

Aidan Hogan · Eva Blomqvist · Michael Cochez ·
Claudia d'Amato · Gerard de Melo · Claudio Gutierrez
· Sabrina Kirrane · José Emilio Labra Gayo · Roberto
Navigli · Sebastian Neumaier · Axel-Cyrille Ngonga
Ngomo · Axel Polleres · Sabbir M. Rashid · Anisa Rula
· Lukas Schmelzeisen · Juan Sequeda · Steffen Staab ·
Antoine Zimmermann

Knowledge Graphs

Knowledge Graphs

Synthesis Lectures on Data, Semantics, and Knowledge

Editor

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

Knowledge Graphs

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann
2021

Web Data APIs for Knowledge Graphs: Easing Access to Semantic Data for Application Developers

Alberto Meroño-Peñuela, Pasquale Lisena, and Carlos Martínez-Ortiz
2021

Designing and Building Enterprise Knowledge Graphs

Juan Sequeda and Ora Lassila
2021

Linked Data Visualization: Techniques, Tools, and Big Data

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

Ontology Engineering

Elisa F. Kendall and Deborah L. McGuinness
2019

Demystifying OWL for the Enterprise

Michael Uschold
2018

Validating RDF Data

José 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. Mixer
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

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

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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2022

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.

Knowledge Graphs

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann

ISBN: 978-3-031-00790-3 paperback

ISBN: 978-3-031-01918-0 PDF

ISBN: 978-3-031-00113-0 hardcover

DOI 10.1007/978-3-031-01918-0

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA, SEMANTICS, AND KNOWLEDGE

Lecture #22

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 2691-2023 Electronic 2691-2031

Knowledge Graphs

Aidan Hogan

DCC, Universidad de Chile; IMFD

Michael Cochez

Vrije Universiteit Amsterdam and Discovery Lab,
Elsevier

Gerard de Melo

HPI, University of Potsdam and Rutgers University

Sabrina Kirrane

WU Vienna

Roberto Navigli

Sapienza University of Rome

Axel-Cyrille Ngonga Ngomo

DICE, Universität Paderborn

Sabbir M. Rashid

Tetherless World Constellation, Rensselaer
Polytechnic Institute

Lukas Schmelzeisen

Universität Stuttgart

Steffen Staab

Universität Stuttgart and University of Southampton

Eva Blomqvist

Linköping University

Claudia d'Amato

University of Bari

Claudio Gutierrez

DCC, Universidad de Chile; IMFD

José Emilio Labra Gayo

Universidad de Oviedo

Sebastian Neumaier

St. Pölten University of Applied Sciences

Axel Polleres

WU Vienna

Anisa Rula

University of Brescia

Juan Sequeda

data.world

Antoine Zimmermann

École des mines de Saint-Étienne

ABSTRACT

This book provides a comprehensive and accessible introduction to knowledge graphs, which have recently garnered notable attention from both industry and academia. Knowledge graphs are founded on the principle of applying a graph-based abstraction to data, and are now broadly deployed in scenarios that require integrating and extracting value from multiple, diverse sources of data at large scale.

The book defines knowledge graphs and provides a high-level overview of how they are used. It presents and contrasts popular graph models that are commonly used to represent data as graphs, and the languages by which they can be queried before describing how the resulting data graph can be enhanced with notions of schema, identity, and context. The book discusses how ontologies and rules can be used to encode knowledge as well as how inductive techniques—based on statistics, graph analytics, machine learning, etc.—can be used to encode and extract knowledge. It covers techniques for the creation, enrichment, assessment, and refinement of knowledge graphs and surveys recent open and enterprise knowledge graphs and the industries or applications within which they have been most widely adopted. The book closes by discussing the current limitations and future directions along which knowledge graphs are likely to evolve.

This book is aimed at students, researchers, and practitioners who wish to learn more about knowledge graphs and how they facilitate extracting value from diverse data at large scale. To make the book accessible for newcomers, running examples and graphical notation are used throughout. Formal definitions and extensive references are also provided for those who opt to delve more deeply into specific topics.

KEYWORDS

knowledge graphs, graph databases, knowledge graph embeddings, graph neural networks, ontologies, knowledge graph refinement, knowledge graph quality, knowledge bases, artificial intelligence, semantic web, machine learning

Contents

	Preface	xv
	Acknowledgments	xix
1	Introduction	1
2	Data Graphs	5
2.1	Models	5
2.1.1	Directed Edge-Labeled Graphs	6
2.1.2	Heterogeneous Graphs	8
2.1.3	Property Graphs	9
2.1.4	Graph Dataset	11
2.1.5	Other Graph Data Models	12
2.1.6	Graph Stores	13
2.2	Querying	13
2.2.1	Basic Graph Patterns	13
2.2.2	Complex Graph Patterns	16
2.2.3	Navigational Graph Patterns	19
2.2.4	Other Features	22
2.2.5	Query Interfaces	22
3	Schema, Identity, and Context	25
3.1	Schema	25
3.1.1	Semantic Schema	25
3.1.2	Validating Schema	27
3.1.3	Emergent Schema	32
3.2	Identity	35
3.2.1	Persistent Identifiers	35
3.2.2	External Identity Links	37
3.2.3	Datatypes	38
3.2.4	Lexicalization	38
3.2.5	Existential Nodes	39

3.3	Context	40
3.3.1	Direct Representation	40
3.3.2	Reification	41
3.3.3	Higher-Arity Representation	41
3.3.4	Annotations	42
3.3.5	Other Contextual Frameworks	44
4	Deductive Knowledge	47
4.1	Ontologies	48
4.1.1	Interpretations and Models	49
4.1.2	Ontology Features	51
4.1.3	Entailment	56
4.1.4	If-Then vs. If-and-Only-If Semantics	56
4.2	Reasoning	57
4.2.1	Rules	57
4.2.2	Description Logics	60
5	Inductive Knowledge	67
5.1	Graph Analytics	67
5.1.1	Techniques	69
5.1.2	Frameworks	70
5.1.3	Analytics on Data Graphs	74
5.1.4	Analytics with Queries	75
5.1.5	Analytics with Entailment	76
5.2	Knowledge Graph Embeddings	76
5.2.1	Tensor-Based Models	78
5.2.2	Language Models	88
5.2.3	Entailment-Aware Models	90
5.3	Graph Neural Networks	91
5.3.1	Recursive Graph Neural Networks	91
5.3.2	Non-Recursive Graph Neural Networks	95
5.4	Symbolic Learning	96
5.4.1	Rule Mining	97
5.4.2	Axiom Mining	100
5.4.3	Hypothesis Mining	102

6	Creation and Enrichment	105
6.1	Human Collaboration	105
6.2	Text Sources	106
6.2.1	Pre-Processing	106
6.2.2	Named Entity Recognition (NER)	107
6.2.3	Entity Linking (EL)	107
6.2.4	Relation Extraction (RE)	108
6.2.5	Joint Tasks	109
6.3	Markup Sources	109
6.3.1	Wrapper-Based Extraction	110
6.3.2	Web Table Extraction	111
6.3.3	Deep Web Crawling	111
6.4	Structured Sources	111
6.4.1	Mapping from Tables	112
6.4.2	Mapping from Trees	114
6.4.3	Mapping from Other Knowledge Graphs	114
6.5	Schema/Ontology Creation	115
6.5.1	Ontology Engineering	115
6.5.2	Ontology Learning	116
7	Quality Assessment	119
7.1	Accuracy	119
7.1.1	Syntactic Accuracy	119
7.1.2	Semantic Accuracy	120
7.1.3	Timeliness	120
7.2	Coverage	121
7.2.1	Completeness	121
7.2.2	Representativeness	121
7.3	Coherency	122
7.3.1	Consistency	122
7.3.2	Validity	123
7.4	Succinctness	123
7.4.1	Conciseness	123
7.4.2	Representational Conciseness	124
7.4.3	Understandability	124
7.5	Other Quality Dimensions	125

8	Refinement	127
8.1	Completion	127
8.1.1	General Link Prediction	127
8.1.2	Type-Link Prediction	128
8.1.3	Identity-Link Prediction	128
8.2	Correction	129
8.2.1	Fact Validation	129
8.2.2	Inconsistency Repairs	131
8.3	Other Refinement Tasks	131
9	Publication	133
9.1	Best Practices	133
9.1.1	FAIR Principles	133
9.1.2	Linked Data Principles	135
9.2	Access Protocols	137
9.2.1	Dumps	137
9.2.2	Node Lookups	138
9.2.3	Edge Patterns	138
9.2.4	(Complex) Graph Patterns	139
9.2.5	Other Protocols	140
9.3	Usage Control	140
9.3.1	Licensing	140
9.3.2	Usage Policies	141
9.3.3	Encryption	142
9.3.4	Anonymization	143
10	Knowledge Graphs in Practice	145
10.1	Open Knowledge Graphs	145
10.1.1	DBpedia	145
10.1.2	Yet Another Great Ontology	146
10.1.3	Freebase	147
10.1.4	Wikidata	147
10.1.5	Other Open Cross-Domain Knowledge Graphs	148
10.1.6	Domain-Specific Open Knowledge Graphs	148
10.2	Enterprise Knowledge Graphs	149
10.2.1	Web Search	149
10.2.2	Commerce	150

10.2.3	Social Networks	150
10.2.4	Finance	151
10.2.5	Other Industries	151
11	Conclusions	153
A	Background	157
A.1	Historical Perspective	157
A.2	“Knowledge Graphs:” Pre-2012	158
A.3	“Knowledge Graphs:” 2012 Onward	161
	Bibliography	165
	Authors’ Biographies	229

Preface

The origins of this book can be traced back to a Dagstuhl Seminar, held in 2018, on the topic of Knowledge Graphs. At the time of the seminar, the topic was quickly becoming mainstream in academia and industry, but there were conflicting messages as to what a “knowledge graph” was. Much of the discussion of the seminar centered on this question, and there were divergent opinions as to how knowledge graphs could (or should) be defined; how they relate to previous concepts such as graph databases, knowledge bases, ontologies, RDF graphs, property graphs, semantic networks, etc.; and how the emerging area of Knowledge Graphs should be positioned with respect to the established areas of Artificial Intelligence, Big Data, Databases, Graph Theory, Logic, Machine Learning, Knowledge Representation, Natural Language Processing, Networks (in their various forms), and the Semantic Web. As the discussion continued, a consensus began to emerge: Knowledge Graphs, as a topic, involves a novel confluence of techniques stemming from previously disparate scientific communities, with the unifying goal of developing novel graph-based techniques for better integrating and extracting value from diverse knowledge sources at large scale.

As a follow-up to the seminar, the attendees agreed that in order to foster this unifying view of Knowledge Graphs, there was a need for a manuscript that would serve as a general introduction to the area. This manuscript would:

- motivate knowledge graphs and the value of abstracting data as graphs;
- survey the historical context of knowledge graphs and the key initiatives leading to their popularization;
- draw together disparate views of knowledge graphs into a unifying definition;
- provide an introduction to the key techniques that knowledge graphs enable, relating to querying, validation, reasoning, learning, refinement, enrichment, quality assessment, and more besides;
- describe how knowledge graphs are used in practice, surveying the companies using knowledge graphs, the applications they are used for, the open knowledge graphs that have been published, etc.; and
- delineate future research directions for knowledge graphs.

The manuscript would then serve as an introductory text for students, practitioners and researchers new to the area, helping to form a consensus in terms of what is a knowledge graph, laying the foundations for future developments.

The goal of preparing this manuscript was an ambitious one, and involved drawing together and distilling down a vast amount of literature on a diverse range of topics into a set of key concepts described in an accessible way. For this reason, the manuscript has been prepared by many authors, who have lent their knowledge and expertise to the preparation of specific sections. A short version of the manuscript was first published as a tutorial paper [Hogan et al., 2021], consisting of an abridged version of the first five chapters of this book, along with a summary of how knowledge graphs are used in practice, and conclusions. However, there was not enough space to describe all of the important developments in the area. This led us to publish this book, which further includes topics relating to the creation, enrichment, quality assessment, refinement and publication of knowledge graphs, as well as formal definitions, a historical perspective, and extended discussion throughout.

The book is divided into ten chapters. Chapter 1 provides a general introduction to the area, defines the concept of a “knowledge graph”, and provides a high-level overview of how knowledge graphs are currently being used. Chapter 2 presents and contrasts popular graph models that are commonly used to represent data as graphs, and the languages by which they can be queried. Chapter 3 describes how the resulting data graph can be enhanced with notions of schema, identity, and context. Chapter 4 discusses how ontologies and rules can be used to encode knowledge, and how they enable deductive forms of reasoning. Chapter 5 delves into how inductive techniques—based on statistics, graph analytics, machine learning, etc.—can be used to encode and extract knowledge. Chapter 6 is dedicated to techniques for the creation and enrichment of knowledge graphs from legacy sources of data. Chapter 7 enumerates a variety of quality measures that can be used to assess a knowledge graph in terms of its fitness for use in a variety of applications. Chapter 8 presents key methods for the refinement of knowledge graphs, with the goal of improving their completeness and correctness. Chapter 9 provides a survey of the open and enterprise knowledge graphs that have emerged in recent years, along with the industries within which, and the applications for which, they have been most widely adopted. Chapter 10 wraps up the book with discussion of the current limitations and future directions along which knowledge graphs are likely to evolve. An Appendix further covers knowledge graphs from an historical perspective, establishing their significance in the broader context of the academic study of data and knowledge, as well as surveying prior definitions of “knowledge graphs” from the literature.

A key aim of this book is to be accessible to a broader audience. While background knowledge of related topics such as Databases, Logic, Machine Learning, Semantic Web, etc., will help to understand some of the particular topics mentioned, such a background is not necessary to follow the general concepts described within. The book aims to motivate and illustrate the various concepts it introduces from a practical perspective, and in order to be as accessible as possible, relies heavily on an example-driven presentation using a graphical notation. For the reader wishing to dig more into the technical minutiae, we complement this discussion with formal definitions throughout; however, the reader more interested in understanding the gen-

eral concepts and their rationale will find the discussion to be self-contained if they choose to skip the definitions presented in visually distinctive boxes.

The book serves as an entry point for those new to the topic, and may thus serve as a useful textbook for university courses, for researchers who are venturing into the topic for the first time, and for practitioners who wish to understand more about how knowledge graphs might be of use within their company or organization, or indeed, how to maximize the value of the knowledge graphs that they are currently developing. Readers who are already active within specific sub-areas of Knowledge Graphs may further appreciate the technical definitions included, the references to other literature provided, and the broader perspective that this book offers in terms of the other related sub-areas and how they complement each other.

By drawing together diverse techniques from disparate areas, Knowledge Graphs has become an exciting topic in terms of both research and applications. We expect to see growing interest on this topic as the years advance, and indeed hope that this book will help to more firmly establish the foundations of this topic, and to foster future developments upon these foundations, potentially by its readers.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann
September 2021

Acknowledgments

We thank the organizers and attendees of the Dagstuhl Seminar on “Knowledge Graphs.” We also thank those who provided feedback on this content.

Hogan was funded by Fondecyt Grant No. 1181896. Hogan & Gutierrez were funded by ANID Millennium Science Initiative Program, Code ICN17_002. Cochez did part of the work while employed at Fraunhofer FIT, Germany and was later partially funded by Elsevier’s Discovery Lab. Kirrane, Ngonga Ngomo, Polleres & Staab received funding through the project “Know-Graphs” from the European Union’s Horizon program under the Marie Skłodowska-Curie grant agreement No. 860801. Kirrane & Polleres were supported by the European Union’s Horizon 2020 research and innovation programme under grant 731601. Labra was supported by the Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R). Navigli was supported by the MOUSSE ERC Grant No. 726487 under the European Union’s Horizon 2020 research and innovation programme. Rashid was supported by IBM Research AI through the AI Horizons Network. Schmelzeisen was supported by the German Research Foundation (DFG) grant STA 572/18-1.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann
September 2021

CHAPTER 1

Introduction

Though the phrase “knowledge graph” has been used in the literature since at least 1972 [Schneider, 1973], the modern incarnation of the phrase stems from the 2012 announcement of the Google Knowledge Graph [Singhal, 2012], followed by further announcements of the development of knowledge graphs by Airbnb [Chang, 2018], Amazon [Krishnan, 2018], eBay [Pittman et al., 2017], Facebook [Noy et al., 2019], IBM [Devarajan, 2017], LinkedIn [He et al., 2016], Microsoft [Shrivastava, 2017], Uber [Hamad et al., 2018], and more besides. The growing industrial uptake of the concept proved difficult for academia to ignore: more and more scientific literature is being published on knowledge graphs, which includes books (e.g., Fensel et al. [2020], Kejriwal et al. [2021], Pan et al. [2017], Qi et al. [2020]), as well as papers outlining definitions (e.g., Ehrlinger and Wöß [2016]), novel techniques (e.g., Lin et al. [2015], Pujara et al. [2013], Wang et al. [2014]), and surveys of specific aspects of knowledge graphs (e.g., Paulheim [2017], Wang et al. [2017]).

Underlying all such developments is the core idea of using graphs to represent data, often enhanced with some way to explicitly represent knowledge [Noy et al., 2019]. The result is most often used in application scenarios that involve integrating, managing, and extracting value from diverse sources of data at large scale [Noy et al., 2019]. Employing a graph-based abstraction of knowledge has numerous benefits in such settings when compared with, for example, a relational model or NoSQL alternatives. Graphs provide a concise and intuitive abstraction for a variety of domains, where edges capture the (potentially cyclical) relations between the entities inherent in social data, biological interactions, bibliographical citations and co-authorships, transport networks, and so forth [Angles and Gutiérrez, 2008]. Graphs allow maintainers to postpone the definition of a schema, allowing the data—and its scope—to evolve in a more flexible manner than typically possible in a relational setting, particularly for capturing incomplete knowledge [Abiteboul, 1997]. Unlike (other) NoSQL models, specialized graph query languages support not only standard relational operators (joins, unions, projections, etc.), but also navigational operators for recursively finding entities connected through arbitrary-length paths [Angles et al., 2017]. Standard knowledge representation formalisms—such as ontologies [Brickley and Guha, 2014, Hitzler et al., 2012, Mungall et al., 2012] and rules [Horrocks et al., 2004, Kifer and Boley, 2013]—can be employed to define and reason about the semantics of the terms used to label and describe the nodes and edges in the graph. Scalable frameworks for graph analytics [Malewicz et al., 2010, Stutz et al., 2016, Xin et al., 2013a] can be leveraged for computing centrality, clustering, summarization, etc., in order to gain insights about the do-

2 1. INTRODUCTION

main being described. Various representations have also been developed that support applying machine learning techniques both directly and indirectly over graphs [Wang et al., 2017, Wu et al., 2019].

In summary, the decision to build and use a knowledge graph opens up a range of techniques that can be brought to bear for integrating and extracting value from diverse sources of data at large scale. The goal of this book is to motivate and give a comprehensive introduction to knowledge graphs: to describe their foundational data models and how they can be queried; to discuss representations relating to schema, identity, and context; to discuss deductive and inductive ways to make knowledge explicit; to present a variety of techniques that can be used for the creation and enrichment of graph-structured data; to describe how the quality of knowledge graphs can be discerned and how they can be refined; to discuss standards and best practices by which knowledge graphs can be published; and to provide an overview of existing knowledge graphs found in practice. Our intended audience includes researchers and practitioners who are new to knowledge graphs. As such, we do not assume that readers have specific expertise on knowledge graphs.

Knowledge graph The definition of a “*knowledge graph*” remains contentious [Bergman, 2019, Bonatti et al., 2018, Ehrlinger and Wöß, 2016], where a number of (sometimes conflicting) definitions have emerged, varying from specific technical proposals to more inclusive general proposals; we address these prior definitions in Appendix A. Herein we adopt an inclusive definition, where we view a knowledge graph as *a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities*. The graph of data (aka *data graph*) conforms to a graph-based data model, which may be a *directed edge-labeled graph*, a *property graph*, etc. (we discuss concrete alternatives in Chapter 2). By *knowledge*, we refer to something that is *known*. Such knowledge may be accumulated from external sources, or extracted from the knowledge graph itself. Knowledge may be composed of simple statements, such as “*Santiago is the capital of Chile*,” or quantified statements, such as “*all capitals are cities*.” Simple statements can be accumulated as edges in the data graph. If the knowledge graph intends to accumulate quantified statements, a more expressive way to represent knowledge—such as *ontologies* or *rules*—is required. *Deductive methods* can then be used to entail and accumulate further knowledge (e.g., “*Santiago is a city*”). Additional knowledge—based on simple or quantified statements—can also be extracted from and accumulated by the knowledge graph using *inductive methods*.

Knowledge graphs are often assembled from numerous sources, and as a result, can be highly diverse in terms of structure and granularity. To address this diversity, representations of *schema*, *identity*, and *context* often play a key role, where a *schema* defines a high-level structure for the knowledge graph, *identity* denotes which nodes in the graph (or in external sources) refer to the same real-world entity, while *context* may indicate a specific setting in which some unit of knowledge is held true. As aforementioned, effective methods for *extraction*, *enrichment*, *quality assessment*, and *refinement* are required for a knowledge graph to grow and improve over time.

In practice Knowledge graphs aim to serve as an ever-evolving shared substrate of knowledge within an organization or community [Noy et al., 2019]. We distinguish two types of knowledge graphs in practice: *open knowledge graphs* and *enterprise knowledge graphs*. Open knowledge graphs are published online, making their content accessible for the public good. The most prominent examples—DBpedia [Lehmann et al., 2015], Freebase [Bollacker et al., 2007b], Wikidata [Vrandečić and Krötzsch, 2014], YAGO [Hoffart et al., 2011], etc.—cover many domains and are either extracted from Wikipedia [Hoffart et al., 2011, Lehmann et al., 2015], or built by communities of volunteers [Bollacker et al., 2007b, Vrandečić and Krötzsch, 2014]. Open knowledge graphs have also been published within specific domains, such as media [Raimond et al., 2014], government [Hendler et al., 2012, Shadbolt and O’Hara, 2013], geography [Stadler et al., 2012], tourism [Alonso Maturana et al., 2018, Kärle et al., 2018, Lu et al., 2016, Zhang et al., 2019], life sciences [Callahan et al., 2013], and more besides. Enterprise knowledge graphs are typically internal to a company and applied for commercial use-cases [Noy et al., 2019]. Prominent industries using enterprise knowledge graphs include Web search (e.g., Bing [Shrivastava, 2017], Google [Singhal, 2012]), commerce (e.g., Airbnb [Chang, 2018], Amazon [Dong, 2019, Krishnan, 2018], eBay [Pittman et al., 2017], Uber [Hamad et al., 2018]), social networks (e.g., Facebook [Noy et al., 2019], LinkedIn [He et al., 2016]), finance (e.g., Accenture [Okorafor and Ray, 2019], Banca d’Italia [Bellomarini et al., 2019], Bloomberg [Meij, 2019], Capital One [Branum and Sehon, 2019], Wells Fargo [Newman, 2019]), among others. Applications include search [Shrivastava, 2017, Singhal, 2012], recommendations [Chang, 2018, Hamad et al., 2018, He et al., 2016, Noy et al., 2019], personal agents [Pittman et al., 2017], advertising [He et al., 2016], business analytics [He et al., 2016], risk assessment [Dalglish, 2016, Tobin, 2017], automation [Henson et al., 2019], and more besides. We will provide more details on the use of knowledge graphs in practice in Chapter 10.

Running example To keep the discussion accessible, throughout the book we present concrete examples in the context of a hypothetical knowledge graph relating to tourism in Chile (loosely inspired by related use-cases [Kärle et al., 2018, Lu et al., 2016]). The knowledge graph is managed by a tourism board that aims to increase tourism in the country and promote new attractions in strategic areas. The knowledge graph itself will eventually describe tourist attractions, cultural events, services, businesses, travel routes, etc. Some applications the organization envisages are to:

- create a tourism portal that allows visitors to search for attractions, upcoming events, and other related services (in multiple languages);
- gain insights into tourism demographics in terms of season, nationalities, etc.;
- analyze sentiment about tourist attractions, including positive reviews, summaries of complaints about events and services, crime reports, etc.;

4 1. INTRODUCTION

- understand tourism trajectories: the sequence of attractions, events, etc., that tourists often visit;
- cross-reference these tourism trajectories with currently available flights, buses, etc., to suggest new strategic routes for public transport;
- offer personalized recommendations of places to visit;
- and so forth.

Outline The remainder of the book is structured as follows.

Chapter 2 outlines graph data models and the languages used to query them.

Chapter 3 describes representations of schema, identity, and context for graphs.

Chapter 4 presents deductive formalisms for representing and entailing knowledge.

Chapter 5 describes inductive techniques for learning from graphs.

Chapter 6 discusses the creation and enrichment of knowledge graphs.

Chapter 7 enumerates dimensions for assessing knowledge graph quality.

Chapter 8 discusses various techniques for knowledge graph refinement.

Chapter 9 introduces principles and protocols for publishing knowledge graphs.

Chapter 10 surveys some prominent knowledge graphs and their applications.

Chapter 11 concludes with future directions for knowledge graphs.

Appendix A outlines the historical background for knowledge graphs.

CHAPTER 2

Data Graphs

At the foundation of any knowledge graph is the principle of first applying a graph abstraction to data, resulting in an initial data graph. We now discuss a selection of graph-structured data models that are commonly used in practice to represent data graphs. We then discuss the primitives that form the basis of graph query languages used to interrogate such data graphs.

2.1 MODELS

Leaving aside graphs, let us assume that the tourism board from our running example has not yet decided how to model relevant data about attractions, events, services, etc. The board first considers using a tabular structure—in particular, relational databases—to represent the required data, and though they do not know precisely what data they will need to capture, they begin to design an initial relational schema. They begin with an Event table with five columns:

Event(name, venue, type, start, end)

where name and start together form the primary key of the table in order to uniquely identify recurring events. But as they start to populate the data, they encounter various issues: events may have multiple names (e.g., in different languages), events may have multiple venues, they may not yet know the start and end date-times for future events, events may have multiple types, and so forth. Incrementally addressing these modeling issues as the data become more diverse, they generate internal identifiers for events and adapt their relational schema until they have:

EventName(id, name), EventStart(id, start), EventEnd(id, end),
EventVenue(id, venue), EventType(id, type)

(2.1)

With the above schema, the organization can now model events with 0– n names, venues, and types, and 0–1 start dates and end dates (without needing nulls).






Along the way, the board has to incrementally change the schema several times in order to support new sources of data. Each such change requires a costly remodeling, reloading, and reindexing of data; here we only considered one table. The tourism board struggles with the relational model because they do not know, *a priori*, what data will need to be modeled or what sources they will use. But once they reach the latter relational schema, the board finds that they can integrate further sources without more changes: with minimal assumptions on *multiplicities* (1–1, 1– n , etc.) this schema offers a lot of flexibility for integrating incomplete and diverse data.

6 2. DATA GRAPHS

In fact, the refined, flexible schema that the board ends up with—as shown in (2.1)—is modeling a set of binary relations between entities, which indeed can be viewed as modeling a graph. By instead adopting a graph data model from the outset, the board could forgo the need for an upfront schema, and could define any (binary) relation between any pair of entities at any time.




We now introduce graph data models popular in practice [Angles et al., 2017].

2.1.1 DIRECTED EDGE-LABELED GRAPHS

A directed edge-labeled graph (sometimes known as a *multi-relational graph* [Balazevic et al., 2019a, Bordes et al., 2013, Nickel and Tresp, 2013]) is defined as a set of nodes—like , , , —and a set of directed labeled edges between those nodes, like . In the case of knowledge graphs, nodes are used to represent entities and edges are used to represent (binary) relations between those entities. Figure 2.1 provides an example of how the tourism board could model event data as a directed edge-labeled graph. The graph includes data about the names, types, start and end date-times, and venues for events.¹ Adding information to such a graph typically involves adding new nodes and edges (with some exceptions discussed later). Representing incomplete information requires simply omitting a particular edge; for example, the graph does not yet define a start/end date-time for the Food Truck festival.

Modeling data as a graph in this way offers more flexibility for integrating new sources of data, compared to the standard relational model, where a schema must be defined upfront and followed at each step. While other structured data models such as trees (XML, JSON, etc.) would offer similar flexibility, graphs do not require organizing the data hierarchically (should venue be a parent, child, or sibling of type for example?). They also allow cycles to be represented and queried (e.g., note the directed cycle in the routes between Santiago, Arica, and Viña del Mar).

A standardized data model based on directed edge-labeled graphs is the Resource Description Framework (RDF) [Cyganiak et al., 2014], which has been recommended by the W3C. The RDF model defines different types of nodes, including *Internationalized Resource Identifiers* (IRIs) [Dürst and Suignard, 2005] which allow for global identification of entities on the Web; *literals*, which allow for representing strings (with or without language tags) and other datatype values (integers, dates, etc.); and *blank nodes*, which are anonymous nodes that are not assigned an identifier (for example, rather than create internal identifiers like EID15, EID16, in RDF, we have the option to use blank nodes). We will discuss these different types of nodes further in Section 3.2 when we speak about issues relating to identity.

¹We draw bidirectional edges as , which more concisely depicts two directed edges:  and . Also while some naming conventions recommend more complete edge labels that include a verb, such as *has venue* or *is valid from*, in this book, for presentation purposes, we will omit the “has” and “is” verbs from such labels, using simply *venue* or *valid from*.

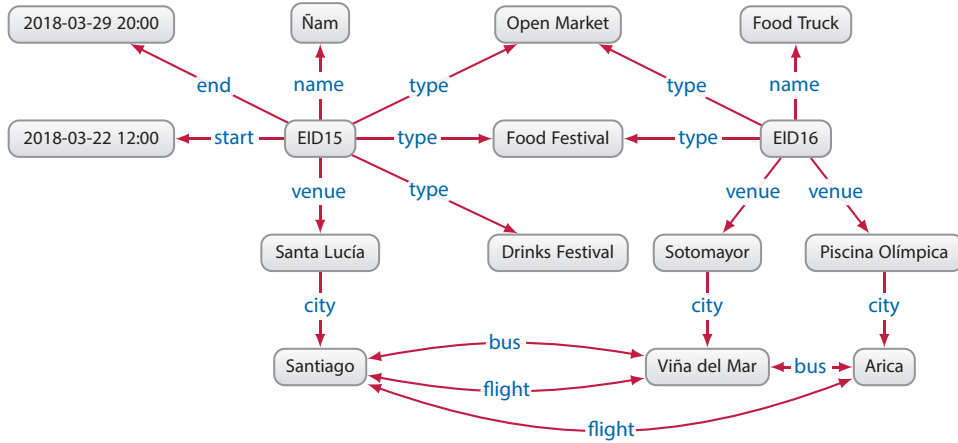


Figure 2.1: Directed edge-labeled graph describing events and their venues.

We now formally define a directed edge-labeled graph, where we denote by **Con** a countably infinite set of constants.

Definition 2.1 Directed edge-labeled graph. A *directed edge-labeled graph* is a tuple $G = (V, E, L)$, where $V \subseteq \mathbf{Con}$ is a set of nodes, $L \subseteq \mathbf{Con}$ is a set of edge labels, and $E \subseteq V \times L \times V$ is a set of edges.

Example 2.2 In reference to Figure 2.1, the set of nodes V has 15 elements, including Arica, EID16, etc. The set of edges E has 23 triples, including $(\text{Arica}, \text{flight}, \text{Santiago})$. Bidirectional edges are represented with two edges. The set of edge labels L has 8 elements, including *start*, *flight*, etc.

Definition 2.1 does not state that V and L are disjoint: though not present in the example, a node can also serve as an edge-label. The definition also permits that nodes and edge labels can be present without any associated edge. Either restriction could be explicitly stated—if necessary—in a particular application while still conforming to a directed edge-labeled graph.

For ease of presentation, we may treat a set of (directed labeled) edges $E \subseteq V \times L \times V$ as a directed edge-labeled graph (V, E, L) , in which case we refer to the graph induced by E assuming that V and L contain all and only those nodes and edge labels, respectively, used in E . We may similarly apply set operators on directed edge-labeled graphs, which should be interpreted as applying to their sets of edges; for example, given $G_1 = (V_1, E_1, L_1)$ and