Jeff Z. Pan · Guido Vetere
Jose Manuel Gomez-Perez
Honghan Wu *Editors*

Exploiting Linked Data and Knowledge Graphs in Large Organizations

Springer

Exploiting Linked Data and Knowledge Graphs in Large Organizations

Jeff Z. Pan · Guido Vetere
Jose Manuel Gomez-Perez
Honghan Wu
Editors

Exploiting Linked Data and Knowledge Graphs in Large Organizations

Springer

Editors

Jeff Z. Pan
University of Aberdeen
Aberdeen
UK

Jose Manuel Gomez-Perez
iSOCO Lab
Madrid
Spain

Guido Vetere
IBM Italia
Rome
Italy

Honghan Wu
University of Aberdeen
Aberdeen
UK

ISBN 978-3-319-45652-2 ISBN 978-3-319-45654-6 (eBook)
DOI 10.1007/978-3-319-45654-6

Library of Congress Control Number: 2016949103

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

When I began my research career as a graduate student at Rensselaer Polytechnic Institute in 1989, the phrase “knowledge graph” was not in use. The use of graphs, however, as a notation for “knowledge representation” (KR) was quite common. CLASSIC, the first real implemented description logic, was just being introduced from Bell Labs, and although it had a linear syntax, the community was still in the habit of drawing graphs that depicted the knowledge that was being represented.

This habit traced its history at least as far as M. Ross Quilian’s work on *Semantic Networks*, and subsequent researchers imagined knowledge to be intrinsic in the design of Artificial Intelligence (AI) systems, universally sketching the role of knowledge in a graphical form. By the late 1980s the community had more or less taken up the call for formalisation proposed by Bill Woods and later his student, Ron Brachman; graph formalisms were perhaps the central focus of AI at the time, and stayed that way for another decade.

Despite this attention and focus, by the time I moved from academia to industrial research at IBM’s Watson Research Centre in 2002, the knowledge representation community had never really solved any problems other than our own. Knowledge representation and reasoning evolved, or perhaps devolved, into a form of mathematics, in which researchers posed difficult-to-solve puzzles that arose more from syntactic properties of various formalisms than consideration of anyone else’s actual use cases. Even though we tended to use the words, “semantic” and “knowledge”, there was nothing particularly semantic about any of it, and indeed the co-opting by the KR community of terms like semantics, ontology, epistemology, etc. to refer to our largely algorithmic work, reliably confused the hell out of people who actually knew what those terms meant.

In my 12-year career at IBM, I found myself shifting with the times as a revolution was happening in AI. Many researchers roundly rejected the assumptions of the KR field, finding the focus on computation rather than data to be problematic. A new generation of data scientists who wanted to instrument and measure everything began to take over. I spent a lot of my time at IBM trying to convince others that the KR technology was useful, and even helping them use it. It was a losing battle, and like the field in general I began to become enamoured of the influential power of empirical evidence—it made me feel like a scientist. Still, however, my allegiance to the KR vision, that knowledge was intrinsic to the design of AI systems, could not be completely dispelled.

In 2007, a group of 12 researchers at IBM began working on a top secret moonshot project which we code-named "BlueJ"—building a natural language question answering system capable of the speed and accuracy necessary to achieve expert human-level performance on the TV quiz show, *Jeopardy!* It was the most compelling and interesting project I have ever worked on, and it gave me an opportunity to prove that knowledge—human created and curated knowledge—is a valuable tool. At the start of the project, Dave Ferrucci, the team leader, challenged us all to "make bets" on what we thought would work and commit to being measured on how well our bets impacted the ability to find the right answer as well as to *understand if the answer is correct*. I bet on KR, and for the first year, working alone on this particular bet, I failed, much as the KR community had failed more broadly to have any impact on any real problems other people had. But in the following year, Ferrucci agreed to put a few more people on it (partly because of my persuasive arguments, but mostly because he believed in the KR vision, too) and with the diversity of ideas and perspectives that naturally comes from having more people, we started to show impact. After our widely publicised and viewed victory over the two greatest *Jeopardy!* players in history, my team published the results of our experiments that demonstrated more than 10 % of Watson's winning performance (again, in terms of both finding answers and determining if they were correct) came from represented knowledge.

Knowledge is not the destination

In order to make this contribution to IBM's Watson, my team and I had to abandon our traditional notion of KR and adopt a new one, that I later came to call, "Knowledge is not the destination". The abject failure of KR to have any measurable impact on anything up to that point in time was due, I claim, to a subtle shift in that research community, sometime in the 1980s, from knowledge representation and reasoning as an integral part of some larger system, to KR&R as the ultimate engine of AI. This is where we were when I came into the field, and this was tacit in how I approached AI when I was working in Digital Libraries, Web Systems, and my early efforts at IBM in natural language question answering.

The most ambitious KR&R activity before that time was Cyc, which prided itself on being able to conclude, "If you leave a snowman outside in the sun it will melt". But Cyc could never possibly answer any of the myriad possible questions that might get asked about snowmen melting, because it would need a person to find the relevant Cyc micro-theory, look up the actual names and labels used in the axioms, type them in the correct and rather peculiar syntax, debug the reasoner and find the right set of heuristics that would make it give an answer, and even with all that it still probably could not answer a question like, "If your snowman starts to do *this*, turn on the air conditioner", Watson might actually have had a shot at answering something like this, but only because it knew from large language corpora that ‘snowman starts to melt’ is a common n-gram, not because it understands thermodynamics.

Working with people from Cycorp, or with anyone in the KR&R world, we became so enamoured of our elegant logic that, without a doubt, the knowledge became our focus. We—and I can say this with total confidence—we absolutely believed that getting the right answer was a trivial matter as long as you had the knowledge and reasoning right. The knowledge was the point.

“Knowledge is not the destination” refers to the epiphany that I had while working on Watson. The knowledge was important, but it wasn't the point—the point was to get answers right and to have confidence in them. If knowledge could not help with this, then it really was useless. But what kind of knowledge would help? Axioms about all the most general possible things in the world? Näive physics? Expert Physics? Deep Aristotelean theories? No.

What mattered for Watson was having millions of simple “propositional” facts available at very high speed. Recognising entities by their names, knowing some basic type of information, knowing about very simple geospatial relationships like capitals and borders, where famous people were born and when, and much much more. Knowing all this was useful not because we looked up answers this way—*Jeopardy!* never asked about a person’s age—but because these little facts could be stitched together with many other pieces of evidence from other sources to understand how confident we were in each answer.

This knowledge, a giant collection of subject-property-object triples, can be viewed as a graph. A very simple one, especially by KR&R standards, but this knowledge graph was not itself the goal of the project. The goal—the destination—of the project was winning *Jeopardy!* So, in fact, we made absolutely no effort to improve the knowledge we used from DBpedia and freebase. We needed to understand how well it worked for our problem in the general case, because there was no way to know what actual questions would be asked in the ultimate test in front of 50 million people.

Knowledge Graphs are Everywhere!

As of the publication of this book, most major IT companies—more accurately, most major information companies—including Bloomberg, NY Times, Microsoft, Facebook, Twitter and many more, have significant knowledge graphs like Watson did, and have invested in their curation. Not because any of these graphs is their business, but because using this knowledge helps them in their business.

After Watson I moved to Google Research, where freebase lives on in our own humongous knowledge graph. And while Google invests a lot in its curation and maintenance, Google’s purpose is not to build the greatest and most comprehensive knowledge graph on Earth, but to make a search, email, youtube, personal assistants and all the rest of our Web-scale services, better. That’s our destination.

Many believe that the success of this kind of simplistic, propositional, knowledge graph proves that the original KR&R vision was a misguided mistake, but an outspoken few have gone so far as to claim it was a 40+ year waste of some great minds. As much as I appreciate being described as a great mind, I prefer a different explanation: the work in KR for the past 40 years was not a waste of time, it was just the wrong place to start. It was solving a problem no one yet had, because no one had yet built systems that used this much explicit and declared knowledge.

Now, *knowledge graphs are everywhere*. Now industry is investing in the knowledge that drives their core systems. The editors of this volume, Jeff Pan, Guido Vetere, José Manuel Gómez Pérez and Honghan Wu, all themselves experts in this old yet burgeoning area of research, have gone to great lengths to put together research that matters today, in this world of large-scale graphs representing knowledge that makes a difference in the systems we use on the Web, on our phones, at work and at home.

The editorial team members have unique backgrounds, yet have worked together before, such as in the EU Marie Curie *K-Drive* project, and this book is a natural extension of their recent work on studying the properties of knowledge graphs. Jeff started at Manchester and has done a widely published work in formal reasoning systems, and moved to Aberdeen where his portfolio broadened considerably to include Machine Learning, large data analysis, and others, although he never strayed too far from practical reasoning, such as *approximate reasoning*, and querying for knowledge graphs. Guido has run several successful schema management projects on large data systems at IBM, and was part of the team that worked to bring Watson to Italy. Jose has done important research in the area of distributed systems, semantic data management and NLP, making knowledge easier to understand, access and consume by real users, and Honghan has been doing research in the area of medical knowledge systems.

After you finish this book, try to find a faded red copy of *Readings in Knowledge Representation* lest we forget and reinvent the Semantic Network.

May 2016

Dr. Christopher Welty
Google Research NYC

Preface

A few years after Google announced that their 'Knowledge Graph' would have allowed searching for *things, not strings*,[1] knowledge graphs start entering information retrieval, databases, Semantic Web, artificial intelligence, social media and enterprise information systems. But what exactly is Knowledge Graph? Where did it come from? What are the major differences between knowledge graphs for enterprise information management and those for Web search? What are the key components in a knowledge graph architecture? How can knowledge graphs help in enterprise information management? How can you build good quality knowledge graphs and utilise them to achieve your goals?

The main purpose of this book is to provide answers to these questions in a systematic way. Specifically, this book is for academic researchers, knowledge engineers and IT professionals who are interested in acquiring industrial experiences in using knowledge graphs for enterprises and large organisations. The book provides readers with an updated view on methods and technologies related to knowledge graphs, including illustrative corporate use cases.

In the last four years, we have been working hard and closely in the K-Drive—Knowledge Driven Data Exploitation—project (286348), which was funded by EU FP7/Marie Curie Industry-Academia Partnerships and Pathways schema/PEOPLE Work Programme. The main purpose of this project was to apply and extend advanced knowledge techniques to solve real-world problems, such as those in corporate knowledge management, healthcare and cultural heritage. Most of the challenges we encountered and techniques we dug into are highly related knowledge graph techniques. This book is a natural outcome of the K-Drive project that reflects and concludes the understanding we accumulated from the past four years of work, the lessons we have learned and the experiences we gained.

Contentwise, we will focus on the key technologies for constructing, understanding and consuming knowledge graphs, which constitute the three parts of this book, respectively. **Part I** introduces some background knowledge and technologies,

[1] Introducing the Knowledge Graph: things, not strings, googleblog.blogspot.com May 16, 2012

and then presents a simple architecture in order to help you to understand the main phases and tasks required during the lifecycle of knowledge graphs. **Part II** is the main technical part that starts with the state-of-the-art Knowledge Graph construction approaches, then focuses on exploration and exploitation techniques and finishes with advanced topics of Question Answering over/using knowledge graphs. Finally, **Part III** demonstrates successful stories of knowledge graph applications in Media Industry, Healthcare and Cultural Heritage; and ends with conclusions and future visions.

It is true that there is no *gold standard* definition of Knowledge Graph (KG). While working on the book, the editors and chapter contributors have debated lively on *what constitutes KG?*, *how is it related to relevant techniques like Semantic Web and Linked Data techniques?* and *what are its key features?* Fortunately, most, if not all, arguments have been settled and the conclusions and agreements have been put into the book, e.g. into the last two sections of Chap. 2. Even luckier, when finalising the book, editors have got the opportunity to collect opinions on *visions, barriers and next steps of Knowledge Graph* from key figures in the community including outstanding researchers, practitioners in leading organisations and start-ups, and representative users of various domains. Such valuable opinions have also been compiled into this book as part of its conclusion and future vision.

We would like to thank all of the chapter contributors as well as all members of the K-Drive project, who have given so much of their time and efforts for this book, in particular Dr. Yuting Zhao, who offered much helpful advice on the organisation of the book.

We had great pleasure in having Chris Welty write a touching Foreword for this book, sharing with us his rich experience and epiphany he had during the compelling BlueJ project, as well as his opinions on the motivation (‘*Knowledge Graphs are Everywhere!*’) and the importance of this book.

We would also like to acknowledge the IBM DeepQA research team for allowing us to use their architecture diagram marked as Fig. 7.1 in the book.

We are grateful to the following experts in the field for sharing with us their visions, barriers and next steps of Knowledge Graph in our concluding chapter: Sören Auer, Riccardo Bellazzi, Oscar Corcho, Richard Dobson, Junlan Feng, Aldo Gangemi, Alfio M. Gliozzo, Tom Heath, Juanzi Li, Peter Mika, Fabrizio Renzi, Marco Varone, Denny Vrandečić and Haofen Wang.

Aberdeen, UK — Jeff Z. Pan
Rome, Italy — Guido Vetere
Madrid, Spain — Jose Manuel Gomez-Perez
Aberdeen, UK — Honghan Wu
June 2016

Contents

Part I Knowledge Graph Foundations & Architecture

Part II Constructing, Understanding and Consuming Knowledge Graphs

Part III Industrial Applications and Successful Stories

Contributors

Panos Alexopoulos Expert System, Madrid, Spain

Ronald Denaux Expert System, Madrid, Spain

Alessandro Faraotti IBM Italia, Rome, Italy

Nuria Garcia-Santa Expert System, Madrid, Spain

Jose Manuel Gomez-Perez Expert System, Madrid, Spain

Marco Monti IBM Italia, Milan, Italy

Alessandro Moschitti University of Trento, Trento, Italy

Hai Nguyen University of Aberdeen, King's College, Aberdeen, UK

Massimo Nicosia University of Trento, Trento, Italy

Jeff Z. Pan University of Aberdeen, King's College, Aberdeen, UK

Fernanda Perego IBM Italia, Milan, Italy

Yuan Ren University of Aberdeen, King's College, Aberdeen, UK

Mariano Rodriguez-Muro IBM USA, Thomas J. Watson Research Center, Yorktown Heights, NY, USA

Kavitha Srinivas IBM USA, Thomas J. Watson Research Center, Yorktown Heights, NY, USA

Kateryna Tymoshenko Trento RISE, Povo di Trento, Trento, Italy

Guido Vetere IBM Italia, Rome, Italy

Boris Villazon-Terrazas Expert System, Madrid, Spain

Andrew Walker University of Aberdeen, King's College, Aberdeen, UK

Gemma Webster University of Aberdeen, King's College, Aberdeen, UK

Honghan Wu King’s College London, London, UK

Yuting Zhao IBM Italia, Milan, Italy

Man Zhu Southeast University, Nanjing, China

Chapter 1
Enterprise Knowledge Graph: An Introduction

Jose Manuel Gomez-Perez, Jeff Z. Pan, Guido Vetere and Honghan Wu

```
A knowledge graph consists of a set of interconnected typed
     entities and their attributes.
```

Compared to other knowledge-oriented information systems, the distinctive features of knowledge graphs lie in their special combination of knowledge representation structures, information management processes and search algorithms. The term ‘Knowledge Graph’ became well known in 2012 when Google started to use knowledge graph in their search engine, allowing users to search for things, people or places, rather than just matching strings in the search queries with those in Web documents. Inspired by the success story of Google, knowledge graphs are gaining momentum in the world’s leading information companies.

The idea of a knowledge graph is not completely new though. The original idea dates back to the knowledge representation technique called the Semantic Network. Later on, researchers in Knowledge Representation and Reasoning (KR) addressed

J.M. Gomez-Perez (✉)
Expert System, Prof. Waksman 10, 28036 Madrid, Spain
e-mail: jmgomez@expertsystem.com

J.Z. Pan
University of Aberdeen, King’s College, Aberdeen AB24 3UE, UK
e-mail: jeff.z.pan@abdn.ac.uk

G. Vetere
IBM Italia, via Sciangai 53, 00144 Rome, Italy
e-mail: gvetere@it.ibm.com

H. Wu
King’s College London, De Crespigny Park, London SE5 8AF, UK
e-mail: honghan.wu@kcl.ac.uk

J.Z. Pan et al. (eds.), *Exploiting Linked Data and Knowledge Graphs in Large Organizations*, DOI 10.1007/978-3-319-45654-6_1

some well-known issues on the Semantic Network when standardising the modern version of Semantic Network, or RDF (Resource Description Frameworks). It turns out that knowledge representation techniques, such as Knowledge Graph or Semantic Network, are useful not only for Web search, but also in many other systems and applications, including enterprise information management. The focus of the book, therefore, is about constructing, understanding and exploiting knowledge graphs in large organisations.

The basic unit of a knowledge graph is (the representation of) a singular *entity*, such as a football match you are watching, a city you will visit soon or anything you would like to describe. Each entity might have various attributes. For example, the attributes of a person include name, birthdate, nationality, etc. Furthermore, entities are connected to each other by *relations*; e.g. you *follow* one of your colleagues in Twitter. *Relations* can be used to bridge two separate knowledge graphs. For example, by saying that your Twitter ID and the ID on your driving license are denoting one and the same person, this actually interlinks Twitter data with the information space in the driver licensing agency of your country. Not surprisingly, each entity needs an identification to distinguish one another. This is the final jigsaw in the knowledge representation of knowledge graphs. Note that to facilitate the interlinking between various knowledge graphs, the entity IDs need to be *globally* unique. Types of entities and relations are defined in some machine-understandable dictionaries called ontologies. The standard ontology language is called OWL (Web Ontology Language).

The quality of a knowledge graph is crucial for its applications. For example, a knowledge graph should be consistent. In the above example, it could be the case that your contact address in your driving license is different than that in your Twitter profile. To create a knowledge graph connecting these two information spaces, such inconsistency should be resolved by keeping the correct one. In addition to consistency, one also needs to consider correctness, and coverage of knowledge graphs, as well as efficiency, fault tolerance and scalability of services based on knowledge graphs. Many of those aspects are related to, among others, the schema (ontology) of a knowledge graph.

> A knowledge graph has an ontology as its schema defining the vocabulary used in the knowledge graph.

1.1 A Brief History of Knowledge Graph

1.1.1 The Arrival of Semantic Networks

Knowledge management in early human history was largely shaped by oral communication before the invention of languages, which then allowed human knowledge to be recorded and passed on through generations. One of the first computer-based knowledge representation approaches are *Semantic Networks*, which represent knowledge in the form of interconnected nodes and arcs, where nodes represent objects, concepts or situations, and edges represent the relations between them, including is-a (e.g. "a chair is a type of furniture") and part-of (e.g. "a seat is part of a chair").

As regards the origin of Semantic Networks [38], some researchers argue that Semantic Networks have come from Charles S. Peirce's existential graphs, while many of them pay tribute to Quillian, who was the first to introduce Semantic Networks in his semantic memory models [194]. Semantic memory refers to general knowledge (facts, concepts and relationship), such as a chair. It is different from another kind of long-term memory, i.e. episodic memory, which relates to some specific events, such as moving a chair. After Quillian, many variants of Semantic Networks were proposed.

Compared to formal knowledge representation and reasoning formalisms, such as predicate logics, Semantic Networks are relatively easy to use and maintain. On the other hand, they suffer from some limitations. For example, there is no formal syntax and semantics for Quillian's Semantic Network. This leaves room for users to have their own interpretations of constructors in Semantic Networks, such as the is-a relation. This approach may be seen as flexible for some, but it is also criticised for making it hard to integrate Semantic Networks while preserving their original meaning. Furthermore, Semantic Networks do not allow users to define the meaning of labels on nodes and arcs.

1.1.2 From Semantic Networks to Linked Data

RDF (Resource Description Framework) is a modern standard from W3C, addressing some of the issues related to classic Semantic Networks in terms of the lack of formal syntax and semantics. For example, the is-a relation can be represented by the subClassOf property in RDF, the semantics of which is clearly defined in the RDF specifications. It should be pointed out that RDF does not address all the limitations of a Semantic Network, e.g. RDF does not allow users to define concepts either. This is, however, addressed by OWL (Web Ontology Language), a W3C standard for defining vocabularies for RDF graphs. In OWL, the part-of relation is not a built-in relation like the subClassOf property. Instead, it is a user-defined relation that can be expressed by using the existential constructor. Description Logics [18, 184] are the underpinning of the OWL standard in the Semantic Web. More details of RDF and OWL can be found in Chap. 2.

Based on RDF and OWL, Linked Data is a common framework to publish and share data across different applications and domains, where RDF provides a graph-based data model to describe objects. OWL offers a standard way of defining vocabularies for data annotations. In the Linked Data paradigm, RDF graphs can be linked together by means of mappings, including schema-level mapping (subClassOf) and object-level mapping (sameAs).

1.1.3 Knowledge Graphs: An Entity-Centric View of Linked Data

In 2012, Google popularised the term *Knowledge Graph* (KG) with a blog post titled '*Introducing the Knowledge Graph: things, not strings*',[1] while simultaneously applying the approach to their core business, fundamentally to the Web search area. Among other features, the most typical one from the user's perspective is that, in addition to a ranked list of Web pages resulting from the keyword search, Google also shows a structured knowledge card on the right, which is a small box containing a summarised information snippet about the entity that probably solves the search. Such a knowledge card contains additional information relevant to the search, contributing to relieving the burden on the user's side to pick up relevant Web pages to find answers manually. Furthermore, relations with other entities in the KG are suggested, increasing the feeling of serendipity and stimulating further exploration by the user. In most cases, such knowledge cards sufficiently fulfil searchers' information needs, significantly improving the efficiency of Web search systems both in terms of time spent per search and quality of the results.

Inspired by the successful story of Google, knowledge graphs are gaining momentum in the World Wide Web arena. In recent years, we have witnessed an increasing industrial take-up by other Internet giants, which include Facebook's Graph Search and Microsoft's Satori, continued effort made in industrial research, e.g. Knowledge Vault [69], posting community-driven events (Knowledge Graph Tutorial in WWW2015[2]; KG2014[3]), entering into academia–industry collaborations and the establishment of start-ups that specialised in areas such as Diffbot[4] and Syapse.[5] All these initiatives, taken in both academic and industrial environments, have further developed and extended the initial Knowledge Graph concept which was popularised by Google. Additional features, new insights and various applications have been introduced and, as a consequence, the notion of knowledge graphs has grown into a much broader term that encapsulates a whole line of community effort in its own right, new methods and technologies.

To explain the subtle differences between knowledge graph and Linked Data better, we first need to introduce some basic concepts. Thus, we will postpone such detailed discussions to Sect. 2.4, after providing an introduction on the background knowledge in Sects. 2.1–2.3.

[1] http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.

[2] http://www.www2015.it/tutorials-19/.

[3] http://www.cipsc.org.cn/kg2/index_en.html.

[4] http://www.diffbot.com/products/.

[5] http://syapse.com/.

1.2 Knowledge Graph Technologies in a Nutshell

A knowledge graph based information system usually forms an ecosystem comprising three main components: construction, storage and consumption. Relevant knowledge graph technologies can be classified into one of these components of such an ecosystem where their contribution is most critical. As regards knowledge graph construction and storage, one finds technologies and tools for:

- knowledge representation and reasoning (languages, schema and standard vocabularies),
- knowledge storage (graph databases and repositories),
- knowledge engineering (methodologies, editors and design patterns),
- (automatic) knowledge learning including schema learning and population.

For the first three items, the majority of technologies are derived from the areas of KR, Databases, Ontologies and the Semantic Web. For knowledge learning, on the other hand, frameworks and technologies from Data Mining, Natural Language Processing and Machine Learning are typically employed.

From the consumption point of view, knowledge graphs' content can be directly accessed and analysed via query languages, search engines, specialised interfaces and/or generation of (domain/application-specific) graph summaries and visual analytics. In many other cases, a knowledge graph can enhance the effectiveness of a traditional information processing/access task (e.g. information extraction, search, recommendation, question answering, etc.) by providing a valuable background domain knowledge.

In this book, we cover knowledge graph technologies of all the above types, ranging from foundational representation languages like RDF to advanced frameworks for graph summarisation and question answering. Some of these technologies are useful for understanding knowledge graphs, while others help in exploiting knowledge graphs to support intelligent systems and applications.

1.3 Applications of Knowledge Graphs for Enterprise

Back in 2008, ongoing and future trends in semantic technologies were forecast to lie at the intersection of three main dimensions:

- natural interaction,
- the Web 2.0,
- service-oriented architectures.

If we abstract away from those particular terms, the actual meaning becomes quite simple:

- ease of *access* to computer systems by end users,

- *empowerment* of user communities to represent, manage and share knowledge in collaborative ways,
- machine *interoperability*.

Since then, countless research challenges have been faced in areas such as Knowledge Acquisition, Representation and Discovery, Knowledge Engineering Methodologies, Vocabularies, Scalable Data Management Architectures, Human–Computer Interaction, Information Retrieval and Artificial Intelligence, where semantic technologies have been involved, contributing to crucial advances in knowledge-intensive systems.

Now, like then, the value of data as the driving force behind intelligent applications remains. However, there is a new trend gaining momentum, which lies at the realisation that such a *value is directly proportional to the interlinkedness of the data* not only in complex, open-ended systems like the Web but also in specific enterprise applications based on combinations of both corporate and open data. More suited to look-up and relatedness operations, poorly formalised but highly interconnected data are becoming more popular than highly formalised but isolated datasets. The current application landscape, more oriented towards mobile and real time, is enforcing this new paradigm shift.

Google understood this very well and in 2012 started driving this trend in the industry by releasing their Knowledge Graph as a way to master such value, a large knowledge base that enhances its search engine's results with semantic-search information gathered from a wide variety of sources. Interestingly, the Knowledge Graph provides a way to connect the dots (entities) by means of explicit relations, with both entities and relations described following formal (but lightweight) models and reusing existing datasets like Freebase. After Google, other knowledge graphs arrived at the Internet scale, including those of Microsoft and Yahoo! Nowadays, it is the turn of enterprises and public administrations to leverage the Knowledge Graph concept at a corporate level in order to describe their data, enrich it by interlinking it with other knowledge bases both within and outside their environment and revitalise the development of knowledge-intensive systems on top of it.

Compared to 2008 [24], the interest in Market Intelligence and data-intensive sectors[6] and the role of knowledge graphs have increased dramatically while others, like corporate knowledge management and open government, are still there, though with slightly different foci. Next, we give an account of some selected applications that use knowledge graphs in such sectors, which will hopefully provide insight into the potential impact and future opportunities of knowledge graphs.

Corporate Knowledge Management

Open Innovation

Nowadays, especially after the recent financial downturn, companies are looking for much more efficient and creative business processes so as to place better solutions in the market in less time with less cost. There is a general impression that communication and collaboration, especially mixed with Web 2.0 approaches within companies

[6] IDG Enterprise Big Data report—http://www.idgenterprise.com/report/big-data-2.

and ecosystems (so-called Enterprise 2.0 [156]), can boost the innovation process with positive impacts on business indicators.

Open innovation [45] within an Enterprise 2.0 context is one of the most popular paradigms for improving the innovation processes of enterprises, based on the collaborative creation and development of ideas and products. The key feature of this new paradigm is that knowledge is exploited in a collaborative way flowing not only between internal sources, e.g. R&D departments, but also between external ones such as employees, customers, partners, etc. In this scenario, corporate knowledge graphs can be used to (i) support the semantic contextualisation of content-related tasks involving individuals and roles and (ii) help in discovering relations between communities of employees, customers and providers, with shared knowledge and interests.

The introduction of the open innovation paradigm in an enterprise entails not just a modification of corporate innovation processes but also a cultural change which requires support by an advanced technological infrastructure. Corporate knowledge has to be made explicit, exchanged and shared between participants, and therefore tools for knowledge management, analysis support and information structuring are required to make these tasks affordable and the knowledge available to all the involved actors. In addition, tools supporting the innovation process need to provide a high degree of interactivity, connectivity and sharing. In a scenario where collaborative work is not supported and members of a community could barely interact with each other, solutions to everyday problems and organisational issues rely on an individual's initiative. Innovation and R&D management are complex processes for which collaboration and communication are fundamental. They imply creation, recognition and articulation of opportunities, which need to be evolved into a business proposition at a subsequent stage. Interactivity, connectivity and sharing are the features to consider when designing a technological framework for supporting collaborative innovations [90]. All these characteristics can be identified in Enterprise 2.0 environments.

However, Enterprise 2.0 tools do not provide formal models which are used to create complex systems that manage large amounts of information. This drawback can be overcome by incorporating corporate knowledge graphs introducing computer-readable, interlinked representations of entities. Open innovation platforms similar to the one described in [1] leverage the concept of a corporate knowledge graph to relate people, interests and ideas in a corporate knowledge management environment throughout sectors, involving employees, clients and other stakeholders.

The impact of knowledge graphs through their application in open innovation is illustrated by their adoption in large corporations belonging to several sectors such as banking, energy and telecommunications (see further details in [45]), with companies such as Bankinter, Repsol and Telefonica, which have positioned themselves at the forefront of these efforts. What all these efforts have in common is the need to connect innovative ideas and people in order to orchestrate a healthy innovation ecosystem, addressing several challenges, like:

- Handling the information created by thousands of employees,
- evaluating their ideas efficiently,
- reducing false positives (ideas that reach the market and fail) and false negatives (valuable ideas which are rejected even before they can reach the market),
- stimulating the communication among people located around the globe, in different languages.

Intra-enterprise Micro-knowledge Management

As seen above, knowledge management is one of the key strategies that allow companies to fully tap into their collective knowledge. However, two main entry barriers usually limit the potential of this approach: (i) the barriers that employees encounter discouraging them from strong and active participation (knowledge providing) and (ii) the lack of truly evolved intelligent technologies that enable employees to easily benefit from the global knowledge provided by the companies and other users (knowledge consuming). In [188], miKrow, a lightweight framework for knowledge management, was proposed based on the combination of two layers that exploit corporate knowledge graphs to cater to both needs: a microblogging layer that simplifies how users interact with the whole system and a semantic engine that performs all the intelligent heavy lifting by combining semantic indexing and search of microblogs and users.

The miKrow interaction platform is a Web application that is designed as per the Web 2.0 principles of participation and usability. miKrow centres interaction around a simple text box user interface with a single input option for end users, where they are able to express what they are doing, or more typically in a work environment, what they are working at. This approach diverges from classical KM solutions which are powerful yet complex, following the idea of simplicity behind the microblogging paradigm in order to reduce the general entry barriers for end users. The message is semantically indexed against the underlying knowledge graph so that it can be retrieved later, as well as the particular worker linked to it. miKrow's semantic functionalities are built on top of the underlying knowledge graph, which captures and relates the relevant corporate entities.

Market Intelligence

According to the consulting company International Data Corporation (IDC) in its 2014 IDG Enterprise Big Data report, on an average enterprises spent $8M on leveraging value out of data in 2014, with penetration levels of 70 % and 56 % for large enterprises and SMEs, respectively. Improving the quality of the decision-making process (59 %), increasing the speed of decision-making processes (53 %), improving planning and forecasting (47 %), and developing new products/services and revenue streams (47 %) are the top four areas accelerating investment in data-driven business initiatives.

This trend is especially acute in the digital content and advertising sector. The communication between brands and consumers is set to explode. Product features are no longer the key to sales and the combination of both personal and collective benefits is becoming an increasingly crucial aspect. As a matter of fact, brands providing such value achieve a higher impact and consequently derive clearer economic benefits. On the other hand, millennials [98] are taking over, inducing a dramatic change in the way consumers and brands engage and what channels and technologies are required to enable the process. As a result, traditional boundaries within the media industry are being stretched and new ideas, inventions and technologies are needed to keep up with the challenges raised by the increasing demands of this data-intensive, in-time, personalised and thriving market.

HAVAS, the fourth largest media group worldwide, seeks to interconnect start-ups, innovators, technology trends, other companies and universities worldwide in one of the first applications of Web-scale knowledge graph principles to the enterprise world and media [46]. The resulting enterprise knowledge graph supports analytics and strategic decision-making for the incorporation of such talent within their first 18 months life span. Such an endeavour involves the application of semantic technologies by extracting start-up information from online sources, structuring and enriching it into an actionable, self-sustainable knowledge graph, and providing media businesses with strategic knowledge about the most trending innovations. While the previous success stories deal with the management of corporate knowledge within corporations, in this case the focus lies in creating competitive intelligence.

As we already know, innovation is often misunderstood and difficult to integrate into corporate mind-set and culture. So, why not activate relevant external talent and resources when necessary? The discovery and surveillance of trends and talent in the start-up ecosystem can be time consuming, though. HAVAS' knowledge graph sets its semantic engineering to run a surveillance monitoring of the entrepreneurial digital footprint, collecting and gathering fruitful insight and information, which provides the staff with clear leads for analysis. By automating part of the research process, analysts can get there faster and more accurately than competitors, leveraging millions of data points, and implementing consistency through a single and shared knowledge entry point. At the moment the knowledge graph is being opened to HAVAS' network, with teams in 120 offices around the world and clients, providing access to knowledge about the best-in-class talent to implement new thinking and cutting-edge solutions to the never-ending and evolving challenges within the media industry. Based on the knowledge graph, teams also rate and share experiences, ensuring that learning can be propagated across the network.

IBM Watson

IBM Watson is a cognitive computing platform available in the cloud, developed by IBM as an outcome of the *Jeopardy!* Q&A challenge[7]; cf. Sect. 7.2 and the Foreword of this book by Chris Welty. Watson uses Natural Language Processing and Machine Learning to discover insights from large amounts of unstructured data and provides a variety of services to work with this knowledge. Knowledge Graphs (such as Prismatic, DBPedia and YAGO) were at the core of the IBM's Q&A system.[8] IBM Watson services available today provide KGs capabilities through many services

[7] http://www.ibm.com/smarterplanet/us/en/ibmwatson/.

[8] IBM Journal of Research and Development, Vol. 56, No. 3/4, May/July 2012.

and application program interfaces (APIs), such as the Watson Concept Expansion and Insight.[9] Ongoing research and development aim at extending the availability of large structured knowledge bases to Dialog Services and other cognitive front ends.

1.4 How to Read This Book

1.4.1 Structure of This Book

This book introduces the key technologies for constructing, understanding and exploiting knowledge graphs. We hope you like reading this chapter so far. The rest of this book contains three parts, as illustrated in Fig. 1.1 (p. 11):

- **Part 1** contains Chaps. 2 and 3, in which we first introduce some basic background knowledge and technologies, and then present a simple architecture in order to help you to understand the main phases and tasks required during the lifecycle of knowledge graphs.
  - **Chapter** 2 introduces the background knowledge for studying and understanding the Knowledge Graph. Furthermore, we include a bit more discussion in the end to clarify the relations between Knowledge Graphs and Linked Data, as well as different purposes of building knowledge groups, e.g. for Web search versus for enterprise information systems.
  - **Chapter** 3 introduces a three-layer architecture of the Knowledge Graph application: (L1) Acquisition and Integration Layer; (L2) Knowledge Storing and Accessing Layer; and (L3) Knowledge Consumption Layer.
- **Part 2** is the main technical part for the Knowledge Graph, which contains Chaps. 4–7.
  - **Chapters** 4 **and** 5 further explain the layer L1 and address the state-of-the-art technology of knowledge acquisition and ontology construction.
  - **Chapters** 6 **and** 7 further explain the layer L3, where Chap. 6 introduces the key technologies of summarisation service, while Chap. 7 introduces the techniques of applying knowledge graphs in question answering (like the IBM Watson DeepQA).

  Based on the level of technical details, we have placed an asterisk on the titles of some chapters and sections, which contain detailed technical descriptions (e.g. formal definitions or formulas) or advanced topics (e.g. statistical/logical reasoning). Specifically, they are Chap. 5, Sects. 6.4 and 7.4.

[9] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/.

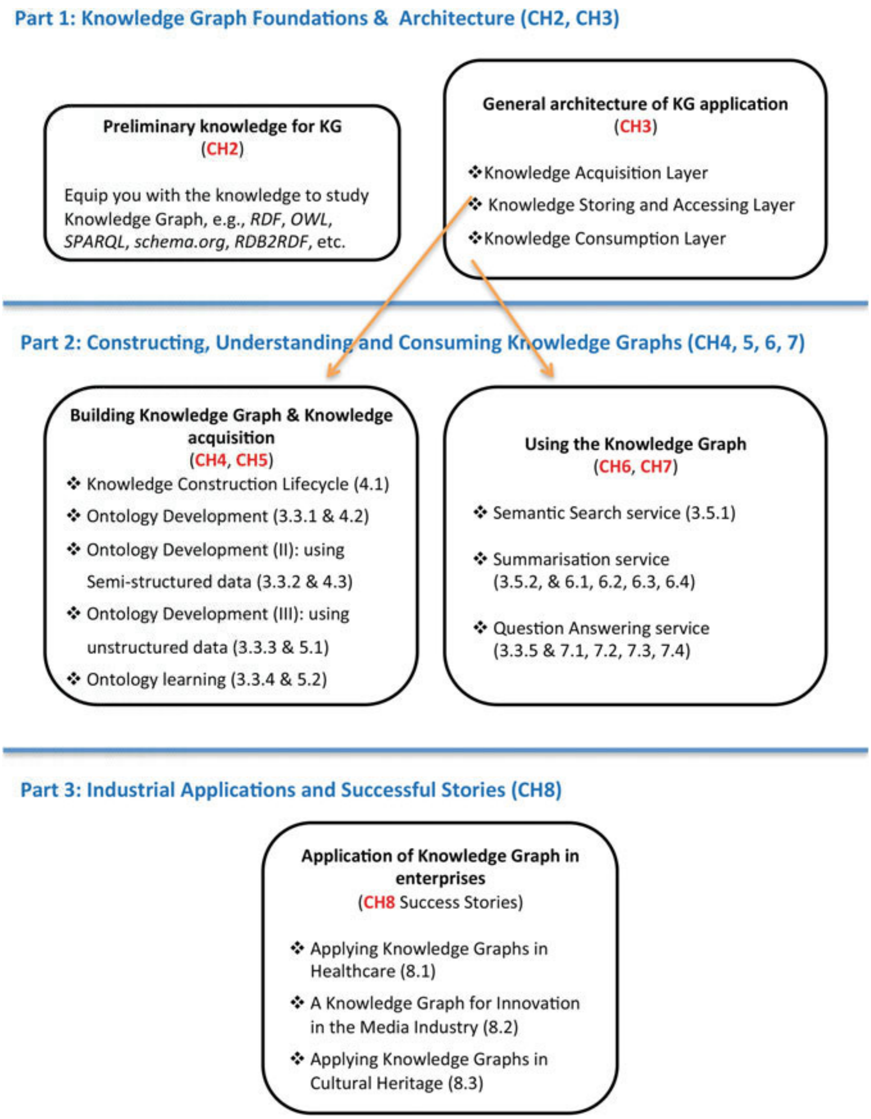
Part 1: Knowledge Graph Foundations & Architecture (CH2, CH3)
Preliminary knowledge for KG
(CH2)
Equip you with the knowledge to study Knowledge Graph, e.g., *RDF*, *OWL*, *SPARQL*, *schema.org*, *RDB2RDF*, etc.
General architecture of KG application
(CH3)
❖Knowledge Acquisition Layer
❖ Knowledge Storing and Accessing Layer
❖Knowledge Consumption Layer
Part 2: Constructing, Understanding and Consuming Knowledge Graphs (CH4, 5, 6, 7)
Building Knowledge Graph & Knowledge acquisition
(CH4, CH5)
❖ Knowledge Construction Lifecycle (4.1)
❖ Ontology Development (3.3.1 & 4.2)
❖ Ontology Development (II): using Semi-structured data (3.3.2 & 4.3)
❖ Ontology Development (III): using unstructured data (3.3.3 & 5.1)
❖ Ontology learning (3.3.4 & 5.2)
Using the Knowledge Graph
(CH6, CH7)
❖ Semantic Search service (3.5.1)
❖ Summarisation service (3.5.2, & 6.1, 6.2, 6.3, 6.4)
❖ Question Answering service (3.3.5 & 7.1, 7.2, 7.3, 7.4)
Part 3: Industrial Applications and Successful Stories (CH8)
Application of Knowledge Graph in enterprises
(CH8 Success Stories)
❖ Applying Knowledge Graphs in Healthcare (8.1)
❖ A Knowledge Graph for Innovation in the Media Industry (8.2)
❖ Applying Knowledge Graphs in Cultural Heritage (8.3)

Fig. 1.1 The three parts of the main content of this book

- **Part 3** (Chap. 8) introduces the successful stories of applying Knowledge Graph in Healthcare (8.1), Media Industry (8.2) and Cultural Heritage (8.3).

In Chap. 9 we conclude this book which shares some valuable experience of the editors and authors about their works on knowledge graphs.