a quick overview can start with the first page of each chapter. When the material in any section becomes more advanced, the reader can skip to the beginning of the next section without losing continuity. Chapter 1 introduces the Web of Linked Data, describing the phenomenon of the production and consumption of LD, the social and economic impact that this data has, and the effect that visualization tools can have in facilitating the understanding and exploitation of such

xii PREFACE

data. Moreover, it presents the principles of LD and the technologies of the Semantic Web Stack. Chapter 2 addresses how data can be presented in visual form, focusing on interactive and specialized visualizations of proportions, relationships, and spatial data. Further, it introduces the new challenges and methods related to Big Data Visualization. Chapter 3 surveys the variety of linked data visualization tools. Chapter 4 defines and models a set of visualization use cases based on the users' requirements in LD exploration. Chapter 5 describes a wide empirical evaluation of the tools introduced in Chapter 3. Here, a practical evaluation of the tools will be shown in order to describe their characteristics and limitations as well as formalize how the tools handle the use cases described in Chapter 4. Chapter 6 reports some conclusions and open issues and suggests research challenges and promising trends for the future.

Laura Po, Nikos Bikakis, Federico Desimoni, and George Papastefanatos March 2020

Acknowledgments

Sincere thanks go to Dr. Jakub Klímek for his careful reading of the manuscript and his constructive remarks. The work carried out in this book was partially supported by the *Networking on Linked Data* project funded by the "Enzo Ferrari" Engineering Department of the University of Modena and Reggio Emilia within FAR2019 as well as by the *VisualFacts* project (#1614) funded by the 1st Call of the Hellenic Foundation for Research and Innovation Research Projects for the support of post-doctoral researchers.

Laura Po, Nikos Bikakis, Federico Desimoni, and George Papastefanatos March 2020

CHAPTER 1

Introduction

Linked data provides the basis for knowledge to be distributed, networked, and shared. The term Linked Data (LD) refers to a set of best practices for publishing and interlinking structured data on the Web. Creating a connection between data and its contexts could lead to the development of intelligent search engines which could explore the Web, moving from a keyword-based approach to a meaning-based approach. Researches can be more accurate by exploiting the relations between words. LD can provide a benefit in several research areas like in the medical field for structuring the connections between various illnesses and the relative cures, in the scientific literature for structuring the citations between the million of documents published online. The potentialities of exploitation of LD are countless.

On the other hand, given the wide availability of LD sources, it is crucial to provide intuitive tools enabling users without semantic technology background to explore, analyze, and interact with increasingly large datasets. Visual analytics integrates the analytic capabilities of the computer and the abilities of the human analyst, allowing novel discoveries and empowering individuals to take control of the analytical process. LD visualization aims to provide graphical representations of datasets with the aim to facilitate their analysis and the generation of insights out of complex interconnected information.

In this chapter, we will introduce why visualization is a powerful means for linked data exploration, then, the principles and technologies that are the bases for the creation of LD are presented, and we also depict the incredible impact that LD can have in the real world.

In the next section, we start illustrating how visualization is a good way of interacting with the corresponding very large amounts of complex, interlinked, multi-dimensional data. The evolution of the web from Web 1.0 to Web 4.0. is depicted in Section 1.2. We highlight the principles of LD in Section 1.3; after this, we describe the Linked Data Cloud (Section 1.4) that draws datasets that have been published according to those principles. Sections 1.5 and 1.6 are devoted to assessing the impact of LD in our life and the opportunities they can generate. Finally, in Section 1.7, we introduce the theoretical basis of LD by describing the Semantic Web technologies.

1.1 THE POWER OF VISUALIZATION ON LINKED DATA

On any kind of data visualization enables serendipity and exploration. On LD it allows users to start understanding data previously unknown and to get the picture of the dataset in their mind, or to penetrate in some portions of the source. Moreover, visualization over LD is probably

the only way to enable users without technical skills to grasp the meaning of the content of LD sources. Furthermore, also domain experts can take advantage of a visual exploration of the dataset resulting in reduction in time.

Following the idea of Tim-Berners Lee, each resource should have a unique name that starts with HTTP. It means that reality can be replicated over the Internet. Each resource of the world can have its digital alter ego. Moreover, the pillar of linked data is that resources should be connected to other resources.

The simplest form of relationship are personal relations; John is a friend of Martin, Martin is the son of Peter, and somehow John is remotely connected to Martin. However, this can be extended to every existing field. Biology, Sociology, and Art are only a few areas in which LD can be deployed. LD has the power to universally express everything. However, how to visualize everything? One possible choice is to learn Semantic Web technologies, write SPARQL queries and then analyze the results. Despite the difficulty of writing SPARQL queries, this approach can be used only when the results are limited, since the information that can be displayed on a screen are limited. The other possible approach is to exploit the power of visualization.

The first tests on graphic visualization date back to 1890. In 1890, Herman Hollerith revolutionized the world of data analysis with a creative and innovative idea: he used punch cards to collect and analyze the U.S. census data. Using punch cards saved two years and five million dollars over the manual tabulation techniques used in the previous census while enabling a more thorough analysis of the data [Blodgett and Schultz, 1969]. We currently face an analogous development in the filed of LD. Since 2006, many researchers developed original solutions for solving the task of LD visualization and now we can exploit different tools and different visualization layouts.

```
Listing 1.1: Query for extracting relations between classes
```

```
SELECT distinct ?cl_sub ?prop ?cl_obj
WHERE {
    ?cl_sub ?prop ?cl_obj.
    ?cl_sub rdf:type ?class.
    ?cl_obj rdf:type ?class.
}
```

For example, how can a user understand the content of the Wikipathways dataset¹? Assuming that the user wants to know the contents of the dataset, he/she could formulate a SPARQL query to extract the classes and relations similar to the Listing 1.1 and then analyze the results, as shown in Figure 1.1.

Adopting a graphical visualization, instead, can simplify a lot the analysis of the results. For example, the previous information can be obtained through one of the visualizations provided by the tool H-BOLD (Figure 1.2). As it can be seen, displaying the same information with a graph, it is easier to understand the connections and paths among the classes.

¹https://www.wikipathways.org/index.php/Portal:Semantic_Web

1.1. THE POWER OF VISUALIZATION ON LINKED DATA 3

sub	prep	obj
http://vocabularies.wikipathways.org/wp#Catalysis	http://www.w3.org/2000/01/rdf-schemal/subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#TranscriptionTranslation	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#Inhibition	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#Binding	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#Stimulation	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#ComplexBinding	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#DirectedInteraction	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#Interaction
http://vocabularies.wikipathways.org/wp#GeneProduct	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DataNode
http://vocabularies.wikipathways.org/wp#Rna	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DataNode
http://vocabularies.wikipathways.org/wp#Metabolite	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DataNode
http://vocabularies.wikipathways.org/wp#Complex	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DataNode
http://vocabularies.wikipathways.org/wp#Protein	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DataNode
http://vocabularies.wikipathways.org/wp#Conversion	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#DirectedInteraction
http://vocabularies.wikipathways.org/wp#pathwayOntologyTag	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#ontologyTag
http://vocabularies.wikipathways.org/wp#diseaseOntologyTag	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#ontologyTag
http://vocabularies.wikipathways.org/wp#celltypeOntologyTag	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://yocabularies.wikipathways.org/wp#ontologyTag
http://vocabularies.wikipathways.org/wp#target	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#participants
http://vocabularies.wikipathways.org/wp#source	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://vocabularies.wikipathways.org/wp#participants

Figure 1.1: Results of the query in Listing 1.1.



Figure 1.2: HBOLD schema visualization of the Wikipathways dataset.

A crucial and impressive aspect of LD is that information is interlinked with different sources. Therefore, starting from the URI of a resource it is possible to display not only the information that describes the resource within the dataset, but also information from outside datasets. Figure 1.3 depicts how a LD visualization tool is able to create a collage of information from disparate sources. In that example, LodView² has been exploited to merge all information about London. Starting from the URI of the resource London in Dbpedia (http://dbpedia.

²https://lodview.it



Figure 1.3: LodView visualization of London.

org/resource/London), the tool looks at all the outcoming links and illustrates all the labels and pictures associated with the URIs of these links. What is displayed is an overview of images and information related to London from different data sources.

1.2 THE WEB OF LINKED, OPEN, AND SEMANTIC DATA

Tim Berners-Lee had a grand vision for the Internet when he began development of the World Wide Web in 1989 [Gillmor, 2004, Chapter 2]. He envisioned a read/write Web. However, what had emerged in the 1990s was an essentially read-only Web, the so-called Web 1.0. The users' interactions with the Web were limited to the search and the reading of information. The lack of active interaction between users and the Web lead, in 1999, to the birth of the Web 2.0. For the first time, common users were able to write and share information with everyone. This era empowered users with a few new concepts like blogs, social media, and video-streaming platforms like Twitter, Facebook, and Youtube.

Over time, users started to upload textual and multimedia content at an incredibly high rate and, as a consequence, more and more people started to use the Web for several different purposes. The high volume of web pages and the higher number of requests required Web applications to find new ways for handling documents. Machines needed to understand what data they are handling. The main idea was to provide a context to the documents in a machine-readable format. This new revolution, the Web 3.0, is called Semantic Web or Web of Data.

1.2. THE WEB OF LINKED, OPEN, AND SEMANTIC DATA 5

With the advent of the Semantic Web, users started to publish content together with metadata, i.e., other data that provide some context about the main data in a machineunderstandable way. The machine-readable descriptions enable content managers to add meaning to the content. In this way, a machine can process knowledge itself, instead of text, using processes similar to human deductive reasoning and inference, thereby obtaining more meaningful results and helping computers to perform automated information gathering and research. Making data understandable to machines implies, anyway, the sharing of a common data structure. To solve this issue, the RDF (Resource Description Framework) was the language proposed by the W3C for achieving a common data structure.

The Semantic Web also allows creating links among data on the Web. So that a person or machine can explore the Web of data. With Linked Data, when you have some of it, you can find other, related, data. Like the Web of hypertext, the Web of data is constructed with documents on the Web and the links between arbitrary things are described by RDF. The URIs identify any kind of object or concept.

Connecting your own data to other information already present on the Web resulted in at least two important consequences. The first is the possibility to add even more information and provide a more extended context, and the second is the creation of a global network of LD, the Giant Global Graph.

Alongside the arise of the Semantic Web, the Web shifted from a web pages-oriented Web to a data-oriented Web (Figure 1.4). Users of the Web started to publish data online and governments foresee in opening data, a way for enrolling the citizen in the governative life of the city.

The volume of data is growing exponentially everywhere. Each minute, 149,513 emails are sent, 3.3 million Facebook posts are created, 65,972 Instagram photos are uploaded, 448,800 Tweets are constructed, and 500 hours of YouTube videos are uploaded. The tremendous increase of data through the Internet of Things (continuous increase of connected devices, sensors, and smartphones), has contributed to the rise of a "data-driven" era. Moreover, future predictions argue that by 2020, every person will generate 1.7 megabytes in just a second.

Each sector is affected by this dizzying increase in available data and this means that Big Data analysis techniques must be implemented for mining data. Big Data is formed of large, diverse, complex, longitudinal, and distributed data sets generated from various instruments, sensors, Internet transactions, email, video, click streams, and other sources, whereas open-linked data focuses on the opening and the combining of data. The data can be released both by public organizations and by private organizations or individuals. Big Data analytics can be used to promote better utilization of resources and improved personalization. Naturally, there are no barriers between Big Data, Linked Data, and Open Data. It means that when a dataset is at the same time open, structured in node-edge fashion, and tremendously big, it can be referred as a BOLD (Big, Open, and Linked Data) source.



Figure 1.4: Transition from the Web of Documents to the Web of Data.

As a consequence, the arisen of the Web of Data gave birth to new specialized figures that can boost the value of those data. Data analysts, which are able to analyze and discover patterns from the data, Data Scientists, which try to predict the future based over past data, or the Chief Data Officer (CDO), who has the duty of defining and governing the data improvements strategy for supporting the achievement of corporate objectives, are only a few figures born for handling with the Web of Data.

Now, in 2019, we are already entering the fourth-generation internet, the Internet of Things, or the Web of intelligence connections. It is talked to be the web of the augmented reality for interacting at the same time with the real world and the online world. Domotic houses, smart domestic appliances, and voice assistants are only a few applications that will take place in the following years. Although interesting, the innovations of the Web 4.0 are out of the scope of this book and will not be addressed.

1.3 PRINCIPLES OF LINKED DATA

The term Linked Data was coined in 2006 from one of the creators of the Web, Sir Tim Berners-Lee. At the same time, he published a note³ listing four rules for publishing LD.

1.3. PRINCIPLES OF LINKED DATA

- 1. Use URIs as names for things. This is the first rule for publishing LD. This rule is the first milestone for creating a system where all resources could be univocally identified. The term resource refers both to real-world objects than web pages.
- 2. Use HTTP URIs so that people can look up those names. The second rule adopted the HTTP protocol as the mean for reaching resources and their information. Thanks to it, users are able to look for a specific object and get all the information they need as a result. Moreover, considering the fact that the resources should also be machine-readable, it is possible to exploit the content negotiation system for obtaining different representations of the requested resource.
- 3. When someone looks up a URI, provide useful information using the standard. This means the resource's information should be returned to the requester in an RDF compliant format.
- 4. Include links to other URIs so that they can discover more things. The last rule emphasizes the fact that resources should be connected to other resources in order to create what can be considered as the successor of WWW, the Giant Global Graph. This rule is the enabler of the great connectivity of Linked Data. Starting from a resource, the users of the Web have the possibility to jump from an object to another resource as they desire.

Some time later, more precisely at the TED⁴ conference in 2009, the same Tim Berners-Lee restated the principles defined in 2006 as three "extremely simple" rules.⁵

- 1. All kinds of conceptual things, they have names now that start with HTTP.
- 2. If I take one of those HTTP names and I look it up ... I will get back some data in a standard format which is kind of useful data that somebody might like to know about that thing, about the event, ...
- 3. When I get back that information it's not just got somebody's height and weight and when they were born, it's got relationships. And when it has relationships, whenever it expresses a relationship, then the other thing that it's related to is given one of those names that start with HTTP, so that I can go ahead and look that thing up.

Shortly before the birth of Linked Data principles, Open Data arose and there were defined some principles. The first appearance of the term "Open Data" dates back in 1995 in a document of American scientific agency. That document stated that geophysical and environmental data transcends political border so they promoted a complete and open exchange of scientific information between different countries. However, a formal definition of the term Open Data wait until 2005 with the Open Definition 2.1.⁶ This document holds several characteristics

⁴https://www.ted.com/

⁵https://www.ted.com/talks/tim_berners_lee_on_the_next_Web#t-960912

⁶http://opendefinition.org/od/2.1/en/

for data to be considered open and it can be summarized as: "Knowledge is open if anyone is free to access, use, modify, and share it—subject, at most, to measures that preserve provenance and opennes." Moreover, a more specific definition of the term *Open Government Data*⁷ had to wait for 2007 where 30 advocates gathered in Sebastopol, California. The meeting was meant to design a set of principles of open government data but the same logic could be inherited by all kinds of Open Data. At the end of the meeting, it was stated that government data is considered open if it is compliance with the following principles.

- *Complete*. All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
- *Primary.* Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
- *Timely*. Data is made available as quickly as necessary to preserve the value of the data.
- Accessible. Data is available to the widest range of users for the widest range of purposes.
- *Machine processable*. Data is reasonably structured to allow automated processing.
- *Non-discriminatory*. Data is available to anyone, with no requirement of registration.
- Non-proprietary. Data is available in a format over which no entity has exclusive control.
- *License-free*. Data is not subject to any copyright, patent, trademark, or trade secret regulation. Reasonable privacy, security, and privilege restriction may be allowed.

Well aware of the advantages both Linked Data and Open Data offered, it didn't take long before someone started encouraging to fuse Linked Data with Open Data. In fact, in 2010, the same Tim Berners-Lee published an extension of its note containing a star rating system for publishing Linked Open Data (LOD). Every rule of this rating system is a specialization of the previous one, it means that a five-star dataset satisfies all the criteria.

- ☆ Available on the Web (whatever format) but with an open license, to be Open Data. Documents are now publicly available online. Everyone can read, edit, save, share, and print them but unless building a custom parser, it is hard to extract data.
- Available as machine-readable structured data (e.g., Excel instead of image scan of a table, ...). Data are now accessible to machines but they remain bound to a proprietary file format. Extracting data means depending on proprietary software.
- Available through non-proprietary format (e.g., CSV instead of Excel, ...). Data are now fully accessible to everyone (both humans and machines) but they are still bound in documents and not freely accessible from the Web.

⁷https://public.resource.org/8_principles.html

1.4. THE LINKED OPEN DATA CLOUD 9

- ☆☆☆☆ Use open standards from W3C (RDF and SPARQL) to identify things so that people can point at your stuff. Every resource has its own URI that identifies it univocally. Users can look them up through HTTP requests and read, edit, and share those data freely. Generally, the data are represented through RDF format however they can be converted in other formats easily.
- AAAA Link your data to other people's data to provide context. Data are now fully connected to other resources and their value increases. Both publishers and consumers benefit from the *network effect*,⁸ the higher is the number of consumers than the higher is the value of the data.

1.4 THE LINKED OPEN DATA CLOUD

The LOD Cloud⁹ is a diagram that depicts the Linked Data datasets publicly available online. The diagram is updated regularly and it is maintained by the Insight Center for Data Analytics¹⁰ which is one of the biggest data science research center in Europe.

Everyone can upload datasets in the cloud but it will only be accepted and added to the cloud if it matches with the LOD Cloud principles, which are a slightly different version of the LD principles described in the section above. In order of being published, a dataset must respect the following rules.

- 1. There must be resolvable http:// (or https://) URIs.
- 2. They must resolve, with or without content negotiation, to RDF data in one of the popular RDF formats (RDFa, RDF/XML, Turtle, N-Triples).
- 3. The dataset must contain at least 1000 triples.
- 4. The dataset must be connected via RDF links to a dataset that is already in the diagram. This means, either your dataset must use URIs from the other dataset, or vice versa. They arbitrarily require at least 50 links.
- 5. Access the entire dataset must be possible via RDF crawling, via an RDF dump, or via a SPARQL endpoint.

Moreover, the maintainers of the LOD cloud developed an ad-hoc rating system for evaluating the quality of the published dataset. Although all the datasets respect the five rules described above, it is not assured that every dataset has the same characteristics. Generally, the evaluation metrics takes into account several metadata associated with the dataset like the presence or the absence of a SPARQL endpoint, the information about the author, the presence

```
<sup>8</sup>https://en.wikipedia.org/wiki/Network_effect
<sup>9</sup>https://lod-cloud.net/
<sup>10</sup>https://www.insight-center.org/
```

or the absence of the information of the author, the presence or the absence of metadata (and eventually the kind of metadata provided) and so forth. At the end of the process, each dataset is associated with a number of stars ranging from 1–5. The higher is the number of stars then higher is the quality of the dataset.

The Linked Data Cloud was, initially, created in May in 2007, at that time, it was composed of only 12 datasets. The LOD cloud contained the following.

- DBpedia which is a Linked Data version of Wikipedia.
- Geonames which contains a Linked Data version of geographical data.
- DBLP which contains a Linked Data version of academic data.
- Project Guttenberg and RDF Book Mashup which contains RDF data about books.
- Revyu which contains reviews in the form of LD.
- MusicBrainz, DBtune, and Jamendo which contain RDF data about the music business.
- FOAF (acronym of Friend of a Friend) which is an ontology containing LD that describes information about people, their relations, their activities, and, more generally, social network data.
- World Factbook and U.S. census data that contain governative data in the form of RDF triples.

The cloud shows which datasets are related to other datasets and a qualitative indication of the number of properties which connect a dataset to another. A thin line indicates that two datasets are connected with a low number of properties while a thick line represents a high number of relations connecting those datasets.

As time passed by, more and more institutions started to publish their data according to the Linked Data Principles described in Section 1.3 and the cloud got immense. After only half a year, the Linked Data Cloud doubled in terms of the number of the dataset published and reached the incredible number of 295 datasets on September 2011. There is no data about the dimension of the cloud in 2012 and 2013 but in 2014 the cloud counted up to 570 datasets. Again, no data is available for 2015 and 2016 but from 2017 the LOD cloud started to be updated regularly and there is plenty of information. The first record of 2017, dated January 26th, reported that the number of datasets increased to 1.146 (the double of the number of datasets present in the cloud in 2014). During the following years, the race of publishing Linked Data slowed down. In fact, recent updates of March 29, 2019 showed that the number of datasets present in the Linked Data Cloud is 1.239. Figure 1.5 represents the actual Linked Data Cloud. Despite the time elapsed and the increasing number of datasets, DBPedia is still the biggest and the best representative datasets of the LOD Cloud.

1.5. WEB OF DATA IN NUMBERS 11



Figure 1.5: Linked Open Data Cloud (March 29, 2019).

As it can be perceived from Figure 1.5 the cloud is depicted as a partially connected graph. Each node of the graph represents a dataset and the link between two nodes indicates that some kind of property connects elements from different datasets. For helping the users during the navigation of the LOD cloud, given the fact that each dataset is different to others both in size and in the domain it covers, the maintainers of the LOD cloud decided to improve the graph through the adoption of visual notation. The number of triples contained in a dataset is used for calculating the dimension of its node while the domain is used to color the inside of the nodes. Moreover, to provide further aid during the navigation, each domain is subsequently divided into distinct subsections.

1.5 WEB OF DATA IN NUMBERS

The Linked Open Data Cloud is probably the best representation of the Web of data. However, Figure 1.5 does not reflect the volume of data it contains. Each node of the graph, despite the small dimension, contains an incredibly amount of data. For example, DBPedia alone contains

more than 9.5 billion triples. Clearly, DBPedia is one of the biggest datasets around but it is not the only one that reaches an incredibly high number of triples. Geonames, LinkedGeoData, BabelNet are only a few other examples of huge datasets. Along with several metadata, the LOD cloud collects the number of triples which compose every single dataset. Unfortunately, the number of the triples is not present for all the datasets, but analyzing in details the dataset whose triples count is given, it results that the mean number of triples for each dataset is approximately equal to 176 million triples. It ends up in a total of 202 billion triples counted over 1151 datasets!

The number of Linked Data has risen in the last years also because the efforts of governments, to be more transparent and responsive to citizens' demands, have been increasing [Attard et al., 2015], and this, in most cases, resulted in the publication of (linked) Open Data. Many online data portals exist and play a fundamental role in the expansion of the Web of Data. Portals like DataHub,¹¹ the EU Open Data Portal,¹² the European Data Portal,¹³ Data.Gov,¹⁴ Asia-Pacific SDG Data Portal¹⁵ act as repositories for all kind of datasets (agricultural, economy, education, environment, government, justice, transports, ...) of different countries so that everyone could freely access those data. The only limitations to the usage of the data are defined by the licenses under which the data have been published but generally, they are not particularly restrictive. Thanks to those portals, the amount of data accessible through the Web is insane. Gathering together all the datasets those portal contains it is easy to exceed the threshold of one million datasets. However, despite the incredibly high number of datasets, their dimension is limited. Some datasets could be pretty big but the greatest part of them occupies a few kilobytes in space.

There is no clear information about the volume of data already present in the Web but it is easy to figure out that the number and the dimension of the datasets could only increase over the time reaching exabytes of information. Since that information hides a real treasure, in monetary terms, during the last period several data analysis tools and big data analytics tools have been developed for supporting the job of data scientists. One of the leading companies in this sector is the apache foundation that developed a high number of applications perfectly suited for handling big data like Apache Hadoop,¹⁶ Apache Spark,¹⁷ Apache Cassandra,¹⁸ Apache Commons RDF,¹⁹ Apache Jena,²⁰ and many others.

¹¹https://datahub.io/ ¹²https://data.europa.eu/euodp/en/home ¹³https://www.europeandataportal.eu/ ¹⁴https://www.data.gov/ ¹⁵http://data.unescap.org/sdg/ ¹⁶https://hadoop.apache.org/ ¹⁷https://spark.apache.org/ ¹⁸http://cassandra.apache.org/ ¹⁹https://commons.apache.org/ ²⁰https://jena.apache.org/

1.6. THE VALUE AND IMPACT OF LINKED AND OPEN DATA 13 **1.6 THE VALUE AND IMPACT OF LINKED AND OPEN** DATA

The impact of Open Data at the economic, political, and social levels has become clear in the recent years. The European Data Portal²¹ publishes every year several studies and reports about the situation of Open Data in Europe. They distinguished the benefits coming from Open Data in direct and indirect benefits. In their study [Carrara et al., 2015], they defined direct benefits as "monetised benefits that are realized in market transactions in the form of revenues and Gross Value Added (GVA), the number of jobs involved in producing a service or product, and cost savings" and indirect benefits as "new goods and services, time savings for users of applications using Open Data, knowledge economy growth, increased efficiency in public services and growth of related markets." In the same document they estimate that the direct value of the Open Data market in the European Union is 55.3 billion Euros, with a potential growth between 2016 and 2020 of 36.9% to a value of 75.7 billion Euros, and that the overall Open Data market reaches is estimated to be between 193 and 209 billion Euros, with an estimated projection of 265–286 billion Euros for 2020. They also quantified the economic benefits by looking other three indicators: number of jobs created, cost savings, and efficiency gains. The forecasted number of direct Open Data jobs is expected to rise from 75,000 of 2016 to nearly 100,000 jobs by 2020. Moreover, thanks to the positive economic effect on innovation and the development of numerous tools to increase efficiency, not only the private sector, but also the public sector is expected to experience an increased level of cost savings through Open Data to a total of 1.7 billion Euros by 2020. They also estimated an augmentation of 7.000 saved lives thanks to a quicker response, a decreasing of 5.5% in road fatalities, a decreasing of 16% in energy usage, etc.

Another important document that assesses the value of Open Data is Manyika et al. [2013]. That document, created in 2013, estimates the value of the world wide Open Data market is about 3 trillions dollar annually (1.1 trillion for the U.S. market, 0.7 trillion for the European market, and 1.7 for the others). The value is calculated over seven domains of interest (Education, Transportation, Consumer Products, Electricity, Oil and Gas, Health Care, Consumer finance). The staggering difference between the previous values implies that calculating the value of Open Data is not an easy task and that the value is highly dependent on the field in study. To the best of our knowledge, there is no actual estimation of the value of the U.S. Open Data market.

1.7 SEMANTIC WEB TECHNOLOGIES

In order to unlock the full potential of Linked Data and to understand how to extract the maximum profit from them, it is important to dive into the technologies that have favored the birth of Linked Data [Bikakis et al., 2013]. Semantic Web is built upon a series of different technologies that have been piled up. All of these technologies form the Semantic Web Stack. Figure 1.6



Figure 1.6: Semantic Web Stack.

represents the stack and highlights the logical structure (Concept and Abstraction) and the technologies adopted (Specification and Solutions) for the creation of the Semantic Web.

The first layer of the stack is clearly the media for the information transfer, the Web platform. The idea behind the Semantic Web was to create a globally distributed database. This means that it is necessary to univocally identify the resources and that is necessary to adopt an universally accepted encoding system in order to identify thing even between countries that adopt different writing systems. This first step was accomplished by the adoption of URI (Uniform Resource Identifies). With the advent of RDF1.1, in 2014, the actual naming convention standard became the IRI (International Resource Identifier). IRIs are sequences of Unicode characters and supports any character of any languages. This is a quite important progress in the multi-cultural context of the Internet.

Once defined how to identify and how to access the resources, it is mandatory to create them and provide additional information. The Resource Description Framework (RDF) is the model adopted for solving this task and it is a general purpose language for representing information about resources. RDF has a very simple and flexible data model, based on the central concept of the RDF statement. RDF statements describe simple facts as triples in the form of *Subject – Predicate – Object* consisting of the resource being described (the subject), a property (the predicate), and a property value (the object). In particular, the subject can either be an IRI or a Blank node, the predicate must be an IRI and the object can be an IRI, Blank node, or RDF Literal. A Blank node is a placeholder that stands for a resource to which no IRI nor literal is given. A collection of RDF statements (or else RDF triples) can be intuitively understood as

1.7. SEMANTIC WEB TECHNOLOGIES 15

a directed labeled graph, where the resources are nodes and the statements are arcs connecting two nodes (from the subject node to the object node). Finally, a set of RDF triples is called RDF Dataset or RDF Graph.

RDF data can be written down in a number of different formats, known as serialization. The first standard serialization format is called RDF/XML and it is based on XML tags system. Although the RDF/XML is still in use, other RDF serialization are now preferred because they are more human-friendly. The other serialization formats include:

- RDFa: notation for embedding RDF metadata in XHTML web pages;
- N-Triples: an intuitive and line-based format. It express each triple of an RDF graph on a different line;
- N3 (Notation 3): a serialization format developed by Tim Berners-Lee and designed to be compact and human-readable;
- Turtle (Terse RDF Triple Language): a compact and human-friendly format. It is a subset of N3;
- TriG: extension of Turtle notation;
- N-Quads: a superset of N-Triples, for serializing multiple RDF graphs. The fouth element of the "triple" contains the name of the graph to which the statement belongs; and
- JSON-LD: the standard JSON based serialization format that superseded RDF/JSON format. It can be used for writing RDF triples in a JSON style.

The third layer of the stack aims at structuring the data. The former RDF model and its extension, the RDFS (RDF Schema), were designed to describe, using a set of reserved terms called the RDFS vocabulary, resources and/or relationships between resources. They provide constructs for the description of types of objects (classes), type hierarchies (subclasses), properties that represent object features (properties), and property hierarchies (subproperty). In particular, a Class in RDFS corresponds to the generic concept of a type or category, somewhat like the notion of a class in object-oriented languages, and is defined using the construct *rdfs:Class*. The resources that belong to a class are called its instances. An instance of a class is a resource having an *rdf:type* property whose value is the specific class. Moreover, a resource may be an instance of more than one class. Classes can be organized in a hierarchical fashion using the construct *rdfs:subClassOf*. A property in RDFS is used to characterize a class or a set of classes and is defined using the construct *rdf:Property*. The Web Ontology Language (OWL) was released in 2004 and is the standard language for defining and instantiating Web ontologies. OWL and RDFS have several similarities. Indeed, OWL is defined as a vocabulary like RDF, however OWL has richer semantics. An OWL Class is defined using the construct owl: Class and represents a set of individuals with common properties. Moreover, OWL provides additional

constructors for class definition, including the basic set operations, union, intersection, and complement that are implemented, respectively, by the constructs *owl:unionOf*, *owl:intersectionOf*, and *owl:complementOf*. Regarding the individuals, OWL allows to specify two individuals to be identical or different through the *owl:sameAs* and *owl:differentFrom* constructs. Unlike RDF Schema, OWL distinguishes a property whose range is a datatype value (*owl:DatatypeProperty*) from a property whose range is a set of resources (*owl:ObjectProperty*). In 2009, an extended and revisioned version of OWL, called OWL 2, became the new W3C recommendation. The OWL 2 Web Ontology Language (OWL 2) has a very similar overall structure with OWL 1 and is backward compatible with it, while it introduces a plethora of new features.

Alongside the developement of OWL, a countless number of vocabularies have been developed. Just to name a few, VoID²² (Vocabulary of Interlinked Dataset) contains terms for providing metadata to a dataset, FoaF²³ (Friend of a Friend) operates in the Social Network domains and contains terms for describing people and their relations, SKOS²⁴ (Simple Knowledge Organization System) is used for sharing and linking knowledge organization systems like thesauri or taxonomies while the RDF Data Cube Vocabulary²⁵ can be used for publishing multi-dimensional data like statistics.

The SPARQL²⁶ Protocol and RDF Query Language (SPARQL) is a W3C recommendation and it is the standard query language for RDF data since 2008. SPARQL is one of the key technologies of the Semantic Web and it is used to retrieve and manipulate RDF data from the knowledge graphs available on the Web. The evaluation of SPARQL queries is based on graph pattern matching. Graph Patterns are templates that consist of a series of triples that the SPARQL engine looks for inside the store.

SPARQL allows four query forms: SELECT, ASK, CONSTRUCT, and DESCRIBE. The SELECT query form returns a solution sequence, i.e., a sequence of variables and their bindings. The ASK query form returns a Boolean value (yes or no), indicating whether a query pattern matches or not. The CONSTRUCT query form returns an RDF graph structured according to the graph template of the query. Finally, the DESCRIBE query form returns an RDF graph which provides a "description" of the matching resources. Thus, based on the query forms, the SPARQL query results may be RDF Graphs, SPARQL solution sequences, and Boolean values.

Unfortunately, this SPARQL version presented different vacancies including the lack of the support to data management operators so, in 2013, the W3C SPARQL working group published SPARQL 1.1²⁷ which extended the original SPARQL query language in several as-

²²https://www.w3.org/TR/void/

²³http://xmlns.com/foaf/spec/

²⁴https://www.w3.org/2004/02/skos/

²⁵https://www.w3.org/TR/vocab-data-cube/

²⁶ https://www.w3.org/TR/rdf-sparql-query/ 27

²⁷https://www.w3.org/TR/sparql11-query/

1.8. CONCLUSIONS 17

pects. Precisely, SPARQL 1.1 introduced features for manipulating the content of the store and introduced the support for nested queries and aggregation functions.

At last, triples need to be stored in a triplestore. Different proposals have been developed over the year. Monolithic Triple Storages are triplestore that store all the triples in a single table. They are sure easy to implement and work for huge number of properties but it requires an intelligent index system and several self join during queries. A slightly lighter version of monolithic storage implies to associate each URI and Literal with a numerical identifier. It ends up in two tables; one holds the association URI/Literal—number and the other contains the triples in a numerical fashion. Property Tables are triplestore which create a table for each class. This way, the tuple with the same characteristics are grouped together. It resembles the structure of relational DB and queries require fewer joins but it potentially contains a high number of NULL values. Vertically Partitioned Table triplestore create a two-column table for every property of the dataset. Each table contains the subject, in the first column, and the object, in the second column, of the triples with that specific predicate. This system grants good performance when then number of properties is low, otherwise, it is particularly expansive in computational terms. Hexastores are particular structures that create an index for each possible combination of triple elements in order to enable efficient processing at the cost of six times the disk required for storing data.

Quadstores are the natural evolution of the triplestores. The main difference between them is that the quadstores store tuples of four elements: *Subject, Predicate, Object,* and *Graph*.

Furthermore, the data contained in these structures (both triplestore and quadstores) tend to be very atomic since the nodes in the graph are primitive data types like strings, integers, date, etc and the relations connect those kind of data. Graph Databases model the graph following an object oriented fashion. The nodes are not simple primitive kinds of data but instances of the graph. Generally, each instance has properties that describe itself (datatype properties) and properties that relates it to other objects (object properties), so the datatype properties are integrated together forming a sort of description for the instance while object properties are treated as the arcs that connect different instances. Therefore, in graph databases, the nodes are not simple strings but pure objects with a moltitude of datatype properties. Some popular graph databases are Neo4j²⁸ and Amazon Neptune.²⁹

1.8 CONCLUSIONS

In this chapter, we have introduced the story of the Web, starting from the description of a Web of interlinked documents till the Web of Data. The potential hidden behind the usage of the meaning of the words can boost the advent of intelligent agents so we explored the fundamentals that gave birth to the Semantic Web ranging from RDF to Linked Data to the SPARQL query language to the storage technologies. Moreover, we have reported some statistics collected by

²⁸https://neo4j.com/ ²⁹https://aws.amazon.com/it/neptune/

different Open Data agencies worldwide about the dimension, value, and impact that Open and Linked Data have been estimated to reach in the global economic market.

CHAPTER 2

Principles of Data Visualization

Information visualization aims at visually representing different types of data (e.g., geographic, numerical, text, network) in order to enable and reinforce cognition. Information visualization offers intuitive ways for information perception and manipulation that essentially amplify the overall cognitive performance of information processing, especially for non-expert users. Visual analytics combines information visualization with data exploration capabilities. It enables users to explore and analyze unknown (in terms of semantics and structure) sets of information, discover hidden correlations and causalities, and make sense of data in ways that are not always possible with traditional quantitative data analysis and mining techniques. This is of great importance, especially given the massive volumes of digital information concerning nearly every aspect of human activity that are currently being produced and collected. The so-called Big Data era refers to this tremendous volume of information collected by digital means and analyzed to produce new knowledge in a plethora of scientific domains.

The LOD cloud is one of the main pillars of the so-called Big Data era. The number of datasets published on the Web, the amount of information and the interconnections established between disparate sources being available in the form of LOD are nowadays ever expanding, making traditional ways of analyzing them insufficient and posing new challenges in the way humans can explore, visualize, and gain insights out of them.

In this chapter, we present the basic principles and tasks for data visualization. We first provide the tasks for the preparation and visualization of data. We then present the most popular ways for graphically representing data according to the type of data and then we provide an overview of the main techniques for users to interact with the data. Finally, we show the main techniques used for visualizing and interacting with Big Data.

2.1 DATA VISUALIZATION DESIGN PROCESS

Information visualization requires a set of preprocessing tasks, such that data can be first extracted from data sources, transformed, enriched with additional metadata and properly modeled before they can be visually analyzed. An abstract view of this process along with its constituent tasks is shown in Figure 2.1. It presents a generic user-driven end-to-end roadmap of the tasks, problems, and challenges relating to the visualization of data in general.

20 2. PRINCIPLES OF DATA VISUALIZATION



Figure 2.1: Process for the visualization of data.