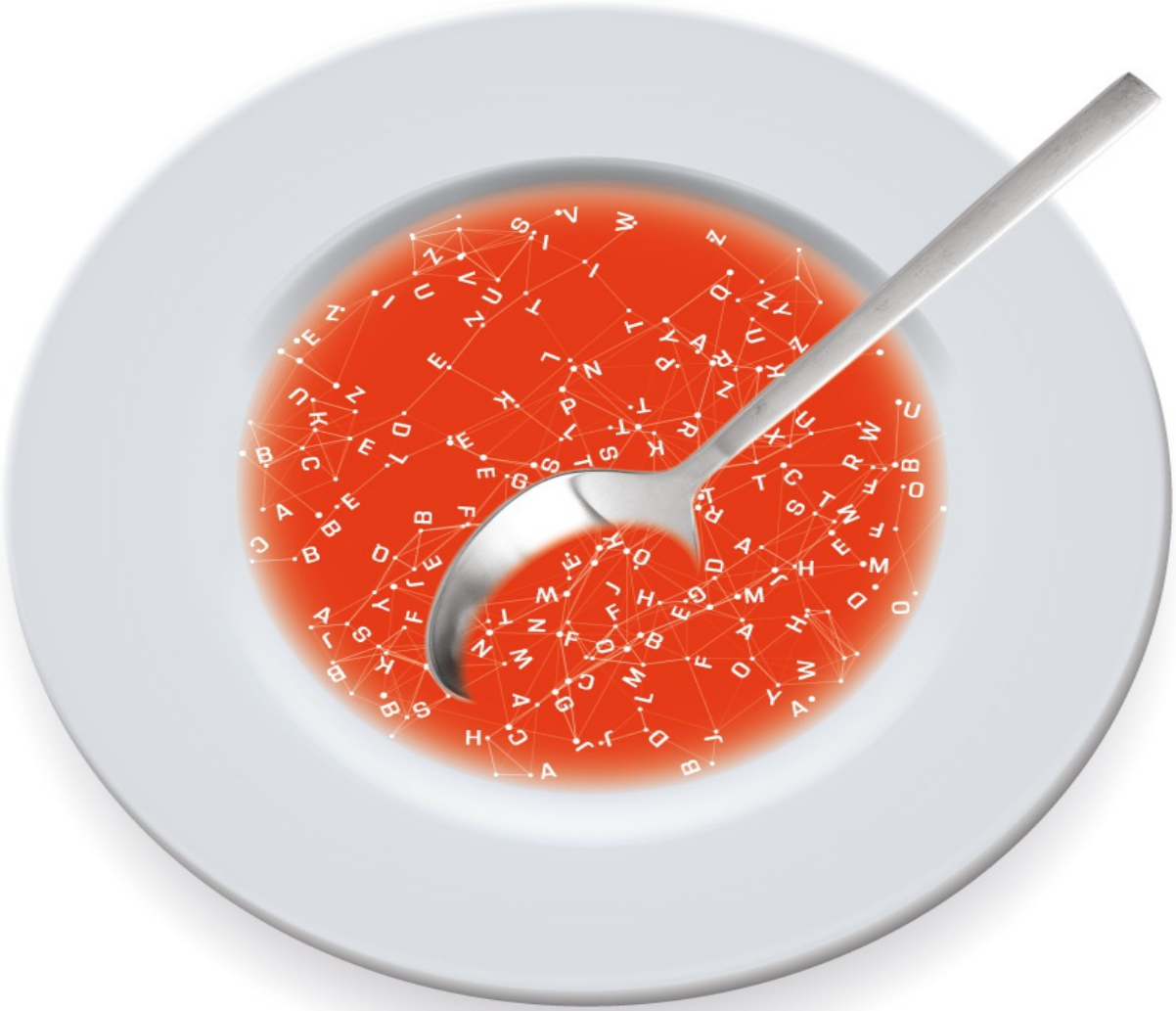
# THE KNOWLEDGE GRAPH COOKBOOK

## RECIPES THAT WORK




**ANDREAS BLUMAUER**

AND **HELMUT NAGY**



1st edition, 2020

# THE KNOWLEDGE GRAPH COOKBOOK

Recipes that work

Andreas Blumauer and Helmut Nagy

**monochrom**

monochrom

Published by:
edition mono/monochrom
Zentagasse 31/8, 1050 Vienna
Austria
fon: +43/650/2049451
edition-mono@monochrom.at

ISBN:
978-3-902796-70-7

Cover design & Layout:
Susan Härtig (Semantic Web Company)

Authors:
Andreas Blumauer, Helmut Nagy

Proofreading:
Anthony Miller

*We would like to thank the global Semantic Web community for their unwavering belief in the unlimited possibilities of connected data, collaborating with other people, networked organizations and societies, and for their tireless work to make this great vision a reality.*

*This is the only way to meet the great challenges facing humanity.*

# Table of Contents

# Why This Book?—Executive Summary

Most companies are increasingly data-driven and suffer from poor data quality due to widespread 'silo thinking.' On the other hand, better contextualized and connected data would help to overcome linear and departmental thinking, for example, to achieve more efficient collaboration between different stakeholders, higher customer satisfaction, or better service levels through holistic views of business objects.

Customer 360 initiatives or Know Your Customer (KYC), for example, involve the use of linked and holistic views of the customer, which are enriched with contextual information, to be able to develop personalized communication, make informed decisions, or put together an accurate product offer.

Knowledge graphs are certainly nothing new, but they have only been in use in industrial environments for a few years now. Accordingly, one speaks of 'Enterprise Knowledge Graphs' (EKGs). This refers to a wealth of approaches and technologies, all of which are aimed at getting a better grip on the chaos in enterprise data management. A central problem here is the siloing of data and the resulting additional costs and inefficiencies that arise along the entire data life cycle.

There are countless articles, slide decks and videos on the Internet about knowledge graphs. The topic is examined from different perspectives, e.g., from the point of view of artificial intelligence or in the context of extended possibilities for data analytics or information retrieval. Various standards, methods and technologies are suggested to the reader, and the moment of overstrain and disorientation typical for the arrival of new technologies quickly arises—so one wonders: “isn't there a step-by-step recipe out there explaining how to create a knowledge graph like there is for preparing a classic dish like Wiener schnitzel?”

This book is intended to help bring together and network different aspects of knowledge graphs. It will serve as a 'cookbook' for upcoming projects that want to integrate knowledge graphs as a central element. Above all, it should provide the reader a quick overview of why every data and content professional should or must deal with the topic in greater detail. The book

should also help to better assess the role of AI, especially that of explainable and semantic AI, in a post-Corona society.

We would like to thank everyone who supported this book, our colleagues at the Semantic Web Company (especially Susan Härtig for the great infographics, Sylvia Kleimann for her outstanding support, and Anthony Miller for the accurate editing and proofreading), and all partners and experts who made their valuable contributions in numerous discussions and were available to us as interview partners.

Once again it turned out that the management of knowledge graphs is above all one thing: a collaboration in which different perspectives have to be networked in order to ultimately create a larger context of meaning.

*Andreas Blumauer and Helmut Nagy*

# WHY WE WROTE THIS BOOK - ABOUT THE AUTHORS

# Andreas

When the right people—with quite different views and approaches—put their heads together, something clever usually emerges. This has always been one of my maxims and this principle can be applied to data and information just as well. Valuable knowledge can only be created through targeted networking: on an individual level ('aha moments'), for organizations, and sometimes even for society as a whole ('Eureka!').

The crux of the story, however, is that human thought and action is by no means predominantly concerned with 'networking' and 'synergies', but is at least concentrated on basic principles of 'specialization' and 'separation.'

This fundamental systemic problem is being increasingly solved—not least because of the penetration of networking technologies at all levels. What remains are people and their Babylonian confusion of language, their respective 'expert views' and their proprietary models of reality. While systems are being broken up, there is a crisis at every interface and the calls that it would be better to close the systems again are getting louder.

*My first PC: ATARI 1040 ST (from 1986)*

When I started programming computers at the beginning of the 1980s, there was no Internet. Like so many other young people at that time, I was inspired by the idea that we might enter a great epoch of humanity, as we now had machines at our side that could relieve us from difficult tasks.

One day I sat in front of the computer again and thought to myself: “What on earth am I doing here, alone in front of the device?” I lost interest in computer science. I was lucky: only a few years later the web started to

develop rapidly. A new fire began to blaze within me, and this time it was fueled by the idea that completely new forms of cooperation would be established through the World Wide Web (WWW).

During this time I completed my studies in business informatics and from then on wanted to inspire organizations with the idea that new forms of collaboration, especially knowledge networking, would be available with the new techniques of the Internet.

Most companies reacted skeptically and were hesitant to invest. Such vast potential could have been perceived as threatening. The young savages who launched the first wave of digital transformation at the time the first digital natives were born were ridiculed by the well established, when they ideally would have been listened to.

> ***"We always overestimate the change that will occur in the next two years and underestimate the change that will occur in the next ten. Don't let yourself be lulled into inaction."***
>
> —BILL GATES

But the Internet and associated technologies and methodologies have entered our lives faster and more powerfully than anyone had anticipated. Organizations are trying to counter this dynamic with agile working methods, and several levels below, data architects are working on new systems that are supposed to be one thing above all: less rigid and more adaptable than our IT legacy from the 80s and 90s.

When I began founding the Semantic Web Company with two friends in the early 2000s, the W3C under the leadership of Sir Tim Berners-Lee was already working on the next generation of the WWW: The Semantic Web should not only be a Web of Documents that is linked with hyperlinks, but also a Web of Data. Scalable data networking, better structuring of all data, machine-readable data, and ultimately, global knowledge networking were and remain the promises of this next generation of the Web.

At the core of the Semantic Web are so-called knowledge graphs[1] that link millions and millions of things together. This is how widely known services such as Alexa, Siri, or LinkedIn are operated. Knowledge graphs drive the current developments in artificial intelligence and make it possible to produce ever more precise AI applications at ever lower costs. Semantic Web technologies, some of which were invented 20 years ago, are now making their way into numerous companies and as always, when disruptive technologies are at the start, there is skepticism based primarily on ignorance.

In all the projects I have been involved in over the past 20 years, it has become increasingly clear to me how important it is for people to understand what is possible with AI—and IT in general—and what is not. Without this knowledge, either fear of uncontrollable AI dominates, or an exaggerated AI euphoria develops, or worse, both.

Knowledge graphs are not only data, but also a methodology for knowledge networking and efficient collaboration. Sound knowledge about why organizations should develop knowledge graphs and how they can do this is the key to success. Knowledge graphs are the result of years of development based on the Semantic Web and are now used in numerous industries; however, several questions about this topic still remain.

So I decided to write this cookbook, and I was lucky to have Helmut as a co-author, who can cover areas of knowledge that I could not. The intention of this book is to gain deeper insights into a highly interesting discipline and unfold its potential to change not only organizations, but also the world, because it is capable of nothing less than networking data, knowledge and people so that we can provide answers to small, large and even global problems.

Andreas Blumauer holds a Master's degree in business studies and information technologies from the University of Vienna. For the last 15 years, he has been CEO of the Semantic Web Company (SWC). At SWC, he is responsible for business development and strategic product management of the PoolParty Semantic Suite.

Andreas has been pioneering the Semantic Web frontier since 2001 and is author of the first comprehensive book in the Semantic Web area of expertise. He is an experienced consultant in the areas of information management, metadata management, linked data, search, analytics, machine learning and semantic technologies.

# Helmut

When I joined Semantic Web Company in 2010, the topic was still in its early stages, but knowledge management, semantic wikis and Enterprise 2.0 were already all over the place. I met Andreas a few years earlier at the SEMANTiCS conference. At that time it was called iSemantics and was held together with a knowledge management conference called iKnow. I was at the iKnow conference, but I was immediately fascinated by the topic Semantic Web, because I saw the connection between the two topics and how they have to interact to be successful.

It puzzled me even then to see these two communities that could have benefited so much from each other, but didn't even talk to each other. There were very few conferences (at least from my point of view) where the connection between these two topics was cultivated together. There was just this one community, but it failed because it tried to implement overly complex knowledge (management) systems that people ultimately avoided. Then next door there was this other community that was still somehow too much in love with academic and technical details to realize that it had the potential to change the entire game.

It was also quite a change for me when Andreas asked me if I wanted to join the company, because it basically allowed me to do what I had been working on for years. To work with people and companies to make the way they communicate and collaborate better and more efficient. I was in the fortunate position of being able to watch the rise of the Semantic Web from the front row and join as an active participant in it. The Semantic Web turned into linked data that eventually became knowledge graphs. The subject unfolded and matured as technologies evolved over time.

How do you know that something has matured? Because it made it into the Gartner's Magic Quadrant? Because there are more and more very large companies you talk to (and never expected to talk to)? Because your business is growing and you have more and more work? Well, most likely for all these reasons and many more. When Andreas asked me if we would like to write this book together, I was honored, but also cautious. Do I have enough relevant things to say? When I started to write it, I realized that this was the case and I hope that others will find it useful as well.

Helmut Nagy holds a Master's degree in journalism and communication studies from University of Vienna. For the last 7 years, he has been COO of the Semantic Web Company (SWC). At SWC, he is responsible for professional services and support and bringing in the business side into the product development of PoolParty Semantic Suite.

Helmut has been in the field of knowledge management for around 20 years and has been working as senior consultant in lots of projects introducing knowledge graphs to industry and public administration since joining SWC in 2010.

# PART 1: INTRODUCTION TO KNOWLEDGE GRAPHS

## HUNGER IS THE BEST SAUCE

# Why Knowledge Graphs?

**How do you "cook" a knowledge graph? Before we discuss specific variants of recipes and dishes, examine the individual ingredients, tools and methods or classify recipes, I would like to explain the main reasons why you should learn how to cook knowledge graphs. This chapter will outline the excellent results you can achieve. Here is a brief preview:**

- **Knowledge graphs (KGs) solve well-known data and content management problems.**
- **KGs are the ultimate linking engine for enterprise data management.**
- **KGs automatically generate unified views of heterogeneous and initially unconnected data sources, such as Customer 360.**
- **KGs provide reusable data sets to be used in analytics platforms or to train machine learning algorithms.**
- **KGs help with the dismantling of data silos. A semantic data fabric is the basis for more detailed analyses.**

Knowledge graphs (KGs) are recognized by many industries as an efficient approach to data governance, metadata management, and data enrichment, and are increasingly being used as a data integration technology. A central promise is that heterogeneous data, i.e., data from unstructured data sources up to highly structured data, can be harmonized and linked so that the resulting higher data quality can be used for subsequent tasks, such as machine learning (ML). KGs are, so to speak, the ultimate linking engine for the management of enterprise data and a driver for new approaches in Artificial Intelligence, from which it was hoped to create trillions of dollars in added value throughout the economy.[2] (Whether and in what form the corona crisis will give AI another strong boost is discussed in the chapter on the future of AI in a post-corona society.)

Typical applications for graph technologies are therefore unified views of heterogeneous and initially unconnected data sources that are generated automatically, such as Customer 360 to build a complete and accurate picture of each and every customer. These "virtual graphs" offer richer and reusable data sets to be used in analytics platforms or to train machine learning

algorithms. On this basis, advanced applications for knowledge discovery, data and content analytics can then be developed by using a semantic layer.

All of these promises sound tempting don't they, perhaps even too good to be true? Can knowledge graphs really do all of this and finally solve data and content management problems that we have been dealing with for decades?

If one analyzes the fundamentals of knowledge graphs, it quickly becomes clear that standing behind them are the promises of the 'Semantic Web.' The Semantic Web was initially designed by Sir Tim Berners-Lee with the aim of organizing nothing less than the entire WWW, resulting in probably the most heterogeneous and decentralized data landscape known to mankind. However, the web as we know it today has developed along a different path, and is characterized by the fact that once again, a few platforms like Facebook lock up content in silos. But parallel to this development, Semantic Web technologies have been able to unfold their potential especially in companies and now help to organize comparatively manageable and controllable data spaces.

As is so often the case, innovations that first took their first development steps on the Web have now arrived in companies. What took the form of the so-called 'Linked Open Data Cloud'[3] just a few years ago is now being readily implemented in companies, partly under different circumstances and with different motivations. We therefore also distinguish between two types of knowledge graphs: open knowledge graphs and enterprise knowledge graphs . Open knowledge graphs are open to the public, are often created and maintained by NGOs, government organizations or research institutions, and in many cases serve as a core element for the development of EKGs.

However, the goals are always very similar:

- Higher data quality, e.g., to train ML algorithms
- Reusability of data, e.g., to reduce the effort for data preparation
- Better interpretability of data and content, both for machines and humans
- Automatable processes for networking and analyzing data
- Find relevant data, personalize and contextualize it, i.e., integrate it into concrete business processes

- Use of data for automatic reasoning

This list could certainly be continued, but what remains at the core is the desire and motivation to adequately cope with the rapidly growing chaos of data.

The leading IT market analyst Gartner highlights knowledge graphs, graph databases and graph analytics as emerging technologies with significant impact on business, society and people over the next five to ten years in the following hype cycles: emerging technologies, analytics and business intelligence, artificial intelligence, data science and machine learning, data management, and for the digital workplace.

Ultimately, knowledge graphs are paving the way from silo-controlled business intelligence based on traditional data warehouses to a holistic approach to augmented intelligence. Augmented means that the Human-in-the-Loop (HITL) design principle is applied, in which various interest groups such as subject-matter experts (SMEs) or business users engage in a continuous mutual dialogue with AI machines throughout their daily work routines, with a knowledge graph becoming the central interface between such a systems' various actors.

# A Brief History of Knowledge Graphs

**Cooking is culture, and culture is based on history. History is not only what has happened, but also what has been piled up—the ground upon which we stand and build. Therefore, we should also have an understanding of where knowledge graphs come from if we want to become a maestro KG chef. Understanding the historical context is always paramount to understanding the possible paths one can take in the future.**

## Fast forward

- In 1736, graph theory was born: Leonhard Euler formulated the 'Königsberg Bridge Problem.'
- In 1976, John F. Sowa published his first paper on Conceptual Graphs.[4]
- In 1982, Knowledge Graphs were invented in the Netherlands. The theory of Knowledge Graphs was initiated by C. Hoede, a mathematician at the University of Twente, and F.N. Stokman, a mathematical sociologist at the University of Groningen.
- In 1999, Resource Description Framework (RDF) Model was published as a W3C Recommendation to lay a foundation for a Semantic Web.
- In 2001, Tim Berners-Lee, Jim Hendler and Ora Lassila published their ground-breaking article 'The Semantic Web'[5] in the Scientific American Magazine.
- In 2006, the DBpedia[6] project created a seed for the emergence of the Linked Open Data cloud by transforming Wikipedia content into linked data.
- In 2012, Google introduced their Knowledge Graph, and since then a lot of companies have started to build their own projects using knowledge graphs in various flavours.
- In 2018, The GQL Manifesto[7] was published to agree on a standard for a property graph query language.
- By the end of 2019 knowledge graphs had become mainstream. For example, Gartner states that "... a semantic knowledge graph can be

used to power other data management tasks such as data integration in helping automate a lot of redundant and recurring activities."[8]

- After decades of developing KGs, the discipline has also been influenced by a lot of other knowledge domains including mathematical logic, graph theory, information retrieval, computer linguistics, knowledge representation and reasoning, and most recently, the Semantic Web and machine learning.

## Semantic Web

In 2001, when the WWW was still in its infancy, its founder Tim Berners-Lee was already talking about the next big step: "The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

20 years later, we all know that things have developed more slowly and somehow in a different direction than expected; nevertheless, the W3C has laid the groundwork for a Semantic Web by publishing several important recommendations:

- 1999: Resource Description Framework (RDF) Model and Syntax Specification as a foundation for processing metadata to provide interoperability between applications that exchange machine-understandable information on the Web.
- 2004: Resource Description Framework (RDF) and RDF Vocabulary Description Language 1.0: RDF Schema (RDFS) as a standard for representing information about resources in the WWW. As a major update, RDF 1.1 was published in 2014.
- 2004: OWL Web Ontology Language as a language for defining and instantiating Web ontologies.
- 2008: SPARQL Protocol and RDF Query Language (SPARQL) to retrieve and manipulate data stored in RDF via so-called SPARQL endpoints. As a major update, SPARQL 1.1 was published in 2013.
- 2009: Simple Knowledge Organization System (SKOS) for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled

vocabulary.

- 2012: OWL 2 Web Ontology Language as an ontology language for the Semantic Web with formally defined meaning.
- 2012: R2RML, a language for expressing customized mappings from relational databases to RDF datasets.
- 2014: JSON-LD as a JSON-based serialization for Linked Data, which is now heavily used by Google for their rich snippets.[9]
- 2017: Shapes Constraint Language (SHACL) for validating graph-based data against a set of conditions.

In addition, the W3C has developed further Semantic Web standards, which are not only used on the Web, but have also led to technology adaptation in the business context. Here are some examples: RDFa, DCAT, Linked Data Platform (LDP) or PROV-O.

Based on this specification, the Linked Open Data Cloud manifested itself in 2006 as the first major success of the Semantic Web and since then, this collection of linked data available on the web has grown steadily and now covers more than 1,200 data sets based on RDF graphs.

Another big leap for the further development of a Semantic Web was the broad adoption of Schema.org.[10] Currently over 10 million sites use this vocabulary to markup their web pages and email messages. Many applications from Internet giants like Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

About 20 years after the beginning of this development, graph databases, many of them based on Semantic Web standards, now play a crucial role in helping companies to bring their data management into the 21st century. “As more companies set out to solve problems that are about the relationships between things, locations, and people, graph will become more popular in enterprises.”[11]

When Amazon announced Neptune[12] as a "fully managed graph database service" in 2017, it also said that "graphs change the world every day." In addition to Gremlin, a language for traversing graphs, Amazon Neptune also fully supports SPARQL 1.1.[13] So the deal was finally sealed: Semantic Web

standards were embedded within the WWW's infrastructure and this change has meanwhile also taken place in companies.[14]

## Labeled Property Graphs

Depending on the specific application scenarios, Semantic Web technologies are often the right choice, while the Labeled Property Graph (LPG) model offers an alternative, especially for analytical use cases, e.g., the analysis of social networks. Some say that property graphs are a stepping stone on the way to knowledge graphs.[15] The RDF standards for knowledge graphs were developed specifically for web-scale interoperability, while property graphs offer other advantages, in particular, they are closer to what programmers are used to.

The LPG model was developed in the early 2010s by a group of Swedish engineers. They developed an enterprise content management system in which they decided to model and store data as a graph.

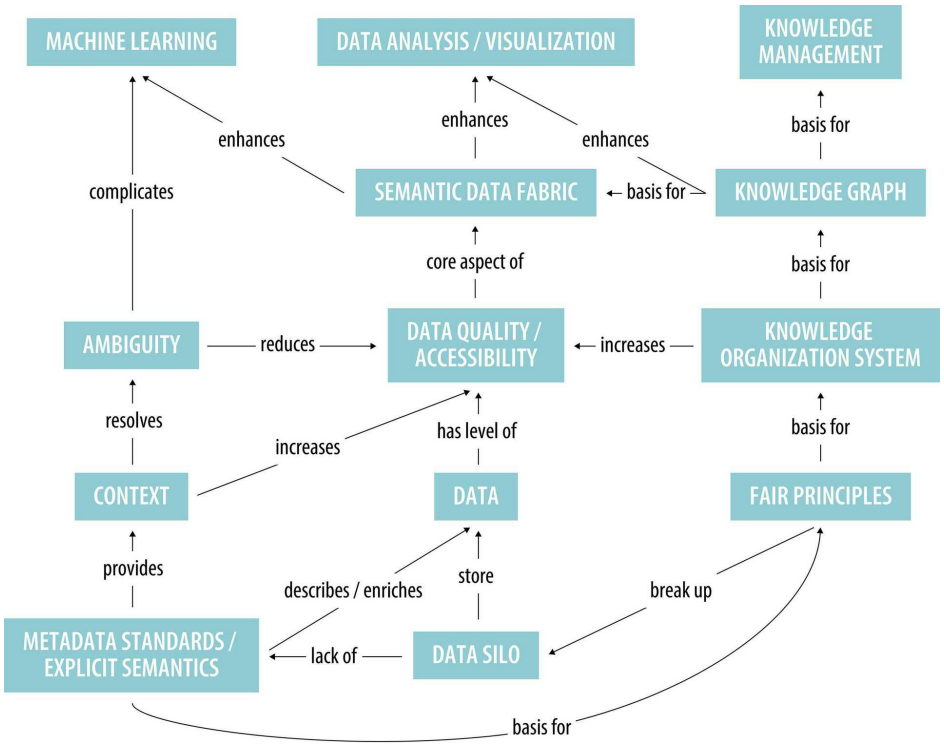
In contrast to the Semantic Web and its standards, property graphs have evolved in an organic way with every Property Graph Database vendor introducing their own query language (Cypher, Gremlin, PGQL, etc.). The GQL manifesto aims to fix this with the development of the GQL standard,[16] a key part of which is to create a fully-featured standard for graph querying, which has been in process since its introduction.

# Core concepts

**This chapter should serve as an introduction to readers and offers a multitude of entry points to the topic of knowledge graphs based on some well known basic concepts. You can start with any of them, no matter which one you read first, you will always traverse a network of concepts in which all are connected.**

- **Metadata should comply with FAIR principles.**
- **Ambiguous data is often a burden on data management. Adding more contextual information is the key to solving this problem.**
- **Data warehouses and data lakes are no longer state-of-the-art paradigms of data integration, but a data fabric will ultimately help dismantle data silos.**
- **Use established standards and methods for knowledge organization instead of developing your own proprietary approaches.**
- **Knowledge graphs are regularly confused with a methodology for knowledge visualization. We will take a closer look at this phenomenon.**
- **It’s all about things, not strings. With knowledge graphs, business objects themselves are placed in the center of data management. Customer 360 initiative should be based on this principle as well.**
- **Only an explainable AI creates trust. KGs play an essential role in any XAI strategy.**
- **Knowledge management often strives to design systems in which knowledge sharing on a large scale becomes possible. See why KGs support this approach.**

Each of the following core concepts sets a focus and thus a view on the whole topic. Which perspective to adopt depends mainly on what is to be improved or achieved with a knowledge graph. We'll see that, above all, the people and roles involved in this process determine which of the basic concepts and aspects are the initial focus. Over the course of the project, all other facets will gradually play a role and provide a holistic approach to knowledge graphs.




*Concept Graph*

## Metadata: Be FAIR

*"IN ALL CASES, METADATA SHOULD BE AS SELF-EXPLANATORY AS POSSIBLE"*

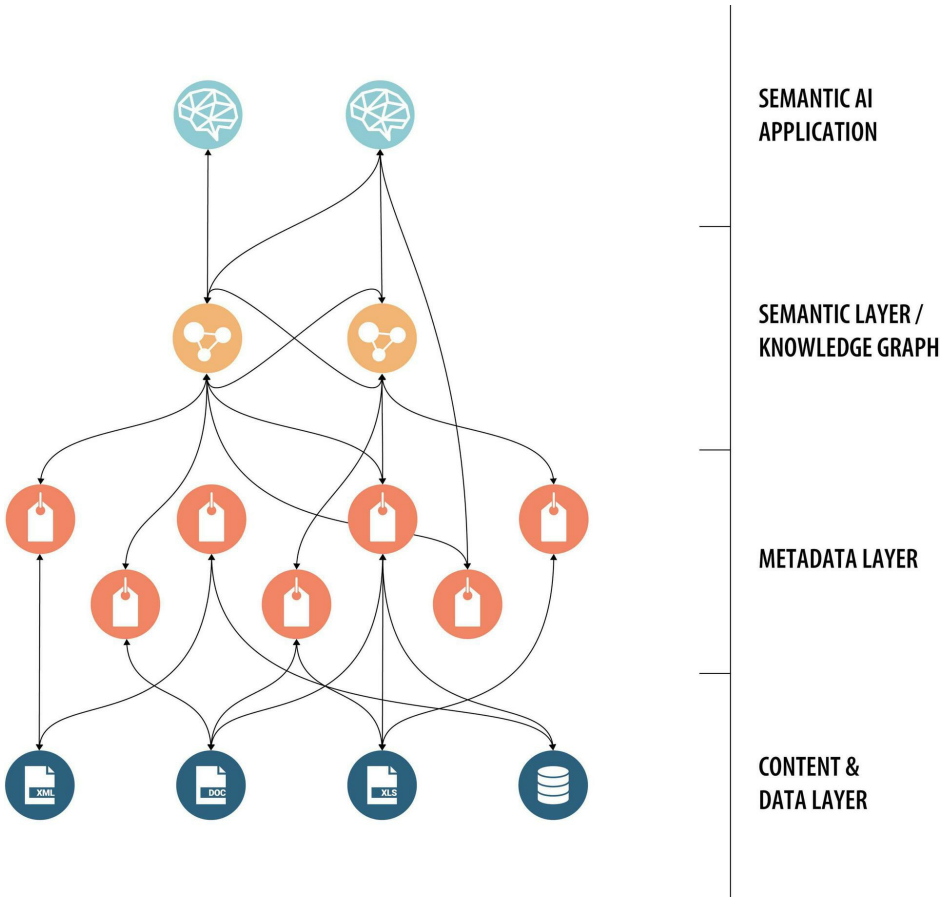
Metadata, i.e., data about data, helps to make data objects or documents more valuable by providing them with handles and entry points for better handling. Originally developed by the scientific community, the FAIR Data Principles provide a systematic overview that explains why metadata plays such an important role in data management. FAIR[17] stands for:

- **Findability**: Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
- **Accessibility**: Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository
- **Interoperability**: Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **Reusability**: Data and collections have a clear usage license and

provide accurate information on provenance.

Gartner differentiates between passive and active metadata.[18] While passive metadata is often generated by the system itself and used for archiving or compliance purposes, active metadata is frequently generated through text mining or automatic reasoning, which is used for further steps within a workflow or for advanced analysis later on. In short, active metadata makes data more valuable by leveraging all four aspects of FAIR as long as it is based on interoperable standards such as the Semantic Web.

In all cases, metadata should be as self-explanatory as possible. The most obvious strategy to achieve all these goals within an enterprise data management framework is to establish a central hub as a reference point that maps all different metadata systems and whose meaning is described in a standards-based modeling language. This central data interface is often referred to as the semantic layer and can be developed in organizations as an Enterprise Knowledge Graph. The relationship between data, metadata, and the semantic layer can be illustrated as follows:




*Four-layered Information Architecture*

Together with the data and content layer and the corresponding metadata, this approach unfolds into a four-layered information architecture, as shown above.

This emphasizes the importance of the semantic layer as a common umbrella for all types of data. Semantics is no longer buried in data silos, but linked to the metadata of the underlying data. It helps to "harmonize" different data and metadata schemata and different vocabularies. It makes the semantics (meaning) of metadata, and of data in general, explicitly available.

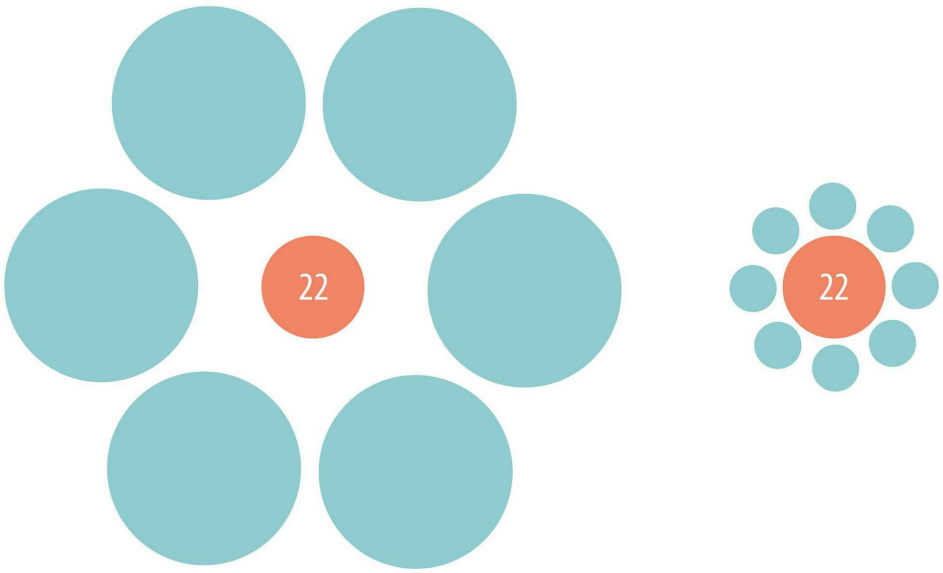
## Context overcomes Ambiguity

Of course, data and knowledge are not the same. So what's the missing link? Suppose someone wants to know what measures the EU has taken in recent years to reduce its $CO_2$ emissions. The figure "22" by itself wouldn't mean much, it is ambiguous. The fact that greenhouse gas emissions in the EU-28

fell by more than 22% between 1990 and 2016 is already interesting. Knowing the reasons for this development, namely the improvement of energy efficiency and the energy mix, gives even more context. It is clear that all of this data and facts are still relatively worthless until the source is known and how GHG and $CO_2$ correlate. Therefore, more context needs to be provided: A $CO_2$ equivalent is a metric measure used to compare the emissions of different greenhouse gases based on their global warming potential (GWP) by converting quantities of other gases into the equivalent quantity of carbon dioxide with the same GWP.

Providing contexts for data enrichment can be important at any point in the data lifecycle and can improve data quality. For example, during the data generation or acquisition phase, additional context can be created by adding metadata about the source and about the way the data was generated. Later, when interpreting and analyzing data or using it for visualization, the value of the context becomes even clearer. It makes a big difference when data can be embedded in a rich context. However, adding contexts can be costly, especially when done on an ad hoc basis, rather than using methods that repeatedly reuse a common knowledge base like an enterprise knowledge graph.

From an end-user perspective, the value of context information attached to a data object depends on the personal context. Looking back to the example from above: While a climatologist is not dependent on additional information about GHG and its correlation to $CO_2$, an average citizen wouldn’t be available to interpret the data at all. And what is valid for humans is even more important for machines: algorithms are highly dependent on context information, to learn from data precisely and unambiguously, even with smaller volumes of training data sets.




*Data is defined through its context*

Finally, let's take a look at the image above and find out how additional context makes a difference. Is 22% a sufficiently high number? That depends.

## Data Fabric instead of Data Silo

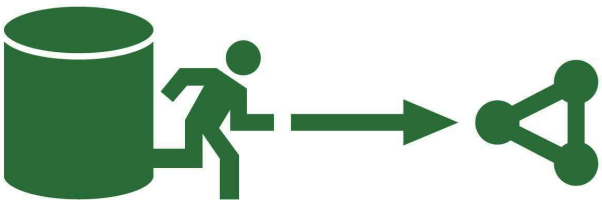
The first step towards a data-driven culture is data access, but many organizations have data silos that hinder this effort. Siloing data has its advantages and disadvantages. While you can maintain full control over the data and establish your own governance processes, data silos reduce speed, accuracy in reporting and data quality. Data silo owners cannot efficiently handle the full range of contexts that are potentially available to enrich their data.

> ***"Poor data quality is the unintended consequence of data silos and poor data & analytics governance."***
>
> (GARTNER, INC: 'THINK BIG, START SMALL, BE PREPARED — MASTER DATA MANAGEMENT', SALLY PARKER AND SIMON WALKER, OCTOBER 2019)

Data silos are isolated islands of data that make it extremely costly and difficult to extract data and use it for anything other than its original purpose. Typically, there is one data silo per application. Contradicting the principles

of FAIR (Findable, Accessible, Interoperable, Reusable) those data silos can have many reasons:

- *Application thinking*: Software applications and associated data structures are optimized for a specific purpose at a certain point in time. Efficient data exchange is rarely a primary requirement, proprietary data models are used instead. Instead of placing data and business objects at the center of system design, applications often continue to be lined up and optimized separately.
- *Political*: Groups within a company become suspicious of others who want to use their data, especially because data is often not self-explanatory. Rather, it must be interpreted with the knowledge of its history and context. Linking data across silos can also lead to undesired results, either because new contexts create new possibilities for interpretation or because problems with data quality become obvious.
- *Vendor lock-in*: Data silos are definitely in the interest of some software vendors. The less the data can be reused outside a platform, the more difficult the transformation to open standards is, the more tedious and unlikely a migration project will be, according to the calculus of some vendors.

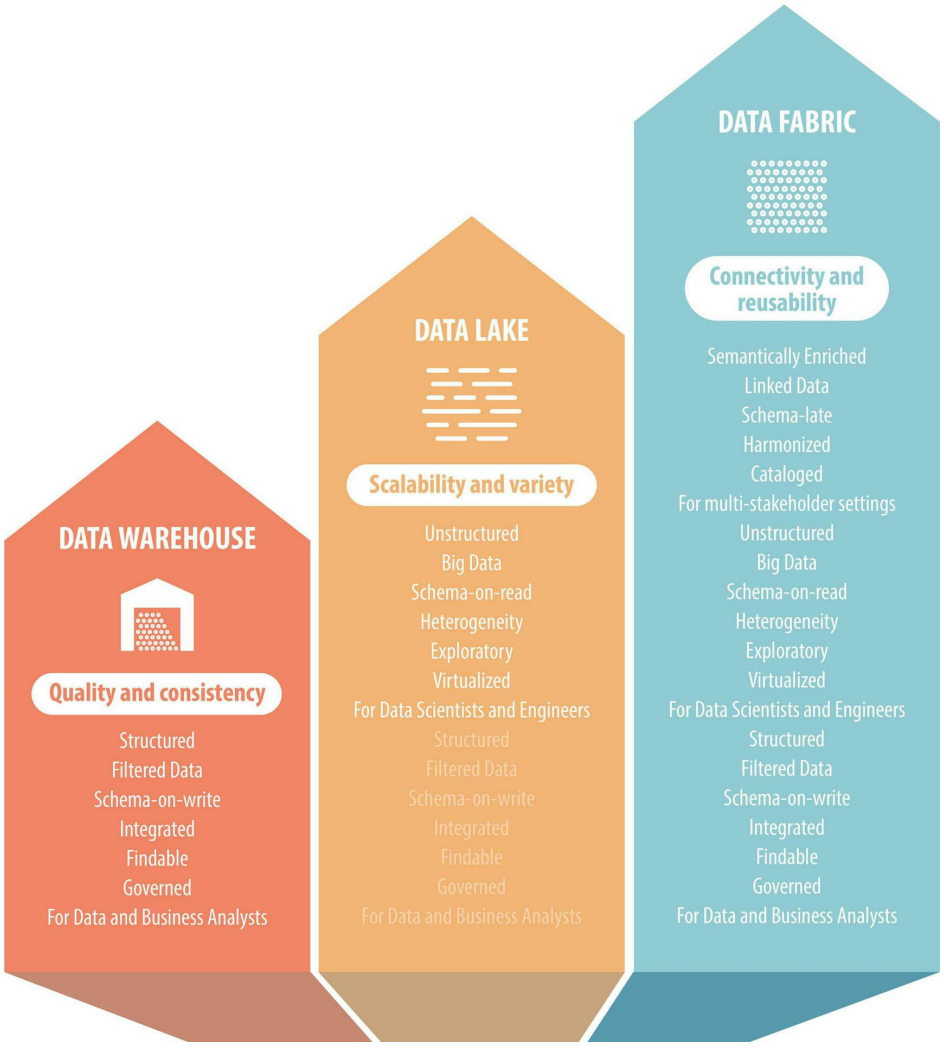


*Escape from Data Silos*

Instead of trying to physically migrate and replace existing data silos, EKGs support a different approach to data integration and linking. Through their ability to translate existing data models into semantic knowledge models (business glossaries, taxonomies, thesauri, and ontologies), knowledge graphs can serve as a superordinate database in which all rules for the meaningful and dynamic linking of business objects are stored.

This approach combines the respective advantages of Data Lakes and Data Warehouses and complements them especially with the advanced linking

methods that Semantic Graph Technologies bring with them.




*Fabrics are the modern successor to warehouses and lakes*

## Knowledge Organization: Make Semantics explicit

The organization of knowledge on the basis of semantic knowledge models is a prerequisite for an efficient knowledge exchange. A well-known counter-example are individual folder systems or mind maps for the organization of files. This approach to knowledge organization only works at the individual level and is not scalable because it is full of implicit semantics that can only be understood by the author himself.

To organize knowledge well, we should therefore use established knowledge organization systems (KOS) to model the underlying semantic structure of a

domain. Many of these methods have been developed by librarians to classify and catalog their collections, and this area has seen massive changes due to the spread of the Internet and other network technologies, leading to the convergence of classical methods of library science and from the web community.

When we talk about KOSs today, we primarily mean Networked Knowledge Organization Systems (NKOS). NKOS are systems of knowledge organization such as glossaries, authority files, taxonomies, thesauri and ontologies. These support the description, validation and retrieval of various data and information within organizations and beyond their boundaries.

Let's take a closer look: Which KOS is best for which scenario? KOS differ mainly in their ability to express different types of knowledge building blocks. Here is a list of these building blocks and the corresponding KOS.

| **Building blocks** | **Examples** | **KOS** |
|---|---|---|
| Synonyms | Emmental = Emmental cheese | Glossary, synonym ring |
| Handle ambiguity | Emmental (cheese) *is not same as* Emmental (valley) | Authority file |
| Hierarchical relationships[19] | Emmental *is a* cow's-milk cheese<br>Cow's-milk cheese *is a* cheese<br>Emmental (valley) *is part of* Switzerland | Taxonomy |
| Associative relationships | Emmental cheese *is related to* cow's milk<br>Emmental cheese *is related to* Emmental (valley) | Thesaurus |
| Classes, properties, constraints | Emmental *is of class* cow's-milk cheese<br>Cow's-milk cheese *is subclass of* cheese<br>Any cheese *has exactly one* country of origin<br>Emmental *is obtained from* cow's milk | Ontology |

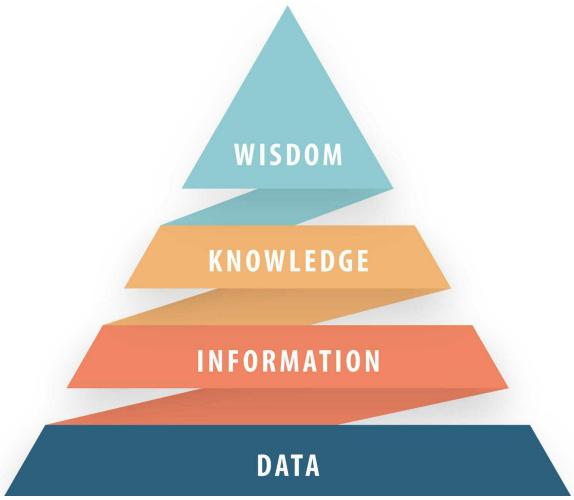
The Simple Knowledge Organization System (SKOS),[20] a widely used standard specified by the World Wide Web Consortium (W3C), combines numerous knowledge building blocks under one roof. Using SKOS, all knowledge from lines 1–4 can be expressed and linked to facts based on other

ontologies.

Knowledge organization systems make the meaning of data or documents, i.e., their semantics, explicit and thus accessible, machine-readable and transferable. This is not the case when someone places files on their desktop computer in a folder called "Photos-CheeseCake-January-4711" or uses tags like "CheeseCake4711" to classify digital assets. Instead of developing and applying only personal, i.e., implicit semantics, that may still be understandable to the author, NKOS and ontologies take a systemic approach to knowledge organization. We will deal with this in more detail in the chapter on Knowledge Organization Systems.

## Knowledge Management—better with Knowledge Graphs

Data is not information, and information is not yet knowledge. For decades there has been a heated debate about the fact that a functioning knowledge management system is not something that can be installed in an intranet like any software system, and that knowledge cannot be stored in documents or databases. With the rise of knowledge graphs, many knowledge management practitioners have questioned whether KGs are just another database, or whether this is ultimately the missing link between the knowledge level and the information and data levels in the DIKW pyramid as depicted here.




*DIKW Pyramid: From Data to Wisdom*

Knowledge graphs stimulate cross-departmental and interdisciplinary communication and help to orchestrate information flows or to link activities

and expertise or ultimately even knowledge workers in larger organizations that are initially isolated from each other, e.g., through mechanisms of semantic matchmaking. Knowledge graphs should therefore be able to fulfill an abundance of long-desired wishes of the knowledge manager community. Can KGs help to turn data and information into knowledge? Let's approacht this systematically—which typical challenge in knowledge management can be met with the help of KGs and how?

| **Challenges in Knowledge Management[21]** | **Knowledge Graph Capabilities** |
|---|---|
| Keeping people motivated to share data and information | Provide controlled vocabularies so that people can trust that their sharing activities will be successful |
| Keeping shared information up to date and accurate | Continuous content analysis as part of the ongoing work on the knowledge graphs keeps both metadata and shared information up-to-date |
| Interpreting data and information effectively | KGs help to ensure that information provided by a person or group is mapped or standardized so that it is meaningful to others in the organization. |
| Ensure relevancy: make it easy for people to find what they are looking for | Algorithms for information retrieval focus mainly on relevance scoring. KGs enable semantic content classification and contextual search, which allows a more precise calculation of relevancy. |
| Rewarding active users | Instead of simply rewarding more active users with stars or thumbs up, they are rewarded directly with the help of knowledge graphs: more active users benefit from more precise and relevant recommendations from the system. Knowledge and interest profiles are continuously updated and expanded using semantic technologies |
| Facilitating collaboration among team members and different teams | Semantic matchmaking on the basis of graphs helps to network people according to their knowledge profiles |
| Providing more user-friendly IT-Systems | KGs are changing the way business users and developers can look at data. It is no longer the regime of database engineers that determines how applications are developed, but rather how we as end users think about and interpret data. KGs provide data as an interface for developers and users along the actual business logic |
| Facilitating individual | Based on personal skills, competencies, interests and learning styles, |

| | |
|---|---|
| learning paths | there are many ways through a curriculum. With a KG, the learning systems are equipped with recommendation systems that help people to identify individual learning paths while combining individual and organizational interests |
| Not-invented-here-syndrome | Overcoming resistance within an organization against external knowledge requires a stronger focus on the principle of "inclusion." Ongoing work on knowledge graphs can be organized in such a way that they are perceived as highly collaborative activities, and thus KGs will be broadly accepted as central knowledge hubs |

It is clear that knowledge graphs will not replace a comprehensive knowledge management program, but they should be embedded as an integral part of such a program. Ultimately, every department and every person involved in a KM program should be included in the process of designing, building and shaping an enterprise knowledge graph, which then not only links data but also brings people and their knowledge together.

Coming back to the DIKW pyramid: knowledge graphs have great potential to finally link the more technically oriented layers of data and information with the human-centric KM topic of knowledge. I fear that the wisdom must originate elsewhere, and the missing link between wisdom and knowledge remains to be found.

***“Knowledge talks, wisdom listens.”***

—JIMI HENDRIX

## Knowledge Graphs are not just for Visualization

*"GOOD VISUALIZATIONS SERVE A SPECIFIC PURPOSE AND CAN ADDRESS SPECIFIC TOPICS OR LEARNING SITUATIONS"*

People who come into contact with knowledge graphs for the first time inevitably think of visualizations of networks, in many cases of social networks. On the one hand, this is a good sign because it confirms the idea that semantic networks (in contrast to relational data models) are very similar

in structure to how people actually think. On the other hand, it often stands in the way of further considerations as to the purpose knowledge graphs may actually serve.

*Remember: knowledge graphs are data.*

Primarily, a knowledge graph represents, among other things, a model of a knowledge domain created by experts using intelligent algorithms of machine learning. It provides a structure and a common interface—not necessarily a visualization—for all important data and allows the creation of multifaceted relationships between databases. The knowledge graph is a virtual data layer on top of the existing databases or data sets to connect all data—whether structured or unstructured—at scale.

*This is a graph, what else?*

## *‘Good’ and ‘bad’ graph visualizations*

Visualizations can support the process of knowledge modeling, especially in the first phase of ontology creation, which is often characterized by collaborative and communicative processes. The benefit of a successful visualization of knowledge graphs in the search, analysis and discovery phase becomes even clearer. How are things linked, e.g., in what hierarchical relation are they to each other? Graphical visualizations can answer such questions directly. Good visualizations serve a specific purpose and can address specific topics or learning situations, e.g., in e-learning systems. Bad visualizations reveal nothing more than a network or a cumbersome graph showing the user that everything is very complex and supposedly chaotic.

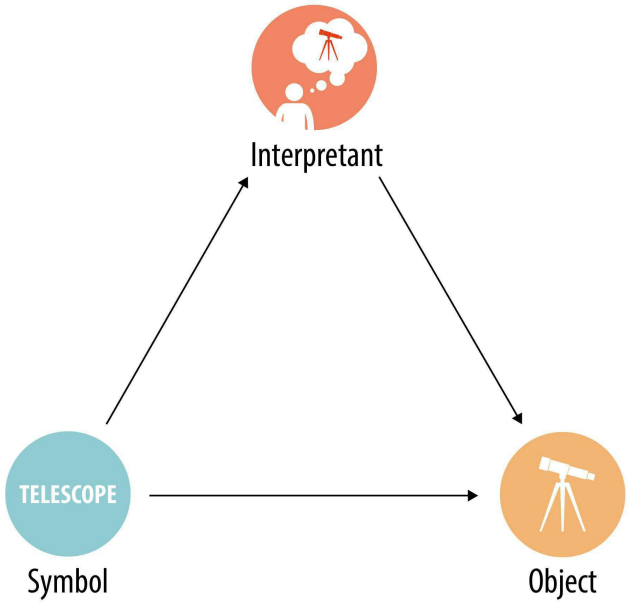
*Knowledge graphs are much more than 'just' visualizations.*

A closer look at the entire life cycle of the knowledge graph creates a holistic view of the creation process and usage options of knowledge graphs. One quickly discovers that visualization supports a certain phase of graph-based data management, namely the analysis of data, very well. Graph visualizations like PoolParty GraphViews[22] therefore support some tasks very efficiently, especially within the user loop of the KG life cycle. But visualization is by far not the only purpose of a knowledge graph.

## Things, not strings

*"SYMBOLS LIKE STRINGS OR NAMES FOR THINGS ARE NOT THE SAME AS THE OBJECTS (THINGS) THEY REFER TO"*

Entity-centric views of all types of data sources provide business users and analysts with a more meaningful and complete picture of all types of business objects. This method of information processing is as relevant to customers, citizens or patients as it is to knowledge workers such as lawyers, doctors, or researchers. In fact, the search is not for documents, but for facts about entities and things needed to bundle them and provide answers to specific questions.



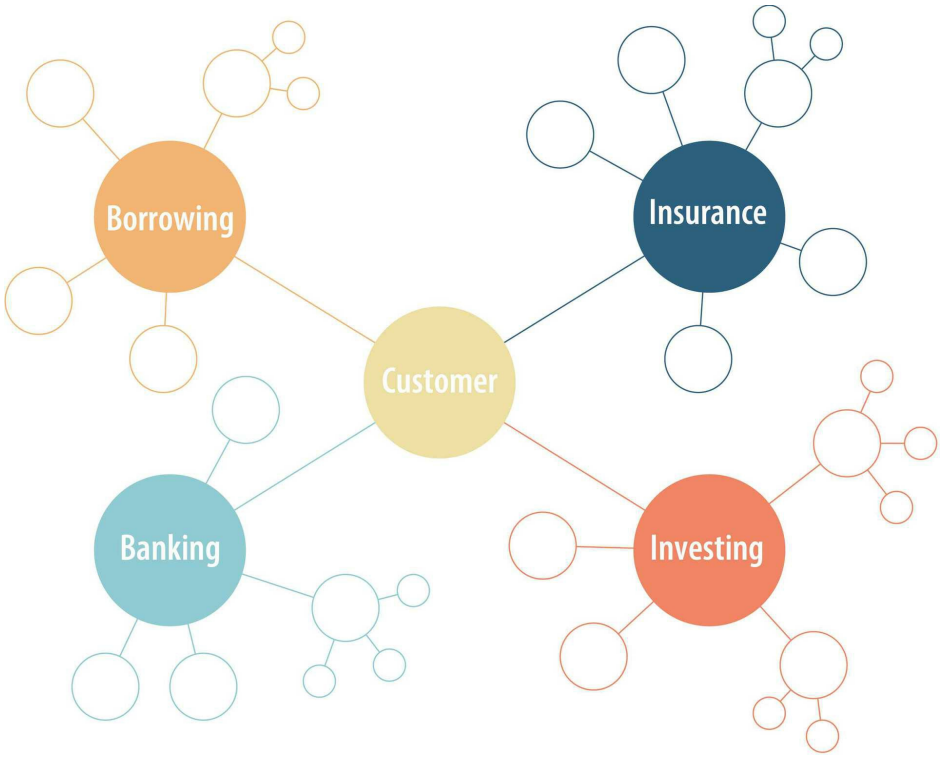

*Semiotic triangle: Things, not only strings*

Symbols like strings or names for things are not the same as the objects (things) they refer to. This is illustrated by the 'semiotic triangle' shown here. However, these two aspects of an entity are regularly confused by the interpreter, furthering the Babylonian confusion. Each string can refer to different things, and each thing can have different names.

Semantic knowledge graphs support a holistic view of all kinds of things, e.g., business objects (products, suppliers, employees, customers, etc.) including their different identifiers, names and relationships to each other. Information about business objects can be found in structured (relational databases), partially structured (XML) and unstructured (text) data objects. However, people are not interested in containers, but in business objects themselves.

For example, we want to get a 360-degree view of the customer based on a consolidated and integrated data set that includes all relevant relationships between a company and its customers. This networked data set may contain information about customer profiles, transactions, preferences or customer relationships with other companies. Companies usually try to build such a holistic view in order to optimize customer satisfaction, customer loyalty and, in turn, sales (Customer 360). Knowledge graphs help to do this in an agile way.

Here is an example from the financial services industry based on the widespread Financial Industry Business Ontology (FIBO).[23]




*Business objects defined as ontology form the basis for Customer 360*

Knowledge graphs based on FIBO help consolidate data from various sources to eventually look at each customer as a whole and in a harmonized way: Virtually on the other side of the "Customer 360" coin is the "Know Your Customer/ Anti-Money Laundering" use case. The challenges for KYC/AML revolve equally around the integration of internal systems, extended by the challenge of networking internal systems with external data sources.

## Machine Learning and Artificial intelligence: Make it explainable

*"MACHINE LEARNING ALGORITHMS LEARN FROM HISTORICAL DATA, BUT THEY CANNOT DERIVE NEW INSIGHTS FROM IT"*

While AI is becoming a part of our daily lives, many people are still skeptical. Their main concern is that many AI solutions work like black boxes and seem to magically generate insights without explanation.

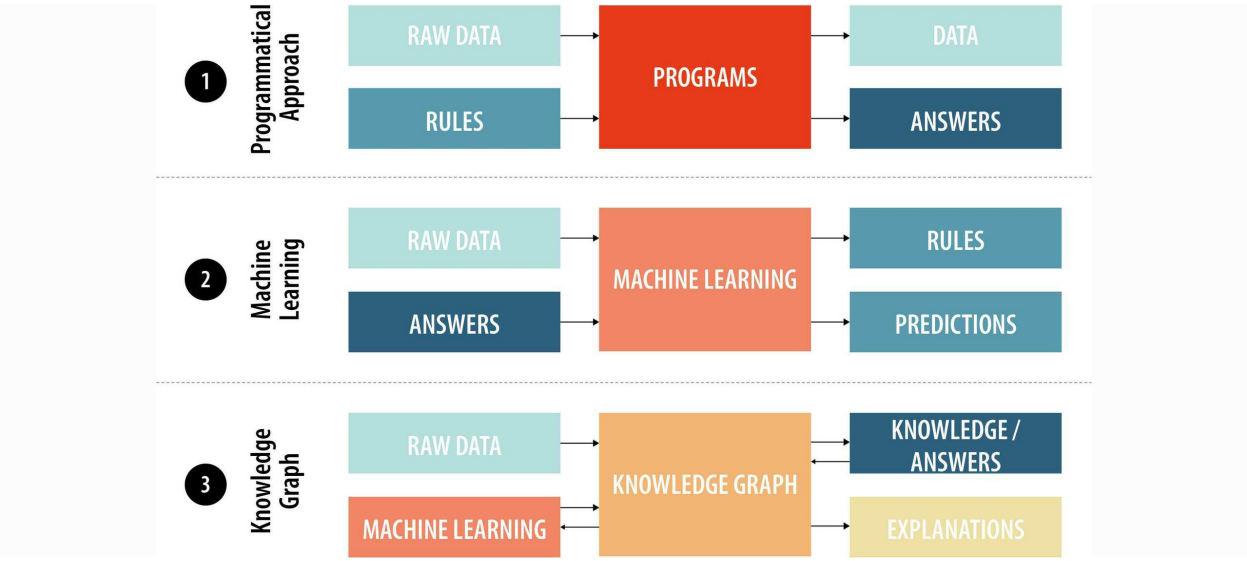
In addition to the benefits they can bring to the area of enterprise data management, knowledge graphs are increasingly being identified as building blocks of an AI strategy that enables explainable AI following the Human-in-the-Loop (HITL) design principle.

*Why does artificial intelligence often work like a black box?*
The promise of AI based on machine learning algorithms, e.g., deep learning, is to automatically extract patterns and rules from large datasets. This works very well for specific problems and in many cases helps automate classification tasks. Why exactly things are classified in one way or another cannot be explained. Because machine learning cannot extract causalities, it cannot reflect on why certain rules are extracted. Deep learning systems, with their hidden world of abstract and unknown layers and patterns, are especially difficult to explain.

Machine learning algorithms learn from historical data, but they cannot derive *new* insights from it. In an increasingly dynamic environment, this is causing skepticism because the whole approach of deep learning is based on the assumption that there will always be enough data to learn from. In many industries, such as finance and healthcare, it is becoming increasingly important to implement AI systems that make their decisions explainable and transparent, incorporating new conditions and regulatory frameworks quickly. See, for example, the EU's guidelines on ethics in artificial intelligence,[24] which explicitly mention the requirement for explainable AI (XAI).[25]

*Can we build AI applications that can be trusted?*
There is no trust without explainability. Explainability means that there are other trustworthy agents in the system who can understand and explain decisions made by the AI agent. Eventually, this will be regulated by authorities, but for the time being the most reasonable option we have is making decisions made by AI more transparent. Unfortunately, it's in the nature of some of the most popular machine learning algorithms that the basis of their calculated rules cannot be explained; they are just “a matter of fact.”

The only way out of this dilemma is a fundamental reengineering of the underlying architecture involved, which includes knowledge graphs as a prerequisite to calculate not only rules, but also corresponding explanations.




*Towards Explainable AI*

This reworked AI architecture based on the Semantic AI design principle introduces a fundamentally different methodology and, thus, additional stakeholders with complementary skills. While traditional machine learning is done mainly by data scientists, knowledge scientists are the ones who deal with semantic AI and explainable AI efforts and are involved in the entire knowledge graph life cycle.

At the core of the problem, data scientists spend more than half of their time collecting and processing uncontrolled digital data before it can be sifted for useful nuggets. Many of these efforts focus on building flat files with unrelated data. Once the features are generated, they begin to lose their relationship to the real world.

An alternative approach is to develop tools for analysts to directly access an enterprise knowledge graph to extract a subset of data that can be quickly transformed into structures for analysis. The results of the analyses themselves can then be reused to enrich the knowledge graph. The semantic AI approach thus creates a continuous cycle in which both machine learning and users are an integral part. Knowledge graphs act as an interface in between, providing high-quality linked and normalized data.

# Application scenarios

**In this chapter readers will find some recipes for different scenarios where knowledge graphs make a difference. We have identified five classes of scenarios, most of which are at the beginning of every knowledge graph initiative that can be used as a blueprint for any project in this area. For each scenario we will also give some more concrete examples.**

- **Orchestrating knowledge workflows in collaborative environment**
- **Unify unstructured and structured data in a Smart Data Catalog**
- **Connecting the dots: Search and Analytics with Knowledge Graphs**
- **Deep Text Analytics (DTA)**
- **Excellent Customer Experience**

The application scenarios described in this chapter give a good overview of most of the known problems we are currently confronted with in our daily work.

Loosely coupled workflows and heterogeneous system landscapes make effective access to information difficult. Structured and unstructured data live in different worlds that are not connected to each other. A complete overview or in-depth analysis with all available data is associated with high costs. And this is especially not possible when it is time-critical. All these systemic shortcomings also prevent the achievement of a consistent customer experience.

The key to solving all these problems lies in the ability of the knowledge graph to link all your data in a meaningful way. So, let's take a look at the different scenarios to see if you get an appetite for them too.

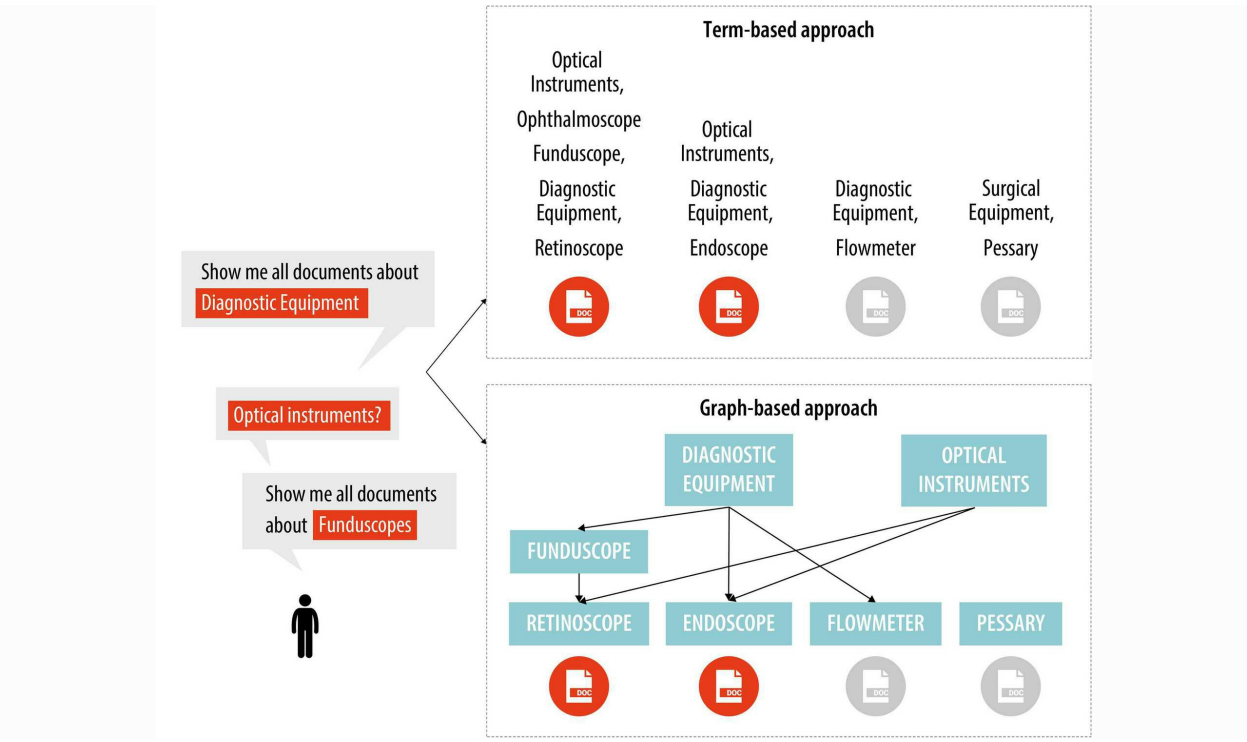
## Orchestrating knowledge workflows in collaborative environments

*"THE BASIC RULE IS THAT CONTENT TAGGING AND*

## *CLASSIFICATION SHOULD TAKE PLACE AS SOON AS POSSIBLE AFTER THE CONTENT IS CREATED"*

Most companies work with a variety of systems that are not well integrated. Information is located in different places and cannot be accessed as a whole. This prevents you from quickly gaining an overview of relevant topics. One of the simplest and most basic application scenarios for a knowledge graph is the integration of semantic or concept-based tagging into your (mostly collaborative) content production environments, be it CRMs, DMSs, CMSs or DAMs, etc.

These integrations basically always follow the same recipe. The basic rule is that content tagging and classification should take place as soon as possible after the content is created. This means that, direct integration into the system is ideal, or even better, into the existing tagging functionality of these systems. Many of them have such a system, but since they are usually based on simple terms and not on semantic (knowledge) graphs, they are of limited value.

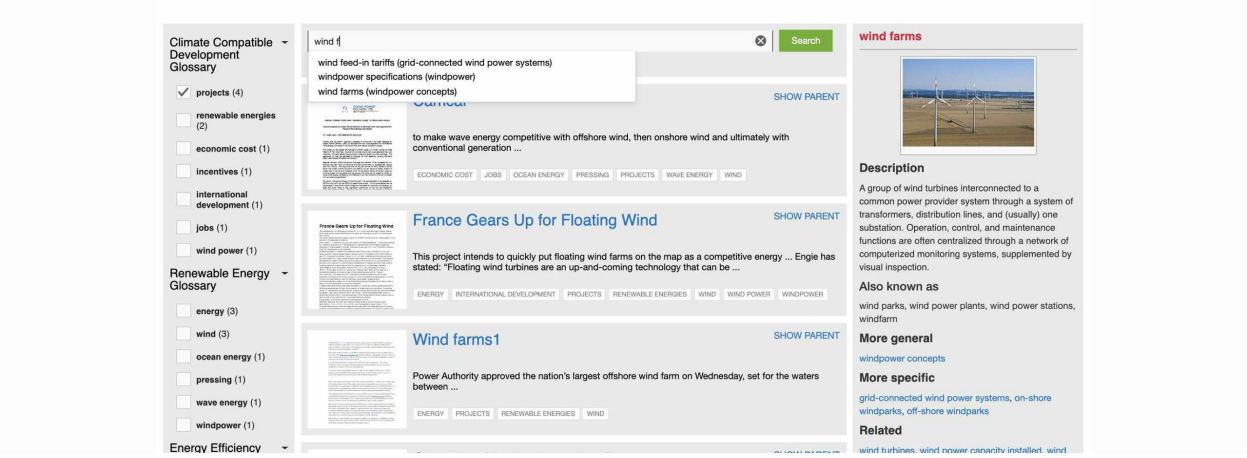



*Term Based vs. Knowledge Graph Based Tagging*

Of course, tagging should be done automatically in the background to allow

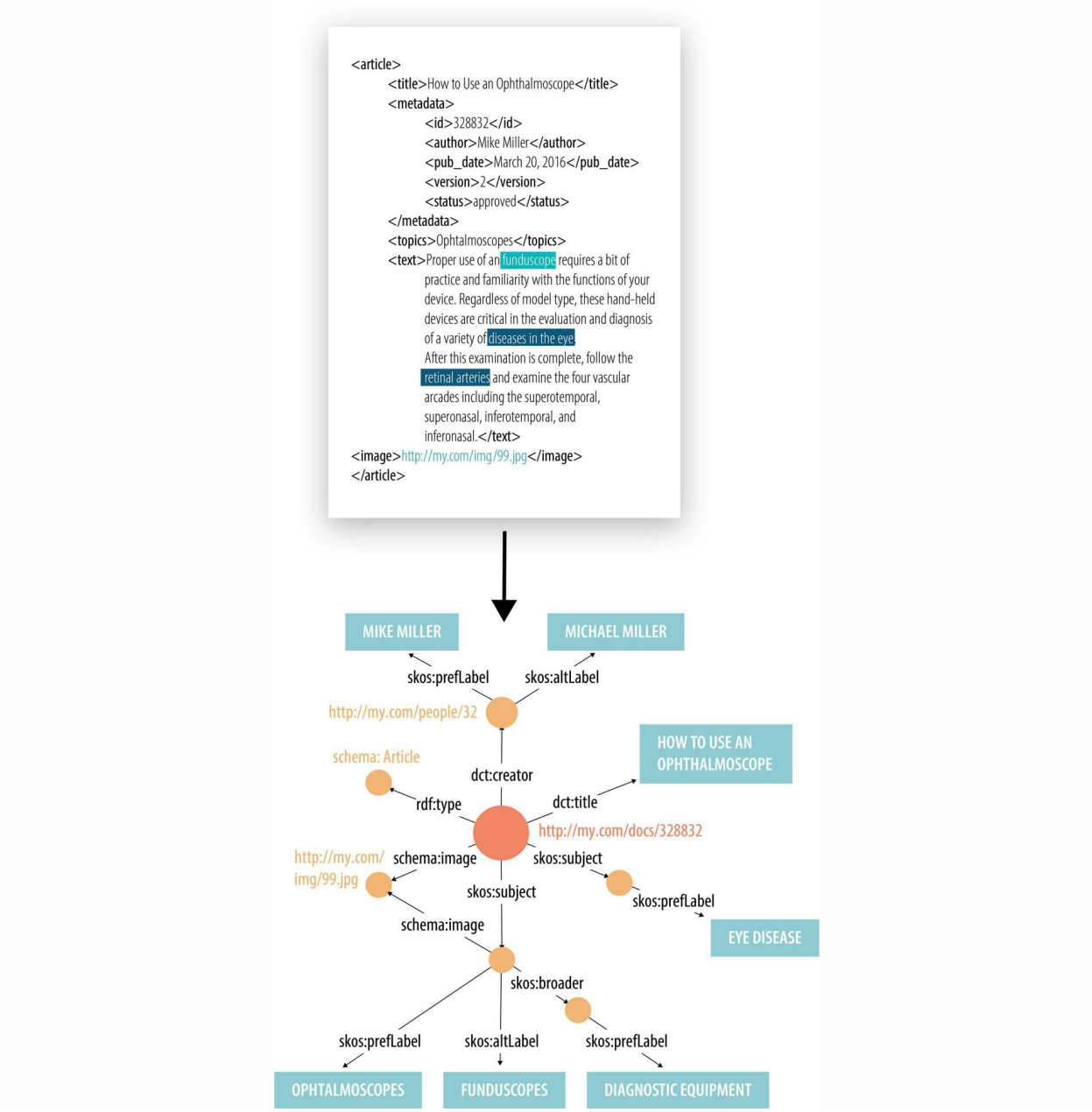
for a smooth integration into current content production workflows and avoid creating additional work for content creators that might prevent adoption. It is recommended to set up a tagging curation workflow to correct false positives or add missing tags. If you have already achieved a good tagging quality in your system, the tagging workflow will typically make extensions to your knowledge graph.

On the basis of the tagged content, the search function in individual systems can be improved in the first step. Based on the knowledge graph, semantic search functions such as facets, search suggestions, cross-language and synonym search are automatically available. Furthermore, the knowledge graph can be used as a search assistant and, similar to the Google Knowledge Graph,[26] provides additional context for the current search result. In this way, it can be used to derive new search paths and to explore a topic.




*PoolParty PowerTagging solution for SharePoint*

Since all information in your systems is tagged and thus linked to the knowledge graph, each digital asset is given a semantic footprint, which is itself a knowledge graph in a smaller form. This enables a precise and sophisticated semantic mapping by allowing similar content to be displayed or by recommending relevant contextual information.




*Semantic Footprint of a Document*

But all of this centers on one system without connecting different resources. Therefore, tag events should be stored in a central, superordinate index or graph for all connected systems (e.g., in a graph database or in a semantic data catalog), and not within their respective silos.

The idea of tagging content and documents using knowledge graphs can of course be applied to all other business objects. Thus, products, projects, suppliers, partners, employees, etc. can be semantically annotated just as

well. As a result, we can establish relationships between different types of business objects and use recommender systems to push, e.g., personally relevant content to the employee's desktop, to link potential suppliers to upcoming projects and thus facilitate the selection process, or to combine products with products to facilitate the customer's purchasing decisions. In other words: things that fit together finally come together more easily along workflows.

This method finally allows a cross-system search for content, people, projects, etc. to fulfill one of the long cherished dreams of knowledge management. Sounds like a pretty good recipe, right?

## Unify unstructured and structured data in a Semantic Data Catalog

The Semantic Data Catalog approach combines the respective advantages of data lakes and data warehouses and complements them especially with active metadata and advanced linking methods that semantic graph technologies bring with them.

> ***"Implementing data catalogs without a strategic plan to link them to broader metadata management needs will lead to metadata silos, making them difficult to effectively manage and integrate in the longer term."***
>
> (GARTNER, INC: 'AUGMENTED DATA CATALOGS: NOW AN ENTERPRISE MUST-HAVE FOR DATA AND ANALYTICS LEADERS', EHTISHAM ZAIDI AND GUIDO DE SIMONI, SEPTEMBER 2019)

The ultimate goal is to unify unstructured, semi-structured, and structured data to make all of it available as if it were the same database. To make this possible, it's necessary to introduce a semantic knowledge graph that describes the meaning of all business objects and topics (and their interrelationships) that can be found in all these data sources. The key to success with this strategy is to look at the importance of metadata more carefully.

In all cases, the metadata should be as self-explanatory as possible. The most obvious strategy for achieving all of these objectives within a framework for managing enterprise data is to establish a central hub as a reference point. This should map all the different metadata systems, describing and making their meanings available in a standards-based modelling language. As we will see, a semantic data catalogue can meet all of these requirements.

This approach emphasizes the importance of the semantic layer as a common umbrella for all types of data. Semantics will no longer be buried in data silos, but linked to the metadata of the underlying data. It helps to "harmonize" different data and metadata with different vocabularies. It makes the semantics (meaning) of metadata and of data in general explicitly available.

Implementing a semantic data fabric or data catalog solution also means that new opportunities for modern enterprise data management will arise. These are manifold:

- Find, integrate, catalog and share all forms of metadata based on semantic data models.
- Make use of text mining: deal with structured and unstructured data simultaneously.
- Graph technologies: perform analytics over linked metadata in a knowledge graph.
- Use machine learning for (semi-) automated data integration use cases.
- Automate data orchestration based on a strong data integration backbone.

In the chapter on semantic data catalogs, we present an example of a system architecture that shows the working methods and components of such a system in more detail. An essential component of this architecture is the 'Semantic Middleware' such as the PoolParty Semantic Suite,[27] which is embedded in a modern data catalog system (such as data.world[28]) and enhances the semantic layer, especially the graph-based text mining capabilities.

## Connecting the dots: Search and Analytics with Knowledge Graphs

### *"GRAPH QUERY LANGUAGES LIKE SPARQL ARE AHEAD OF THE GAME"*

IDC predicts[29] that our global data sphere will grow from about 40 zettabytes in 2019 to an astronomical 175 zettabytes[30] in 2025. We know that you have read such sentences before, but wait, has your organization really started to react to it appropriately and has it ever looked in a fundamentally different direction from the one that traditional answers would cover, such as “Well, let's add a data lake to our data warehouse and with that we’ve certainly covered all our data analysis needs.”

Just keep on counting doesn't work. As our brain develops, we do not just accumulate new neurons, but rather we *link* them together. A human brain has an estimated $10^{15}$ synaptic connections based on ‘only’ 80-90 billion neurons.

Efficient search and analysis in enormous information and data spaces first requires the ability to quickly limit search spaces to those areas where the probability of finding solutions is high. This requires large-scale networked structures in which supposedly large distances can be quickly bridged in order to follow up with a detailed examination of the limited search spaces in a second step.

Both steps are supported by graphs and thus offer both precision and recall at the same time: first, to quickly break down the analysis or search to only relevant documents, data sets or individual nodes of a graph (high recall), and then to also provide less qualified data users with tools that enable them to perform advanced queries and analyses (high precision) by refining, faceting, and filtering guided by semantic knowledge models that are also part of the knowledge graph.

Graph query languages like SPARQL are ahead of the game: “Unlike (other) NoSQL models, specialised 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."[31] Put simply, knowledge graphs and corresponding query languages can be used to identify entities, facts and even more complex relationships based on path finding algorithms, node similarity metrics, etc.

In summary, while conventional technologies like data lakes or data warehouses support either high recall (large data) or high precision (filtered data), graph-based data management with data fabrics combines both into one formula: precision and recall (F1 score). We will now look at some more concrete application scenarios that revolve around this topic, perhaps generating new ideas for your workspace.

### *Semantic Search*

When asked how users can benefit from knowledge graphs, a common answer is "by better search." Some also call "semantic search" the low-hanging fruit of efforts to create a knowledge graph in a company.

Semantic search has been around for many years. In the early years, it was based purely on statistical methods and helped users to refine their search results using automatic clustering techniques. The quality of the clustering results was regularly below the threshold value that is still useful for end users due to the great heterogeneity and the relatively small document volumes typical for a company search scenario. A still unsolved problem of NLP is the meaningful automatic labeling of the resulting clusters,[32] which is of course, important for the user experience.

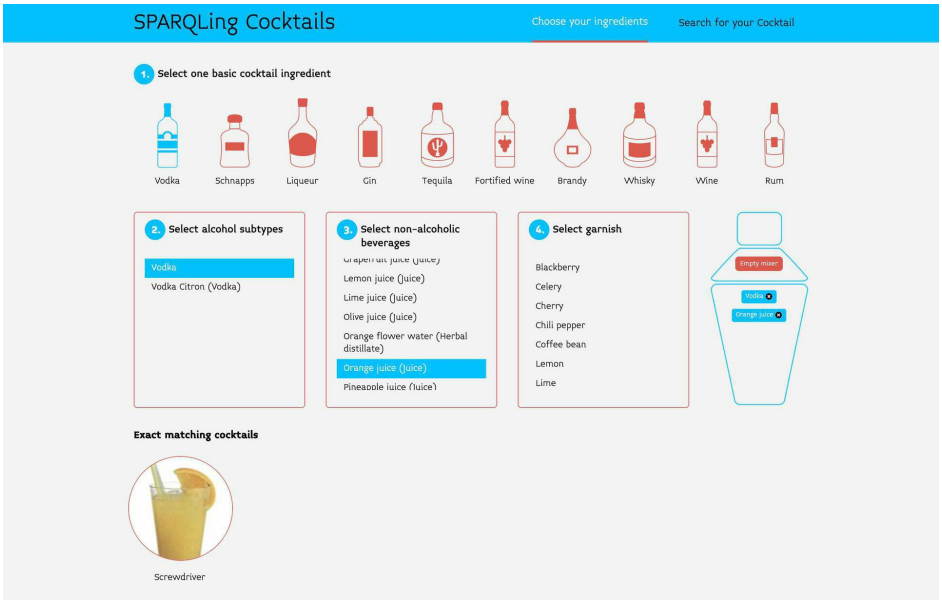
What exactly is meant by "semantic search" can by no means be clearly defined, since the range of such search solutions is enormous. Nevertheless, they all have a common goal: to create a search application that

1. understands the intent of the user in a way that is close to human understanding,
2. links all relevant information (e.g., documents or text passages) to this search intent, and
3. delivers results that are as understandable as possible and well prepared for further user interaction such as faceted navigation.[33]

This approach often includes the ability to understand more sophisticated search queries than the simple keyword search. For example, you can enter such queries into a search form or ask: "Show me all cocktails made with Bacardi, Coke, and citrus fruits", and still *Cuba Libre* will be found as a result, even if the recipe literally says something else: "Put 12cl Cola, 5cl white rum and 1cl fresh lime juice into a highball glass filled with ice."

Semantic search engines, chat bots, intelligent help desks and most solutions related to conversational AI are currently converging rapidly. Search applications based solely on simple input forms and full text indexes have become rare even in enterprise environments. The 'semantic magic' happens either in one or more of these steps or components:

1. on the user side, when the frontend benefits from enhanced Natural Language Understanding (NLU) technologies, or
2. at processing time, e.g., when a semantic index or graph is generated that contains not only terms, but also concepts and their relations, or
3. at output time, when the user benefits from an intelligent and interactive arrangement of search results, e.g., in the form of a graph or of some more domain-specific search and knowledge discovery interface.[34]




*Example for an intelligent search visualization*

### *Drug discovery*

Even in processes such as reverse pharmacology or targeted drug development, it takes, on average, more than 10 years and costs several billion US dollars to develop a new drug.[35] Drug discovery involves various scientific disciplines, including pharmacology, chemistry and biology. Each of them generates large amounts of data, which are often not interconnected. The amount of genomic, molecular and other biomedical data describing diseases and drugs continues to grow exponentially. The reasons for such relatively long periods of time are manifold and sound all too familiar, as they are by no means industry-specific:

- It is difficult to collect and integrate biological data (highly fragmented and also semantically redundant or ambiguous).
- You need to link structured data records with data that has little to no structure.
- There are no automated means for in-depth analysis.

Thus, the actual process of data integration and the subsequent maintenance of knowledge therefore requires a considerable amount of time and effort. Semantic knowledge graphs can help in all those phases of the data life cycle: they provide means for data integration and harmonization, and they use automated inference mechanisms, e.g., to deduce that all proteins that fall into a pathway leading to a disease can be identified as targets for drugs.[36]

### *Fraud detection*

In fraud detection, financial institutions try to interrelate data from various sources including locations over time (geospatial and temporal analysis), previous transactions, social networks, etc. in order to identify inconsistencies, patterns, and take appropriate action.

Deep Text Analytics based on knowledge graphs enables more comprehensive detection of suspicious patterns, e.g., it helps to precisely disambiguate locations or persons. In particular, inferencing mechanisms based on ontologies are essential to uncover relationships undiscoverable by traditional name/place matching algorithms.

### *Digital Twins and Web of Things*

The Web of Things (WoT), considered as a graph, can become the basis of a comprehensive model of physical environments that captures relevant aspects of their intertwined structural, spatial, and behavioral dependencies. As such, it can support the context-rich delivery of data for network-based monitoring and control of these environments, and extend them to cyber-physical systems (CPS).

An example for an application in this area is the Graph of Things (GoT) live explorer,[37] which makes a knowledge graph for connected things navigable. GoT provides not only sensor data, but also the understanding of the world around physical things, e.g., the meaning of sensor readings, and sensing context and real world relationships among things, facts and events. GoT serves as a graph-based search engine for the Internet of Things.[38] This approach is particularly interesting for Smart City initiatives.

On a smaller scale, especially for industrial production lines, the digital twin models have evolved into clones of physical systems that can be used for in-depth analyses, sometimes in near real-time. Industrial production lines usually have several sensors to generate status information for production. The resulting industrial ‘Web of Things’ data sets are difficult to analyze so that valuable information can be derived such as sources of failures, estimated operating costs, etc. Knowledge graphs as digital twin models based on sensor data are a promising approach to improve the management of manufacturing processes by inference mechanisms and the introduction of semantic query techniques.[39]

A good starting point to explore ways to build knowledge graphs for this purpose is the Semantic Sensor Network Ontology,[40] which has been a W3C recommendation since October 2017.

The idea of using a 'digital twin' to improve the quality of decision-making and predictions originally comes from industrial production. Enhancing the underlying models with the help of knowledge graphs is not only obvious, but also has the potential for transference to other economic sectors. Approaches where knowledge graphs are developed as 'digital twins' are therefore increasingly common.[41]

## Deep Text Analytics (DTA)

*"TRADITIONAL TEXT ANALYTICS HAS ONE MAJOR WEAKNESS: ITS METHODS DO NOT BUILD ON BROAD KNOWLEDGE BASES AND CANNOT INCORPORATE DOMAIN KNOWLEDGE MODELS AS EASILY"*

Gartner predicts that "by 2024, companies using graphs and semantic approaches for natural language technology projects will have 75% less AI technical debt than those that don't."[42]

Traditional text analytics has one major weakness: its methods do not build on broad knowledge bases and cannot incorporate domain knowledge models as easily. They instead rely on statistical models while even more advanced technologies such as word embedding are not yet able to understand the larger context of a given text accurately enough.[43]

Another disadvantage of this approach is that the resulting and often more structured data objects are not based on a standard and cannot be easily processed together with other data streams, i.e., to be linked and matched with other data.

In contrast, Deep Text Analytics (DTA) makes heavy use of knowledge graphs and semantic standards and is therefore able to process the context of the text being analyzed, which can then be embedded in an even broader context. It is a very advanced methodology for automated text understanding, based on a number of technologies that are being fused together: NLP techniques such as

- text structure analysis,
- extraction of entities from text based on knowledge graphs,
- extraction of terms and phrases based on text corpus statistics,
- stemming or lemmatization;
- recognition of named entities and text classification based on machine learning enhanced by semantic knowledge models;
- optionally also the extraction of facts from text; and finally,

- the automated sense extraction of whole sentences, which is based on the extraction of data and entities and validation against a set of conditions using knowledge graphs.

To summarize, here is a list of the advantages that DTA offers compared to traditional text analysis methods:

- Instead of developing unique semantic knowledge models per application, DTA relies on a knowledge graph infrastructure, and thus on more reliable and shared resources to efficiently develop Semantic AI applications embedded in specific contexts.
- It merges several disciplines like computer linguistics and semantic knowledge modelling to help computers understand human communication (e.g., to create fully functional chatbots).
- Human communication generates a large amount of unstructured data mostly hidden in textual form. Deep Text Analytics helps to resolve the ambiguity of unstructured data and makes it processable by machines.
- It performs extraction and analysis tasks more precisely and transforms natural language into useful data.
- The technology is used for more precise intent recognition of human communication in the context of so-called natural language understanding (NLU). The basis for this is automatic sense extraction and classification of larger text units, e.g., entire sentences.
- Deep Text Analytics is text mining based on prior knowledge, i.e., on additional context information. This increases the precision in extracting relevant data points from unstructured content.

In the following subchapters we describe three concrete application examples based on DTA.

### *Contract Intelligence*

Contracts are often difficult to administrate and are filed and forgotten until a problem arises. The reason for this is that the manual management of contracts, including the creation of new agreements and tracking the expiration of contracts, is very time-consuming and person-dependent. Existing contracts can also often contain risks that are difficult to detect using

manual methods.

There are quite a few applications out there labeled as providing contract intelligence solutions and aiming to give better access and control over legal contracts by making them interpretable and searchable in an intelligent way. This is a perfect use case for making use of knowledge graphs supporting DTA to make the information within large volumes of contracts easier to find and access.

The first step in this process is to make contracts more accessible by arranging them into a meaningful structure. Most contracts are only available in unstructured formats like MS Word or PDF. In the first step, this unstructured information can be brought into a generic structure like XML or RDF based on the document structure (headings, paragraphs, lists, tables, sentences). Based on this, an initial semantic analysis can be conducted using the knowledge graph to determine which sections of the contract should be further analyzed by entity extraction, categorization, and classification. In this step, the generic structure is then converted into a semantically meaningful structure.

Now that you know exactly which parts of the contract relate to which subjects (e.g., confidentiality, guarantees, financial conditions, etc.), an in-depth analysis of the specific subjects can be carried out, applying rules that are in line with the conditions, through tests defined on the basis of the knowledge graph. This provides you with greater insight into your contracts and allows you to check the compliance of contracts along your own guidelines using the automated sense extraction of entire sentences.

### *Automated understanding of technical documentation*

Technical documentation is usually very structured and quite often very difficult to access. "Read the manual!" Why should I do that? I usually can't find anything and don't want to search all the documentation for the one little thing I’m looking for. Do we have to keep it that way?

Because technical documentation is highly structured, it is a perfect use case for applying deep text analysis to significantly improve the user experience. In the context of documentation, it is not only important to find the right

place, but also the right kind of information. Do I want step-by-step instructions for a specific topic or do I prefer all warnings related to a functionality?

There are XML standards like DITA[44], which are often used in technical documentation. These can be used as a basis for a corresponding ontology. The content of the documentation also provides an excellent basis for creating a taxonomy of all addressed topics, components, roles, problems, etc.

Utilizing the automatic tagging and extraction of named entities allows for content to be better filtered, found, and linked. Combining different types of documentation such as manuals, tutorials, FAQs with the same knowledge graph allows the right information from different sources to be linked and displayed as a whole, and also to recommend related content, e.g., the part of a manual that matches a question in the FAQs.

The problems around the current versioning of manuals and resulting inconsistencies can also be addressed with the help of a knowledge graph. More advanced scenarios, using Q&A systems or chatbots as the best possible access to technical documentation for example, can be realized on the basis of a well-structured knowledge graph.

***Intelligent Robotic Process Automation***

*"THE NEXT GENERATION OF RPA PLATFORMS IS JUST AROUND THE CORNER, AND THEY WILL CONTAIN MUCH MORE AI THAN THEIR PREDECESSORS"*

With the introduction of robotic process automation (RPA), organizations are striving to use a noninvasive integration technology to eliminate tedious tasks so that the company's employees can concentrate on higher-value work. However, RPA rarely uses any AI or ML techniques, but rather consolidates a large number of rule-based business process automation and batch jobs to organize them in a more intelligent way.

The next generation of RPA platforms is just around the corner, and they will contain much more AI than their predecessors, and much of it will be based on Deep Text Analytics. Thus, RPA seems to be only a stopgap en route to intelligent automation (IA), which eventually automates higher-order tasks that previously required the perceptual and judgment capabilities of humans, for example:

- On-boarding processes (new customers or employees)
- Complaint and claims handling
- Risk analysis (e.g., financial reports)
- Optimization of helpdesk
- Monitoring and verification of compliance
- Due diligence processes

## Excellent Customer Experience

What is it that makes an outstanding customer experience? The aim is always to ensure that, throughout the entire customer journey, the customer always has access to the information that will enable him or her to optimize his or her purchase decisions (including possible improvements after the first transaction), the operation of the product or possible bug fixes. It is also about minimizing the resources used, especially the time spent for both the customer and seller. Personalization techniques play a major role in this process.

You will see that you can achieve an improved customer experience around your offerings if the knowledge graph is integrated into your support processes and content management workflows. Last but not least, you and your users will benefit from semantic technologies by gaining more knowledge about clients from structured and unstructured data, as described in the previous section, thus continuously increasing customer satisfaction.

### *Customer 360*

*"A CUSTOMER KNOWLEDGE GRAPH OFFERS THE POSSIBILITY TO CREATE SUCH A UNIFORM VIEW OF ALL CUSTOMER INTERACTIONS AND*

## *RELATIONSHIPS"*

Marketing campaign and automation managers are tasked with finding out what draws people to a website. Regardless of how well networked the data for this already is, with or without graph technology, it involves analyzing data from various sources, Twitter, e-mail, Google Ads, etc. The aim is to obtain the most complete picture of the users possible, and this is referred to as "Customer 360."

The other side of this user-*centered* view of the analysts is a radically user-*oriented* view of all content and offerings. The more complete the customer model that is available to a provider, e.g., for personalizing offers, the more the customer feels "in good hands" and the better the quality of service will be. A customer knowledge graph offers the possibility to create such a uniform view of all customer interactions and relationships. This contextual 360° view of the customer, in which all his activities can be aggregated across the entire spectrum, can also reveal previously hidden relationships between people, content, and products.

An example of a graph that provides holistic views of users/customers on both sides of the system, i.e., from the perspective of the end-user as well as from the perspective of the analyst/operator, is the Economic Graph,[45] which, as a central element of the LinkedIn platform, enables some essential services:

- connect people to economic opportunities
- spot trends like talent migration or hiring rates
- identify ‘in-demand’ tech skills for organizations or whole regions

### ***Recommender Systems***

When you connect your business objects to the knowledge graph, each of them receives a semantic footprint. This footprint can be generated automatically and regardless of the type of object, it can in many cases be used to describe end-users in greater detail. This is done, for example, with the help of the documents they have created or tagged, the business objects or products they are interested in, their resume, etc.

The comparison of the semantic footprints of the individual objects in a knowledge graph makes it possible to point out similarities and also to recommend complementary things, e.g., a follow-up product in which a person might be interested based on the products already purchased and the resulting footprint.

One example of a graph-based recommendation system you can test out online is a wine-cheese recommendation system[46] that is able to select complementary products to a specific wine or cheese. The system is based on a domain-specific knowledge graph and is also able to derive semantic footprints of each new product using text mining based on the graph.


*Recommender system for HR*

PoolParty HR Recommender,[47] a demo application of a semantic recommendation system for another area, namely for the human resources department, shows how a comprehensive taxonomy like ESCO[48] can be used as the basis of a knowledge graph to automatically link people in companies with knowledge assets such as projects, open positions or other experts.

### *Conversational AI*

Despite some disappointments after the initial hype, chatbots and conversational AI are still on the rise. However, the underlying system architecture is still evolving. Gartner explains that “by 2022, 20% of all new chatbot and virtual assistant implementations will be done on conversational AI middleware that supports multiple NLP back ends, up from less than 5% today.”[49] This means that there is no longer just a monolithic system running the chatbot as a whole, but rather a 3-tier architecture embedded in a larger AI infrastructure.

As part of this architecture and in view of the need for AI middleware, reusable knowledge graphs serve as a basis for advanced NLU and NLP capabilities. They help to identify the intent of requests and interactions by extracting terms and entities that are placed in a larger semantic context. Early on, this primarily helps to provide more accurate answers. In addition, this approach is more transparent to subject matter experts and helps them to improve the flow of dialogue while ensuring compliance with laws and regulations.

### *Search Engine Optimization (SEO)*

One of the main goals of any online marketing department is to optimize the content of a website in order to achieve the best possible ranking on the search engine result pages (SERP). There are many strategies to achieve this goal, but an important one is to feed search engines like Google and its crawlers with information that is available in a machine-processable form.

Once this is in place, Google can display the crawled information as featured snippets, PAA boxes (‘people also ask’), as answers to ‘how-to’ queries, or as knowledge panels.[50] The semantic metadata, which is typically embedded as JSON-LD into HTML, can even be used as an input for virtual assistants like Google Assistant. All of that increases visibility on (Google's) search platforms, which in turn increases customer satisfaction.

For search engine optimization (SEO), the concepts used in an online article should be classified and marked up with Schema.org and be linkable to knowledge graphs such as DBpedia, Wikidata or the Google Knowledge Graph. In this way, search engines are informed about why and when a certain content may be relevant for a certain search intent.

Let’s assume you have just published an article about “How to cook a Wiener Schnitzel,” as for example can be found on *The Spruce Eats*,[51] and now you want to boost your visibility on the web. A step-by-step guide[52] that describes how you can enrich this article with semantic metadata to be highly ranked can be found on Google Search web developer's guide.

The use of Semantic Web technologies within an SEO context initially appears to pursue other goals than e.g., semantic search in enterprises, but search engines like Google are also constantly striving to improve the user experience and search results. In this respect, the SEO strategies of online marketing professionals are increasingly similar to methods for optimizing enterprise search. In both cases, the heart of the problem is networked and high-quality content, consisting of entities (instead of words) that are linked in the background via knowledge graphs.

# PART 2: SETTING THE STAGE

## PREPPING THE KITCHEN

# Introducing Knowledge Graphs into Organizations

**A little soup is quickly cooked. But if many chefs are working together on a larger menu that will eventually be appreciated by a banquet of guests, good preparation is key. In this chapter we will focus on the preparation phase and outline what it means to introduce knowledge graphs in a company firsthand and from an organizational perspective:**

- **When do you know that you need a knowledge graph?**
- **Assessing the semantic maturity level of an organization**
- **Overcome segregation and specialization**
- **The importance of involving the right stakeholders**
- **Why develop a knowledge graph in an agile way?**

*"IT IS NOT ENOUGH TO PURCHASE BLACK BOX AI OR SIMPLY HIRE TEN DATA SCIENTISTS OR DATA ENGINEERS TO CREATE A KNOWLEDGE GRAPH"*

The introduction of knowledge graphs into organizations is not necessarily comparable to the introduction of any new technology. It is not just a question of which graph database to use and which tools to use for managing the knowledge graphs. The best equipped kitchen will not cook the best food by itself. Of course, it is important to choose the best technology and equipment, but this is best done following one's own experience.

The introduction of knowledge graphs is a data management initiative that requires appropriate change management as scaling increases. This means that it must start with the careful planning of goals and strategies. It requires a change in the way of thinking when dealing with data. It requires learning new standards, methodologies and technologies by your technical teams. It requires new skills for the people working on these projects. It is not enough to purchase black box AI or simply hire ten data scientists or data engineers to create a knowledge graph. If the knowledge graph is to become a strategic asset in your organization, then you need to treat it as such.

## When do you know that you need a Knowledge Graph?

## *"AT THE END OF THE DAY, KNOWLEDGE GRAPHS DON'T JUST LINK DATA, THEY ALSO LINK PEOPLE"*

This question may sound strange, but you should ask yourself before you start. Because a successful implementation of an Enterprise Knowledge Graph is a course-setting for the future. That's not to say that like in the classic waterfall model you have to plan the implementation meticulously before you start. On the contrary. But it should at least be clear whether a knowledge graph is the right way to solve existing problems.

If one or more of the following aspects sound familiar, you are on the right track:

- You often face the problem of having to translate or rephrase your questions
    - across languages, because you work in an international environment,
    - across domains, as your departments have different views on things,
    - across organizations, as your partners have their own language, and
    - because the language has changed and things today are named differently than two years ago.
- You often want to get information out of your systems but you do not succeed because
    - there are so many systems but they do not talk to each other,
    - they all have different data models and you need help to translate between them,
    - you need experts to help wrangle the answers out of your systems, and
    - your experts tell you this is not possible because of the relational data model in place.
- You often can't identify the right person or expert in your company, so you have to start from scratch.

- After you have completed a project or work, you have often found that something similar already existed. You have often had the feeling that you have reinvented the wheel.
- You always use Google instead of internal tools to find things.

Now might be the right time to think about how to change and develop the organizational culture in terms of access to information and work with information or the development of knowledge. But when people should “go where no one has ever been before”, they also need to be prepared and open-minded. At the end of the day, knowledge graphs don’t just link data, they also link people.

## Assessing the Semantic Maturity Level of an Organization

The next step is to take a look at your organization and see how well it is prepared for this change. When assessing the maturity level of your organization, you should again consider two aspects:

### *Organizational Aspects*

Knowledge graphs are traditionally based on the Open-World assumption, which implies that knowledge graphs are never complete. This seems to be a strong contrast to the reality of many organizations and the way they do projects. So if you might find yourself characterizing your organization as "a highly specialized, relatively old industry" that "deals with complex, costly, and potentially hazardous facilities and processes," you may find it difficult to introduce knowledge graphs and convince people why they should spend time on such an adventure.

If, on the other hand, you characterize your organization in such a way that "we are open to new things and like to learn and explore" and "we deal with complex information and processes are important, but we have also learned to change and adapt when necessary," then it will most likely not be difficult for you to spark the interest of your teams.

Specialization normally also means segregation into knowledge silos and interfaces to translate between them. A knowledge graph approach means to establish a unified translation layer on top of those silos, so they speak to

each other in a common language that is easy to understand and explore. But what happens if people are not used to, or trained, or open to explore and "talk to each other in a common language"? They will not understand or use those systems. Therefore, of course the necessary skills must be built up and simple applications that improve everyone's working life must be made available as quickly as possible to convince people. In addition, a change in mindset and culture is also required to ensure that employees become accustomed to the following principles:

- Systems are easily extendable.
- Systems can be linked and connected.
- Systems allow us to think/explore beyond silos.

In addition to defining the technologies for linking systems and merging data, the corresponding processes must also be established. After all, you want your people to soon be able to cook from memory without a cookbook.

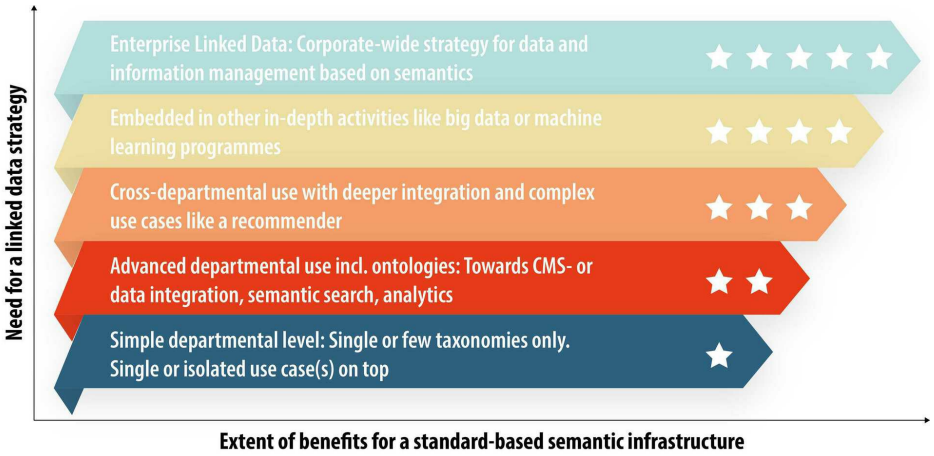
### *Technical Aspects*

Building an enterprise knowledge graph is an agile thing. It’s alive you have to grow and mature it, and you have to feed it well so it becomes strong and healthy. So it is not a typical project you plan, implement, and then you are done. Rather, we strongly encourage you to develop it in an agile way:

- Starting small and growing continuously based on examples and use cases.
- Trying to show benefit as early as possible.
- Learning from successes and failures and establishing the necessary know-how and skills along the way.

An enterprise knowledge graph cannot be implemented without support throughout the whole organization. Also SysOps, security and infrastructure have to embrace the change. This is a potential problem because exactly those departments are frequently enemies of change as they have to guarantee stability and continuity of operations. In addition to changes within the organization, new roles/personas with new skills and knowledge should be introduced as well in order to support this transition. In the next chapter we will outline different personas you will typically need to set the stage.

The enterprise semantics maturity model below clearly outlines that the need for a linked data and knowledge graph strategy becomes more evident as your knowledge graph infrastructure matures.




*Enterprise Semantics Maturity Model*

So the success of a knowledge graph initiative strongly depends on establishing the right methodologies and getting important stakeholders on board, i.e.,

- start simple and grow,
- develop your knowledge graph in an agile way,
- build up the necessary skills and roles, and
- understand that it is not a replacement, but an extension.

## Embedding Knowledge Graph Building in a Change Management Processes

As explained in the previous section, the introduction of a knowledge graph is not (only) a technical implementation of a new technology. It must include both strategic and organizational aspects to be successful. It is also a change management initiative because it changes the way your organization works with and values data. In a sense, you could say that data has always been hidden in a complex infrastructure and technology. It wasn't about data, it was about the technology to store data securely so that in the end, no one knows it exists.

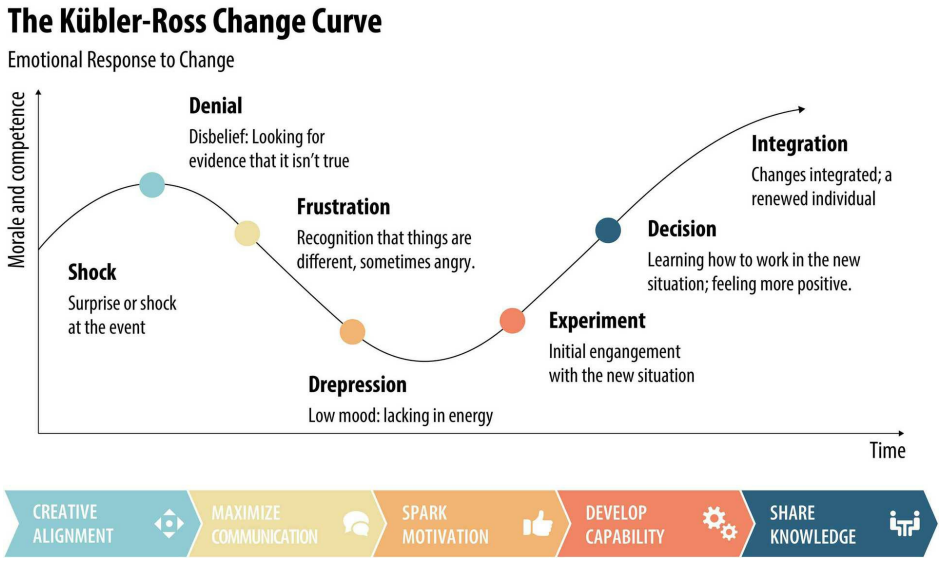
We remember very well a DMS conference years ago, where a provider used

a safe as a symbol for his security level. We understand that security is important for all companies, but there is also another interpretation: "keep your data safe and forget about it, that way everything will be so complicated that nobody would dare to ask if they can do this or that with the data." Well, the answer is in the middle, and we should think about how we want to handle our data. So make your data a first-class citizen that you can explore through an enterprise knowledge graph, and let technology be the means, not the driver.

As with any change management initiative, you will need to deal with the different phases of the emotional response to change. Realizing that you are in a situation where you have locked your data away for years and when you want to use it, you can no longer access it in a meaningful way, will lead to shock and denial. Here's what you're going to hear a lot:

- “But this is how we've always done it!”
- “But we can do all of this already!”
- “We can do all of that with [XML|Lucene|SharePoint|...] anyway, so what is new?”
- “Isn’t that what our Big Data initiative is for?”

So you will have to find creative ways to show the value and benefit and convince people to bring them on board.




*The Kübler-Ross Change Curve*

There will also be frustration and depression when people realize that they have to move and change. Their comfort zone is at stake and change will come. The most important and helpful argument we found in those situations is: “Look, you do not have to throw away anything. You just put the knowledge graph on top of your existing data infrastructure to make it connected and accessible.” That already relaxes most involved stakeholders a bit. If they are involved from the beginning, well informed and also motivated because they soon experience the value of the initiative, they can eventually be convinced to join forces.

Now you are in a critical phase, as you may want to try to make the big change and plan it for the next 20 years. Don't do that! Experiment in order to make valid decisions based on experience. Learn that experiments are not bad things or even a sign of immaturity, but rather the only chances to learn, to become better, to improve continuously and to develop skills. If all this smells of agility, then it is. "Agile" is everywhere these days. We know that, but we also need agile access to data to make better use of it, so we need agile data management. A knowledge graph project must always be an agile data management project. Knowledge is a living thing that is constantly changing.

***“Change is the only constant in life.***

—HERACLITUS OF EPHESUS

So if you've done this part right and haven't forgotten to keep on experimenting, you can start to integrate your findings into your existing productive data landscape and enjoy the change that comes with it. When people realize that they are no longer slaves to data locked up in systems, but masters of a unified data landscape that can create a lot of knowledge in the right hands, they will become more productive and they will begin to think in new directions and discover things that were not possible before.

## Knowledge Graph Governance

Knowledge graphs are not “just another database,” they rather serve as a vehicle to rethink and rework the existing data governance model while a governance model for the KG management itself has to be developed at the

same time. Here are some key questions that help to form the basis for a KG governance model:

- Which parts of the graph have to be managed centrally, which are more driven by collaborative and decentralized processes?
- How can all the different requirements be met, including the different notions of what a high-quality knowledge graph actually is?
- Which parts of the KG can be generated automatically without affecting the defined quality criteria, and which elements have to be curated by humans?
- What kind of data produced by the users, e.g., their navigation behaviour, can be used for further processing in the knowledge graph? Or, as another possible part of the user loop, could users be involved in crowd sourcing activities, e.g., to tag content elements or data sets?
- Which data elements, e.g., structured data sets or already existing taxonomies could potentially be included in the emerging KG and who can determine this?
- Which already existing data governance models, e.g., for taxonomy governance should be embedded in the overall KG governance model?

Ultimately, the development of knowledge graphs as an agile approach to managing significant portions of the overall data landscape implies the need to extend the existing data governance framework. Each graph project triggers changes on different levels of an organization and its information and data architecture, here are a few examples:

- New roles, their interplay and their responsibilities have to be defined.
- Content and data authoring/curation processes will be extended and partially automated.
- Diversification of access points to data and knowledge have a direct impact on the existing data governance model.
- New ways to gain insights into enterprise data will be developed, e.g., automated generation of links between data points which were not initially connected.
- In return, these new insights trigger new questions related to GDPR

compliance.

- Algorithms that automatically generate personalized views on content and data enhance the customer experience.
- Customers then become more active data producers who generate data for further processing in the knowledge graph.
- The use of linked data principles and a more standards-based approach to data management in general opens up the possibility of making greater use of open data sources, which in turn increases the need for a more stringent data quality management system.
- New ways to filter and contextualize data objects will be available "as a service.
- As new technologies get implemented, new and diversified perspectives on data quality and compliance make the necessity to establish a Data Governance Board even more obvious.

# Personas: too many cooks?

**Any data and AI program including knowledge graphs as a cornerstone also includes a number of projects that in turn require the participation of various stakeholders. So what are the best practices for developing semantic AI and the underlying knowledge graphs? To better understand this, we should first look at the people involved, their typical responsibilities and tasks, and their particular (potential) interest in knowledge graphs.**

- **How do you put together a working team to roll out an enterprise knowledge graph?**
- **Which stakeholders are involved and what are their interests?**
- **How can they be inspired to support a KG project?**

In recent years, companies have carried out numerous Proof of Concepts (PoC) to develop the appropriate recipe for setting up AI systems. Depending on who sponsored these pre-production projects, either predominantly bottom-up or more top-down approaches were chosen to roll out the topic. Many of these PoCs also had a strong bias towards one of the three loops of the knowledge graph lifecycle, rather than allowing the three areas to interact and be considered equally. In any case, our experience with all these relatively one-sided approaches is mixed. The best chances of success in terms of an efficient learning curve are when the topic is approached from several perspectives, since ultimately a collaborative and agile environment must be practiced and rolled out.

In this chapter we describe how the potential interest of individual interest groups in knowledge graphs could be described or awakened. We design a line of argumentation for each roll, and in order to specifically address the decision makers, we also outline "elevator pitches." All this helps to quickly reach the point where an informed discussion can take place with anyone who might be involved in a subsequent KG project.

For example, a precise and detailed view of the roles involved will also help to define appropriate skills and tasks to bridge mental differences between departments that focus on data-driven practices on the one hand, and

documents and knowledge-based work on the other. Similarly, we will also address the question of how subject matter experts with strong domain knowledge (and possibly little technical understanding) can work together with data engineers who are able to use heavily ontology-driven approaches to automate data processes as efficiently as possible.

Also, involving business users and 'citizen data scientists' as soon as possible is essential, since users will become an integral part of the continuous knowledge graph development process, nurturing the graph with change requests and suggestions for improvement.

### *Chief Information Officer (CIO)*

Among many other responsibilities (e.g., information security), CIOs want to develop the right organizational model to achieve better results from their AI initiatives. CIOs develop teams to implement their AI strategy with the awareness that AI is a much broader discipline than just ML, e.g., knowledge representation, rule-based systems, fuzzy logic or Natural Language Processing (NLP).

- ***Why KGs?***
  KGs form a robust backbone for every AI and analytics platform by establishing a common semantic foundation for your enterprise architecture. They help to provide high-quality data based on enriched and linked metadata for ML, involving different people and roles from the AI team and lines of business. KGs are also essential for any explainable AI strategy.
- ***What is a KG?***
  A knowledge graph provides linked data containing all business objects and their relationships to each other. To create the knowledge graph, all possible databases of a company are typically linked and stored in a graph database where they are then enriched with additional knowledge. Text documents can also be docked to the knowledge graphs with the help of NLP. This creates 360-degree views of all relevant business objects in the company.
- ***How to apply KGs?***
  The use of KGs can have enormous effects on various systems and processes. Active metadata, as formulated by Gartner as “...a key to

more efficient use of enterprise data," can be managed with KGs. And NLP also benefits enormously, where much more sophisticated text analysis methods can be used when MLs are combined with KGs.

- ***What if?***
  If your company already had a full-blown EKG available, then the interaction of all important stakeholders within your AI team would have matured a bit more. KGs also serve as a central reference point in a company where all business objects and their semantics are managed. This is made possible by a high degree of collaboration and thus allows a more agile handling of data even along complex compliance regulations.

### *Chief Data Officer (CDO) / Data & Analytics Leaders*

A CDO as the leader of the data and analytics team wants to create business value with data assets. "Enhance data quality, reliability and access," "Enhance analytical decision making" and "Drive business or product innovation" are the top three business expectations for the data and analytics team in Gartner's most-recent CDO study.[53] CDOs take more and more responsibilities from CIOs, as evidenced by the transfer of ownership of metadata, for example, which we are currently seeing in many companies.

- ***Why KGs?***
  Without having to radically change existing data landscapes and infrastructures, knowledge graphs, as non-disruptive technologies, form the basis for significantly enhancing the value of data for several reasons: metadata from different sources can be harmonized and enriched, structured and unstructured data can be dynamically and cost-effectively networked and better analyzed, cross-silo tests for data quality can be automated reasonably, and NLP technologies based on knowledge graphs become more precise.
- ***What is a KG?***
  The knowledge graph is a virtual layer on top of the existing metadata and data. Since it describes business objects, topics, and their interrelationships in such a way that machines can also access them, it greatly supports numerous ML and NLP technologies. To

guarantee high data quality, smaller parts of the knowledge graph have to be created and curated by experts, but much of the creation process can be automated using ML.

- ***How to apply KGs?***
  Knowledge graphs can play a central role in any initiative to improve data quality. All repositories, from master data, to records and document management, to unstructured parts of the intranet, and thus all kinds of metadata, are harmonized and enriched with additional knowledge with the help of KGs, making them machine-readable and easier to analyze.
- ***What if?***
  What if every knowledge worker and business user in the company could create a networked view of all relevant business objects with a few mouse clicks? Knowledge graphs as an innovative method of making business data more accessible combine the advantages of data lakes and data warehouses.

***AI Architect***

*"KNOWLEDGE GRAPHS ARE A KIND OF MASTER DATA SYSTEM WITH SUPERPOWERS"*

AI architects play the central role in realizing an end-to-end ML and AI pipeline. They are the owners of the architectural strategy. They connect all relevant stakeholders to manage and scale the AI initiatives. Unlike the Enterprise Architect, who is responsible for a wide range of functions, the AI architect focuses only on the transformational architecture efforts that AI introduces. To select the right components, an AI architect must have deep knowledge of tools and technologies within the AI industry, as well as the ability to keep up with rapidly evolving trends.

- ***Why KGs?***
  Any AI strategy must, of course, focus on ensuring the accessibility, reusability, interpretability and quality of the data. With the existing infrastructure this is regularly a major challenge. Knowledge graphs can be used to address all these issues without having to make major changes to existing systems. Even better: the limits of machine

learning, traditional NLP technologies, and statistical AI in general become evident again and again. Semantic knowledge models in the form of symbolic AI can efficiently enrich and enhance data sets. Strategies that have set explainable AI as a building block can also be implemented with knowledge graphs.

- ***What is a KG?***
  First of all, knowledge graphs are data. This is data that can describe how all other data and metadata in the company can be classified and related to each other. Knowledge graphs are a kind of master data system with superpowers. They describe the meaning (semantics) of all business objects by networking and contextualizing them. In addition, they can also be used to more efficiently process the naming diversity of all the things, products, technologies, policies, etc., in an organization. In-depth text mining and the cross-linking of databases are two fields of application for enterprise knowledge graphs.
- ***How to apply KGs?***
  Data engineers and ML engineers are busy extracting and preparing data, and often data silos and data quality issues are the biggest hurdle to overcome before the "AI magic" can kick in. KGs serve as a universal access point to all your data, it's like a multidimensional index that is standards-based and machine-processable. With little preparation and in an efficient way, data sets can be extracted from the entire collection and made available as training data. Knowledge graphs also contain knowledge about specific areas of expertise that could not be found in the data in the form they are presented. This 'ontological' and 'terminological' knowledge linked to the enterprise data enables additional analyses, as well as more precise and scalable AI applications (e.g., semantic chatbots), and also enriches your data. Training data sets, even in small amounts, can be better processed by ML algorithms.
- ***What if?***
  In the semantic layer all AI and KG services of your AI architecture are developed to make data interoperable with each other and to significantly improve human-machine communication. These services should be positioned as an enterprise-wide asset and should not be developed again for each application individually. The synergies are obvious: with the knowledge graph, the 'cerebrum' of a

company is created, which can be linked to different data streams in an intelligent and dynamic way.

***Data/Information Architect***

## *"COMBINE DATA CATALOGS AND VIRTUALIZATION TO CREATE A SO-CALLED SEMANTIC DATA FABRIC"*

The Data/Information Architect is the technical leader and key strategist for aligning all technologies and architectures, as well as the underlying standards and processes for data management across the enterprise.

By balancing the interests of business and IT, he or she ensures that the data architectures are sustainable in meeting both business and IT objectives. He/she defines best practices for enterprise data management, especially for data quality management.

- ***Why KGs?***
  Data architects today find themselves in an almost hopeless dilemma. Most companies cannot afford to dismantle and replace systems built over years. The architecture is outdated because it is based on the principle "data logic follows application logic." The underlying business logic is still valid, but over time it has been ripped apart into countless data silos and the applications that access them. “The current Enterprise Information System paradigm, centered on applications, with data as second class citizens, is at the heart of most of the problems with current Enterprise Systems.”[54]
- ***What is a KG?***
  Since knowledge graphs are at the heart of a next-generation data architecture, the proposed solution to this challenge is to combine data catalogs and virtualization to create a so-called semantic data fabric. This means that the data stays where it is and is accessed via the semantic layer, with the data catalog pointing to the underlying data storage systems.
- ***How to apply KGs?***
  However, this conceptual architecture with its focus on data virtualization does not exclude an actual movement of data when

necessary. Data architects are required to fine-tune the balance between these two possibilities, focusing on the key element of this approach, namely the ability to add meaning (or semantics) to data in a controlled, standardized and, if possible, automated way.

- ***What if?***
  What if your company didn't simply copy tech giants' strategy, but returned to its core competence? Your exorbitant business know-how, which guarantees you a competitive edge in specific knowledge domains, can only be further developed with knowledge-driven semantic AI approaches.

"You can't out-tech Big Tech. But you can out-knowledge them in your specific business domain."[55]

### *Data Engineer*

At their core, data engineers have a programming background. They are responsible for providing the data scientists with the corresponding data. They use this engineering knowledge to create data pipelines. Creating a data pipeline for large amounts of data means bringing numerous data technologies together. A data engineer understands the different technologies and frameworks and how to combine them into solutions to support a company's business processes with appropriate data pipelines.

In the context of systems based on enterprise knowledge graphs, data engineers mainly work within the automation loop and take care of the continuous (further) development of the knowledge graph as a service. In a graph environment, a major challenge for them is understanding knowledge graphs in the first place (why knowledge graphs? We have XML technologies!) and to learn new technologies, such as languages like SPARQL or GraphQL,[56] in order to combine them with conventional means like XSLT.

### *ML Engineer (MLOps)*

ML engineers are at the intersection between software engineering and data science. They bring data science models into production and ensure that

business SLAs are met. They are part of the continuous feedback loop essential to improving the validity and accuracy of AIs. Similar to the Citizen Data Scientist, ML engineers will increasingly be able to take over areas of the traditional data scientist with the help of AutoML tools.

### *Knowledge Engineer / Metadata Specialist*

Knowledge engineers such as taxonomists or ontologists either have a strong background in information management or library science, or they have evolved from a "classical" data or content manager to a metadata specialist with a special focus on organizing knowledge. They strive to introduce a general framework for organizing data and content within an organization on a larger scale, avoiding solutions that only work for an isolated area. They develop and maintain KOS or knowledge graphs and bring along corresponding methodological knowledge and tool expertise. Within the expert loop, they often interact with data and content managers, with SMEs, and, when it comes to the strategic development of a governance model for knowledge graphs, with AI architects or CDOs.

### *Subject Matter Expert (SME, Domain Expert)*

In many cases, SMEs have extensive expertise but little methodological knowledge to develop knowledge models. This makes the use of intuitive modelling tools all the more important, and it requires a governance model that will include the domain expert in a collaborative process.

Often there are also people who can fill the role of both SME and knowledge engineer at the same time. If this ideal case occurs, the knowledge acquisition bottleneck can be overcome most quickly. "Taxonomists with expertise in a particular subject area more often work on the larger taxonomies for indexing or retrieval support and especially on more complex thesauri. Ontologists are also typically subject matter experts, with perhaps some additional background in linguistics."[57]

### *Data Scientist / Data Analyst*

Data Scientists aim to use data to understand, predict and analyze relevant

events and their interrelationships while extracting knowledge and insights from structured and unstructured data. The key to this is obviously the availability of meaningful, high-quality data sets. Limitations in the availability of 'classical' Data Scientists and their often limited knowledge about the actual business domain have led to the fact that 'Citizen Data Scientists' are increasingly taking over the tasks of a Data Scientist, often with the help of AutoML tools. A related role is the 'Knowledge Scientist,' who increasingly acts as an intermediary between SMEs, Data Scientists and the business users.

***Business user / Customer / Citizen***

Normally, end users do not even notice that an application is based on a semantic AI or that they are vertices in a knowledge graph and are constantly feeding it with data as part of the user loop. KGs are a key to meeting the growing demand for recommender systems, self-service portals and applications. Digital transformation programs of many public administrations or companies aim at (partially) automating analytics tasks or knowledge-oriented dialogues. End users can wear many hats, e.g., in their role as employees they want to gain more transparency about which projects or open positions within their company correspond to their career ideas, or as a learner they want to get suggestions for personalized learning paths from the system, as a patient they want to benefit from an automatic symptom checker, as a business analyst they want to use intuitive and self-explanatory data dashboards, etc.

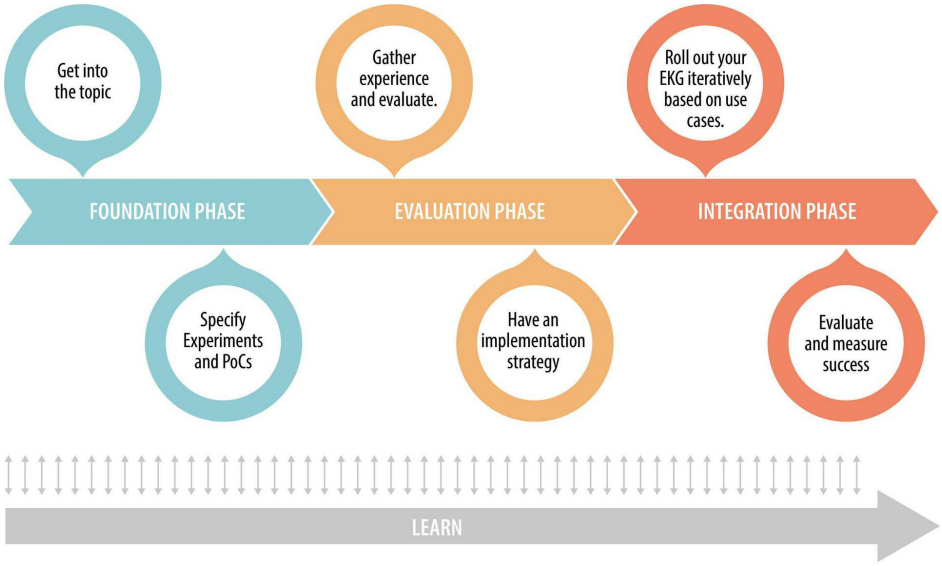
In any case, users are becoming increasingly demanding, especially with regard to the desired level of service via digital media, while at the same time user-related data is becoming more and more protectable. Under these circumstances, semantic AI with its ability to support precise automation even on the basis of smaller amounts of training data seems to provide the ideal methodological toolbox.

# Setting up an Enterprise Knowledge Graph Project

So how should one begin? Much like cooking, you can take courses and learn from people who know how to do it, or you can read books and try it yourself. Where and how you start depends very much on where you are in terms of the semantic maturity model of an organisation.

- Are you an enthusiast who wants to become a prophet of change in your organization?
- Do you belong to a group of people who have identified this as the next strategic goal to be achieved (and are in a position to achieve it)?
- Are you the one your management has chosen to evaluate this strange new promising thing and implement it based on the results?

Wherever you start—get into the subject; get a feel for what it is and what it means.




*Typical process for the initial phase of a KG project*

Then, try to find allies in your organization who will support you and more importantly, help provide real business cases and data that can be used in the evaluation phase. If you want to accelerate progress, find partners who can help you, especially during this initial phase. Evaluate and decide which toolset you want to use to begin your evaluation phase. Get your system operations involved early on so you don't get bogged down with discussions about tool selection and compatibility with the organization's policies later.

Set up a framework of agile tools to support your project management process. Collaboration systems such as wikis, workplace communication tools, etc., help to handle communication and documentation in an agile way. An agile development methodology like Scrum or Kanban helps with an iterative approach. If you are not yet familiar with these methods, now is the right time to try them out and learn. Find partners once more who work in this way to help you and learn from them.

Now you are prepared and ready for the evaluation phase. Let the fun begin:

- Make sure you have clearly defined business cases that generate value.
- Establish clearly defined success criteria that can be evaluated later.
- Make sure that you do not bring too many aspects into a PoC or experiment.

For example, it is not a good idea to try out what a knowledge graph can do in terms of better search and information retrieval, and combine this with an in-depth performance analysis of such a system. Both are legitimate evaluation criteria, but during a PoC they should be performed separately. Eventually, you should have found answers to the following key questions:

- What business values do knowledge graphs bring to my organization and how can I make them transparent and measure them?
- What skills and changes are required in my organization to bring the knowledge graph initiative into production?
- What tools and infrastructure are needed to bring the knowledge graph initiative into production?
- What are the first two or three initiatives to start with and are the necessary stakeholders on board?

Write a strategy paper or implementation strategy that covers the above points, as it will help you focus and sell the initiative in your organization.

Now you are ready and can start cooking for others. Get your evaluated knowledge graph infrastructure in place and start implementing it based on the initiatives you have chosen. Bring people in as soon as possible to get feedback, train them and learn from them. It must not end up with only one

"knowledge graph" specialist and a single department working on it. SMEs and business users must be involved so that the knowledge graph can grow, otherwise the knowledge graph will end up in an ivory tower.

# Circumvent Knowledge Acquisition Bottlenecks

The so-called "bottleneck in knowledge acquisition" is a well-known problem resulting from the fact that taxonomies and ontologies in each knowledge domain require input from SMEs and those who have this domain knowledge, and at the same time, the necessary knowledge engineering know-how, are scarce. The sources of information are huge, but to classify and structure them, there is often a lack of specialists who could form the ontological basis for the realization of enterprise knowledge graphs. Nevertheless, there are various strategies to avoid such bottlenecks:

- **Purchasing pre-built ontologies and taxonomies from the market**: the frequent problem with this approach is that any organization won't sufficiently benefit from an off-the-shelf product like this, and in most cases they have to be refined.
- **Automatic creation of taxonomies and ontologies from data available within an organization**: the promise of fully automatically generated ontologies is as old as the discipline of knowledge organization itself. It has not yet been fulfilled, and will most likely never be reached. The crux of the matter is that enterprise data does not contain the right information to be able to derive ontologies and taxonomies from it. Even if that's the case, using unstructured information to train machine learning algorithms in order to generate ontologies from it still needs some human in the loop, who is still capable of curating the results from an SME perspective.
- **Decomposition of complexity and using various tools to enable collaborative workflows**: this approach is related to AutoML and seems to be most promising and has been adopted by many organizations. Workflows and tools are used to enable SMEs, business users and knowledge engineers to collaborate and view knowledge models from their respective perspectives, while also enabling them to communicate better with each other.

# How to Measure the Economic Impact of an Enterprise Knowledge Graph

As with most quality-oriented initiatives, success is difficult to measure. Similar to cooking, tastes are different and influenced by cultural conditions. I was surprised to learn that the biscuits we love so much in Austria are far too sweet for my Chinese colleague, while the sweets he brings back from holiday are not what we would think of as sweet in our country. So the question is analogous: "how and, above all, who, can objectively measure the economic impact of a knowledge graph initiative?”

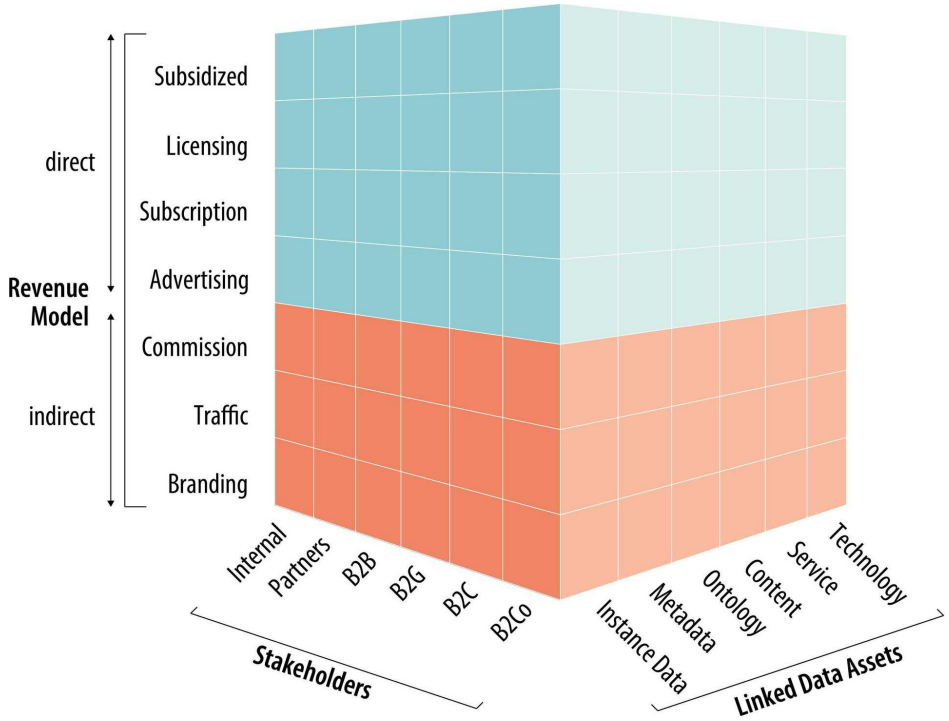
One of the main criteria could, of course, be to reduce the time needed for experts/knowledge workers to obtain information. According to McKinsey, employees in knowledge-intensive industries spend 20% of their time searching for internal information or finding the right contact person.[58] Since one of the main application scenarios for implementing knowledge graphs is improved search and retrieval, one way to calculate the ROI of such an initiative could be to calculate the reduction in time spent searching for information and people. But as we have seen, the benefits of knowledge graphs go far beyond those application scenarios where search and retrieval is the only focus of interest; instead, they can even fundamentally transform enterprise data management.

Within the framework of the European research project ALIGNED,[59] which is concerned with improving software development and the data lifecycle, we have carried out an integration of all the information about our development process in a search application based on a knowledge graph. In an empirical evaluation, we were able to show that the time required to find information in this integrated system could be reduced by about 50% when compared to searching through four different locations and manually combining the information. It should be noted that this was only a prototypical implementation for a research project. We would expect even better results in a productive system that is constantly being improved.

Closely related to this is another way of measuring the economic impact, namely, the evaluation of the time needed to integrate different data sources. Here too, the evaluation and thus the calculation of a return on investment, can be based on figures. The cost of integration by traditional means should

be determined from experience. An evaluation of the same integration using a knowledge graph could produce surprising results.

A third option to evaluate the economic impact is the fact that combining information from different systems via a knowledge graph will allow us to combine information in new ways that were not possible before. That, of course, allows us to identify new knowledge and based on that, offer new services and products. The Linked Data Business Cube,[60] which was developed in the course of the ground-breaking LOD2 project,[61] provides an integrated view on stakeholders (x-axis), revenue models (y-axis), and linked data assets (z-axis). This allows for the systematic investigation of the specificities of various linked data or knowledge graph-based business models.




*Linked Data Business Cube*

The creation of an enterprise knowledge graph should therefore not just reduce existing costs. It can be combined with the development of new business models for the knowledge assets, which should be available as a further result of this initiative and included in the ROI calculation.

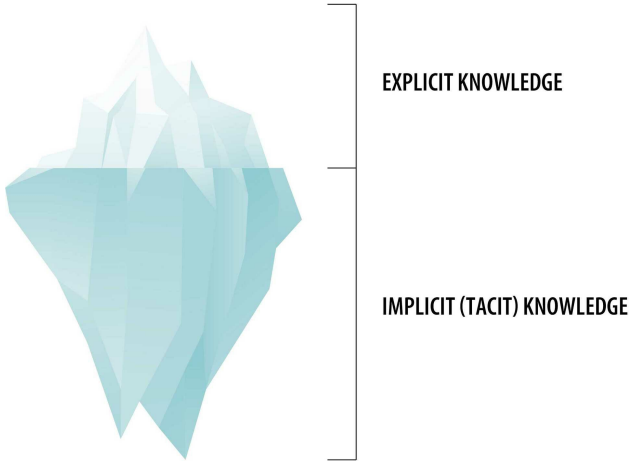
# PART 3: MAKE KNOWLEDGE GRAPHS WORK

## THE PROOF IS IN THE PUDDING!

# The Anatomy of a Knowledge Graph

## *"THE MAJORITY OF THE DATA IN EVERY ENTERPRISE KNOWLEDGE GRAPH IS ALWAYS GENERATED AUTOMATICALLY"*

Why are knowledge graphs such a hot topic lately? Because they take up and possibly solve one of the long-standing problems of knowledge management: they make implicit knowledge in people's heads explicit. Many of us may still have an image of an iceberg in our minds, where this small part protruding from the water reflects explicit knowledge, while a titanic-sinking amount of implicit knowledge lurks beneath the surface.




*How can organizations benefit from implicit knowledge?*

Let's be honest, it's not just because we in many aspects continue to have a working culture where knowledge sharing is not the standard and is only seen as a costly and time-consuming effort. But more and more organizations know that they can't continue burying their heads in the sand and ignore the problem of losing more and more knowledge.

Their knowledge is buried in a heterogeneous system without good opportunities to use it. We would like to suggest a more positive view and interpret it as a hidden treasure, whereby the knowledge graph offers a great opportunity to unearth this treasure. The knowledge graph thaws the iceberg and, above all, helps to finally let its hidden part enter the flow of work and knowledge.

But what do these knowledge graphs consist of? They reflect the way we think: how we collect, link and abstract facts. So, like a child, they have to learn from scratch what the world or a particular area is all about. And like a child, there are two fundamental possibilities of how this knowledge is learned. One is through experience, by looking at the world, by acquiring information about an area, or by experimenting and working in an area. The other is by getting help or guidance from experienced and knowledgeable people.

What does this mean for the creation of our knowledge graph? When we take a closer look at all the available information and experience from a field of knowledge, we can identify all the categories, types, things and objects that are important for that field, and we then understand more and more how they relate to each other and what information is available to describe them even more accurately. We call this the ‘conceptual model,’ and in a semantic knowledge graph this is represented by a schema or ontology.

Since we express knowledge not only schematically, but also, and above all, through human language, very individually and in different languages, we must also provide a ‘linguistic model’ for our knowledge graph. The linguistic model serves to label and further describe and contextualize the individual elements of the conceptual model and their individual instances. In a semantic knowledge graph this is made possible by controlled vocabularies such as taxonomies. The linguistic model is derived from the analysis of existing information from a domain and its instance data as well as from the experience gained in this field.

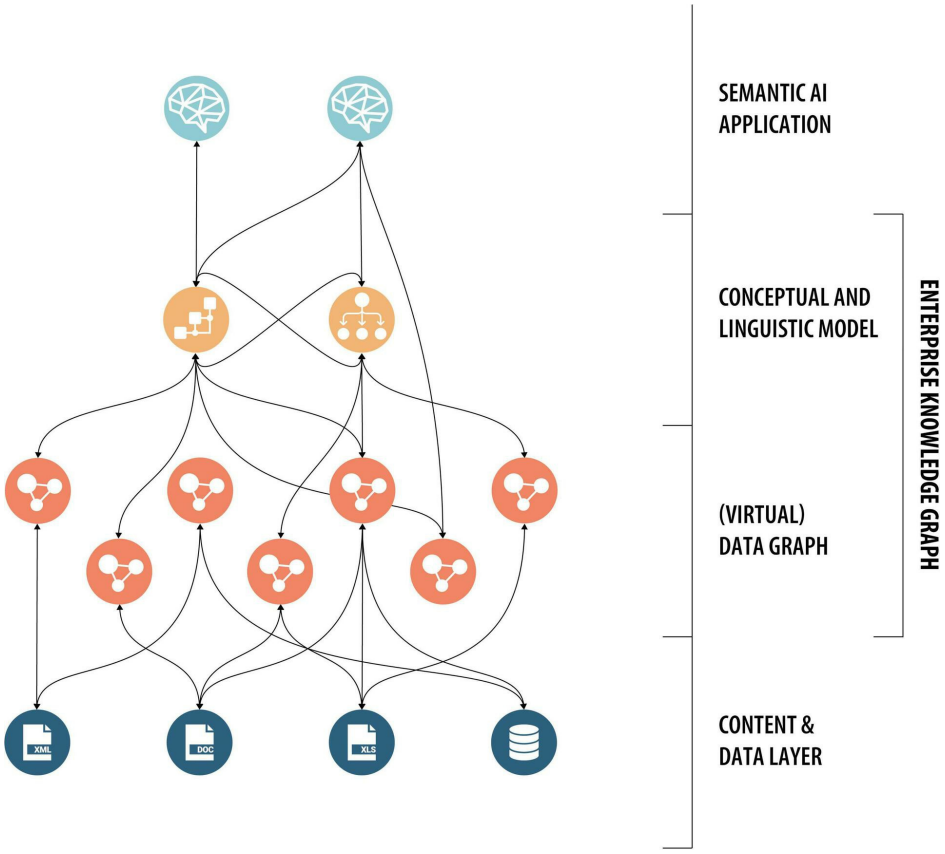
Part 1, the "information analysis" can be largely automated by machine learning, while part 2, the "deriving knowledge from experience" can be performed by domain experts. Ideally, these two elements are combined to build the conceptual and linguistic model based on information and experience in a given domain.

A scope for the domain to be represented by the conceptual model can be calculated by means of a reference text corpus or by means of so-called “key questions,” which are specified by potential business users.

Subsequently, domain knowledge can be made available as a machine-

processable domain model in order to largely automate the extraction and linking of instance data and documents from any given repository. We have now identified the three key elements of a knowledge graph:

- **Ontology**: conceptual model
- **Taxonomy**: linguistic model
- **Data Graph**: instance data and metadata; documents and annotations




*Four-layered Information Architecture*

At this point it should also be noted that the majority of the data in every enterprise knowledge graph is always generated automatically. Ontology and taxonomy behave similarly to DNA and RNA: the sentence "DNA is the blueprint for all genetic information, while RNA converts the genetic information contained in DNA into a format used to build proteins" can be translated into "ontology is the blueprint for all information within a domain, while taxonomy converts the information contained in ontology into a format used to generate actionable data and information."

# Basic Principles of Semantic Knowledge Modeling

Semantic knowledge modeling is similar to the way people tend to construct their own models of the world. Every person, not just subject matter experts, organizes information according to these ten fundamental principles:

1. Draw a distinction between all kinds of things: ‘This thing is not that thing.’
2. Give things names: ‘This thing is a cheese called *Emmental*’ (some might call it *Emmentaler* or *Swiss cheese*, but it’s still the same thing).
3. Create facts and relate things to each other: ‘Emmental *is made with* cow’s milk’, Cow's milk *is obtained from* cows’, etc.
4. Classify things: ‘This thing *is a* cheese, not a ham.’
5. Create general facts and relate classes to each other: ‘Cheese is made from milk.’
6. Use various languages for this; e.g., the above-mentioned fact in German is ‘Emmentaler wird aus Kuhmilch hergestellt’ (remember: the thing called ‘Kuhmilch’ is the same thing as the thing called ‘cow’s milk’—it’s just that the name or *label* for this thing that is different in different languages).
7. Putting things into different contexts: this mechanism, called "framing" in the social sciences, helps to focus on the facts that are important in a particular situation or aspect. For example, as a nutritional scientist, you are more interested in facts about Emmental cheese compared to, for example, what a caterer would like to know.

With named graphs you can represent this additional context information and add another dimensionality to your knowledge graph. Technically spoken, the context information is added to your triples as an additional resource (URI) to make a quadruple out of the triple.

8. If things with different URIs from the same graph are actually one and the same thing, merging them into one thing while keeping all triples is usually the best option. The URI of the deprecated thing must remain permanently in the system and from then on point to the URI of the newly merged thing.

9. If things with different URIs contained in different (named) graphs actually seem to be one and the same thing, mapping (instead of merging) between these two things is usually the best option.
10. Inferencing: generate new relationships (new facts) based on reasoning over existing triples (known facts).

Many of these steps are supported by software tools. Steps 7–10 in particular do not have to be processed manually by knowledge engineers, but are processed automatically in the background. As we will see, other tasks can also be partially automated, but it will by no means be possible to generate knowledge graphs fully automatically. If a provider claims to be able to do so, no knowledge graph will be generated, but a simpler model will be calculated, such as a co-occurrence network.

# Basic ingredients of Knowledge Graphs

## URIs and Triples

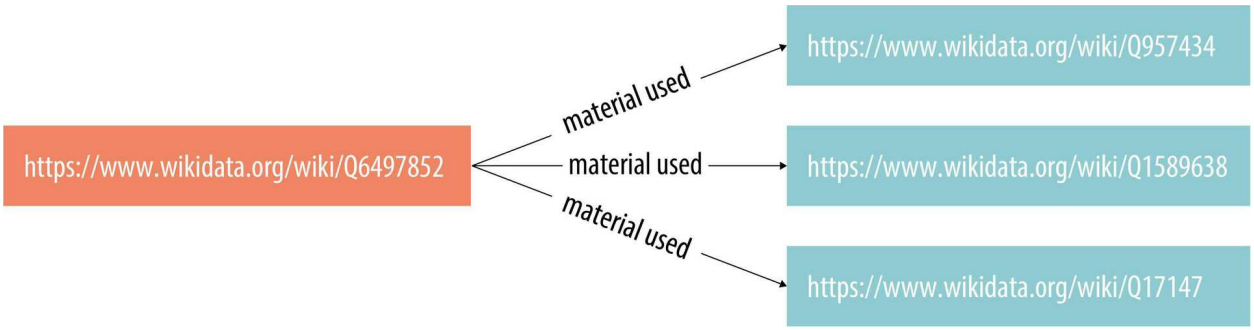
Knowledge graphs are primarily about things and therefore, when it comes to business, about business objects. Technically, each thing is represented and addressed by a Uniform Resource Identifier, a URI. So URIs are the foundational elements of your knowledge graph and you should treat them carefully. URIs are typically dereferencable and thus often HTTP URLs. For example: https://www.wikidata.org/wiki/Q6497852 is the URI of a 'thing' which is frequently called 'Wiener schnitzel.'

Now let's put things together into triples to express facts about things. The fact that the thing with the URI from above is called 'Wiener schnitzel' is expressed by a triple. Any triple consists of a subject, a predicate and an object or a literal (string, numerical value, boolean value, etc.):

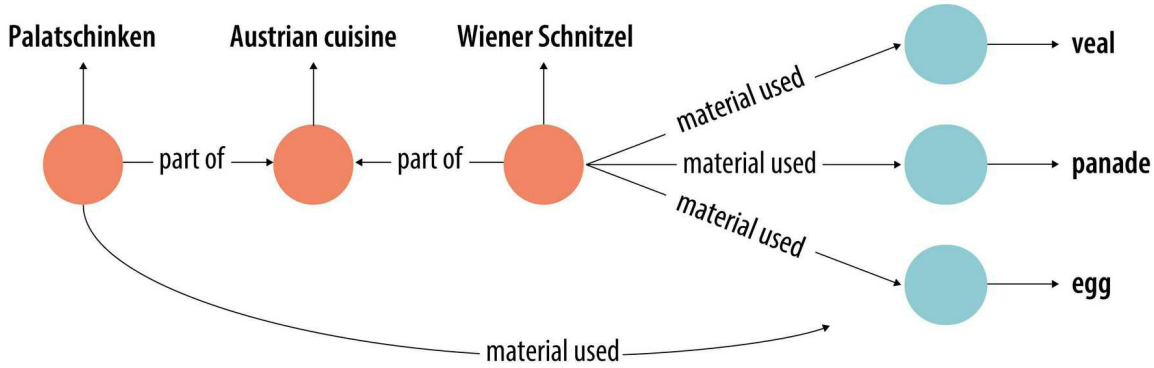

Another fact about this dish is that it's part of the Austrian cuisine, let's create another triple, now consisting of a subject, predicate, and an object (whereas the object on the right side is first the URI of a thing called 'Austrian cuisine'):


Another typical dish from Austrian cuisine are Palatschinken:

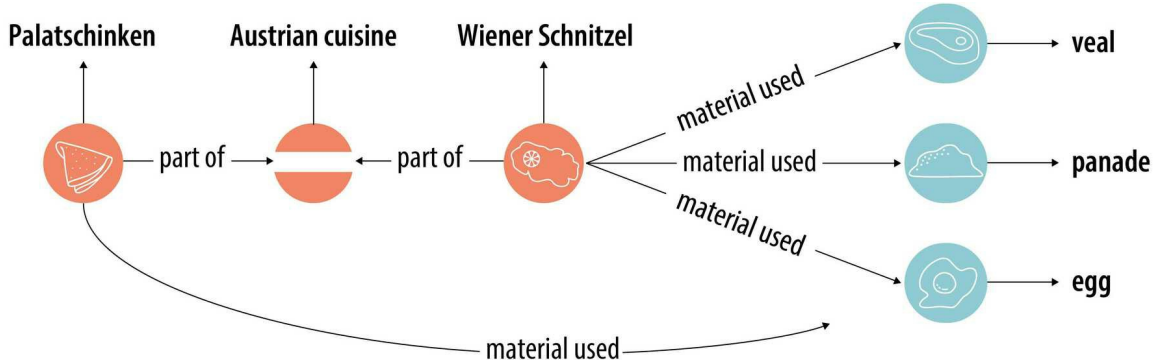

The fact that a Wiener schnitzel is made of (at least) three ingredients (veal, panade, and egg) is correspondingly expressed by the following three triples:




Let us now summarize all of these triples including the label information in a (small) knowledge graph, with the URIs in this version omitted for better readability. We also add the fact that Palatschinken also use eggs in their typical recipe:




The same knowledge graph could be visualized in an even more humane way:




In contrast to other graph models like labeled property graphs (LPG), RDF uses URIs for nodes and edges in directed graphs, and in doing so, they can be dereferenced to obtain further information, thus creating a network of linked data.

## RDF Triples and Serialization

What has been depicted as a human-friendly version above should be made machine-readable as well. To make triples available to be stored and further processed by RDF graph databases (a.k.a., 'Triple Stores'), RDF is used, being the most fundamental part of the W3C Semantic Web standards stack. RDF data can be serialized in different formats while representing at any time the exact same set of triples, for example: Turtle (TTL), JSON-LD, N3, or RDF/XML. Following the knowledge graph from above, we can express via Turtle[62] that 'Wiener schnitzel' uses veal meat and is suitable for people with lactose intolerance:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix schema: <http://schema.org/>.
@prefix wikidata: <http://wikidata.org/wiki/>.

<wikidata:Q6497852> rdf:type <schema:Recipe>.
<wikidata:Q6497852> skos:prefLabel "Wiener schnitzel".
<wikidata:Q957434> schema:isPartOf <wikidata:Q6497852>.
<wikidata:Q6497852> schema:suitableForDiet <wikidata:LowLactoseDiet>.
```

To any given set of triples, an additional URI can be added, making quadruples out of triples. This results in so-called 'named graphs', allowing descriptions to be made of that set of statements such as context, provenance information or other such metadata. TriG[63] is a W3C recommendation and an extension of the Turtle syntax for RDF to define an RDF dataset composed of one default graph and zero or more named graphs.

## Knowledge Organization Systems

As we saw in the previous chapter, all things can be connected to each other via networks of triples. At the instance level, this mechanism is quite simple and works intuitively. But let's take another look at the basic principles of semantic knowledge modeling: we have not yet started with tasks 4–10. Without classifying things, and if only arbitrary predicates are used to link things together, a knowledge graph quickly becomes messy and remains as flat as a typical mind map. With taxonomies, thesauri, and ontologies, we are now beginning to introduce additional dimensionality into every graph, and

we are standardizing the meaning of instance data, resulting in better machine readability.

### *Taxonomies and Thesauri*

Which of my recipes are good for vegetarians? The filter or constraint we have to apply is that the recipe must not contain meat or fish. Furthermore, if we want to find out what is good for vegans, we also need to know which ingredients belong to the class of animal products. The graph above does not contain such information, so what options do we have to introduce this into the model?

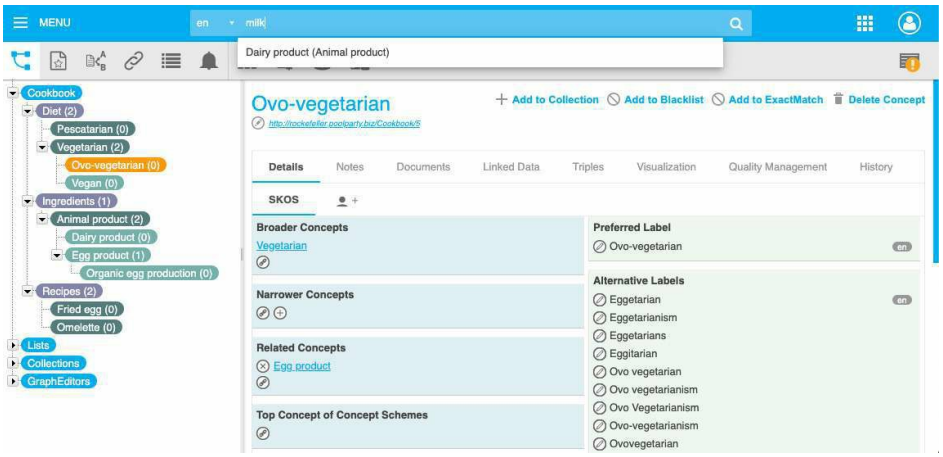
Option 1: we keep lists of things and start to annotate them.

- We add the attributes vegan/vegetarian = yes/no per meal.
- Alternatively, we add these attributes per ingredient and infer that only dishes that do not contain such ingredients are good for vegetarians/vegans.

Option 2: we start to build a thesaurus and put things into hierarchical orders.

- We introduce a new thing called "animal product" and begin to build hierarchies of things to bring order and more meaning to our list of things. For example, we add "dairy product" or "egg product" under "animal product", and further down the hierarchy we find that "mayonnaise" is an "egg product", etc.
- We start by introducing consumer types such as "vegetarian" and below "ovo-vegetarian" and associate this concept with "egg product", expressing that all egg products can be eaten by ovo-vegetarians and anything else that is acceptable to vegetarians in general.

We choose option 2, and here is the resulting taxonomy:



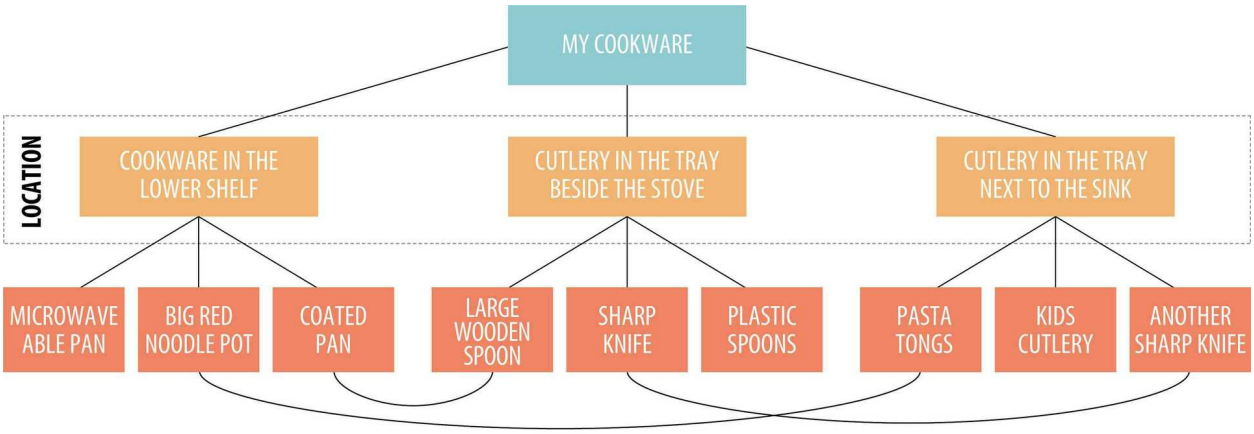

*Screenshot: PoolParty Thesaurus Server*

With this simple taxonomy we have laid the foundation for a scalable knowledge graph. No matter how many recipes, ingredients, or consumer types we add to the model later, all applications, reports, and underlying queries will still work. Thus, it is no longer necessary to make additions or changes at the attribute level for each thing, for example, if we want to introduce a new consumer type like 'Pescetarian.' Instead, this category is simply added to the knowledge graph as a new sub-concept of 'Vegetarian' and linked to the appropriate ingredient categories.

*What are taxonomies?*

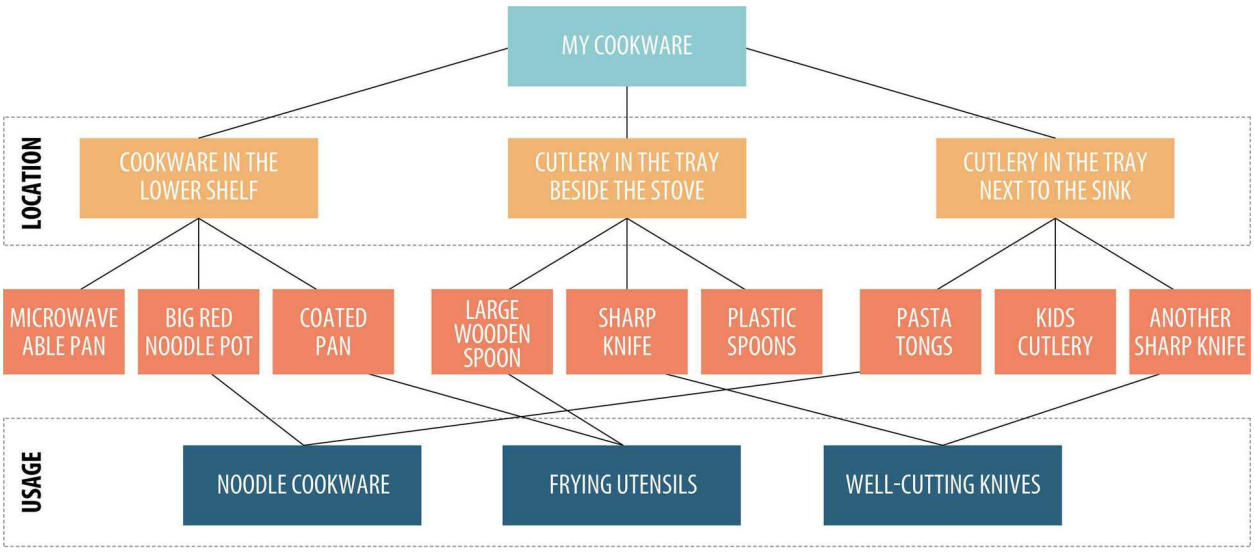
## *"THE WORLD IS POLYHIERARCHICAL"*

At first glance taxonomies sound like strange animals. In reality, we all use taxonomies to bring order into our lives. For example, to master our cooking adventures, we first bring order to our kitchen by following a taxonomy like this one:




Taxonomies are used to find things (documents, experts, cookware, etc.) and help classify them. The tricky thing is that the world is more complex than a mono-hierarchical taxonomy could express, as shown above for example. There are many ways to classify things into meaningful categories. The world is polyhierarchical.

As already indicated by the dotted lines, the "big red noodle pot" and the "noodle tongs", for example, also fall into the category of "noodle cookware", not just into the single category to which they are currently assigned. Accordingly, we can extend the taxonomy from above and introduce new categories by already presenting the taxonomy in a graph instead of a simple tree—we make the model poly-hierarchical.




Taxonomies therefore contain as many categories as necessary, and each thing can be assigned several times. With categories, we give a thing an

additional context. In our example, we introduced two types of contexts to help the cook classify and find things in relation to their location in the kitchen, and the second type of context is about the ways of using the cooking utensil.

Before we describe methodologies to build and manage taxonomies, which we will outline with more detail in the Taxonomy Management chapter, we will take a closer look at the SKOS[64] data model which is broadly used to represent taxonomies and thesauri.

*Concepts, concept schemes and relations*

At the center of each SKOS-based taxonomy there are so called ‘concepts’ (`skos:Concept`). A concept can represent any kind of entity or business object. Concepts are organized within so-called ‘concept schemes’ (`skos:ConceptScheme`) which should contain only concepts of the same kind. A taxonomy about cooking could consist of several concept schemes, for example: cookware, ingredients, dishes, and consumer types. Concepts within the same concept scheme are typically either hierarchically (`skos:broader/narrower`) or non-hierarchically (`skos:related`) related, concepts across two concept schemes have typically only non-hierarchical relations between them.

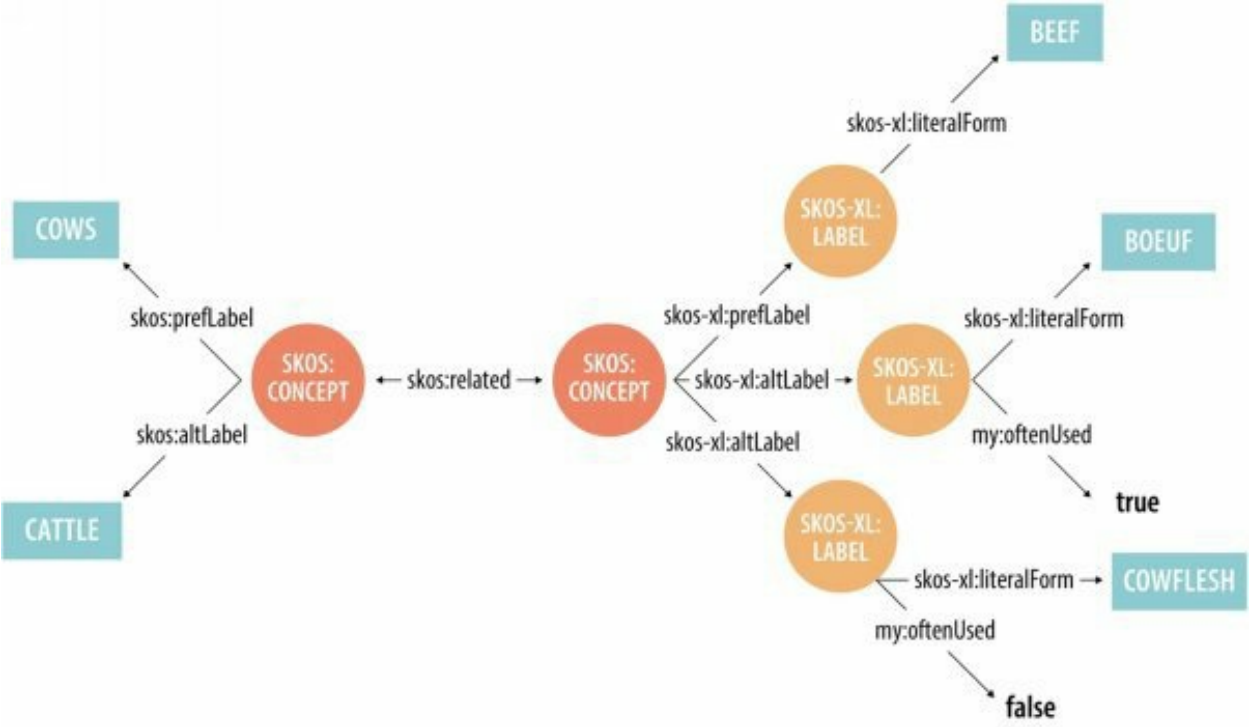
*Top Concepts*

Taxonomies and concept schemes can have as many hierarchies as necessary. When using a graph database instead of a relational database, a deep hierarchy would not cause any problems, since it is in the nature of graphs that there are no limits in this aspect. However, there is an outstanding type of concept, the so-called ‘top concepts’, which are located at the first level of a concept schema. In our example from above, "cookware on the bottom shelf" or " noodle cookware" would be top concepts, which serve as categories and structural elements of the taxonomy and are not an entity by themselves.

*Concept labels*

Each concept has at least one label per language, the so-called ‘preferred label’ (`skos:prefLabel`), but any number of synonyms, also called

alternative labels (`skos:altLabel`). Labels give names to concepts and should cover all the identifiers and surface forms of all the things that are found in an organization's digital assets. Labels are leaves of the graph, so additional triples cannot be attached to a label.




*Example of a SKOS/SKOS-XL graph*

An extension of SKOS is SKOS-XL, which essentially contributes to the fact that concept labels can also be addressed as resources or nodes in order to make further statements about labels. With SKOS-XL you can say, for example, that a certain alternative label should only be used for marketing or for internal purposes, or that one label is the successor of another label, which is important for technical documentation, for example.

### *Ontologies*

*"ONTOLOGIES ARE USED TO GIVE MORE DIMENSIONALITY TO A KNOWLEDGE GRAPH"*

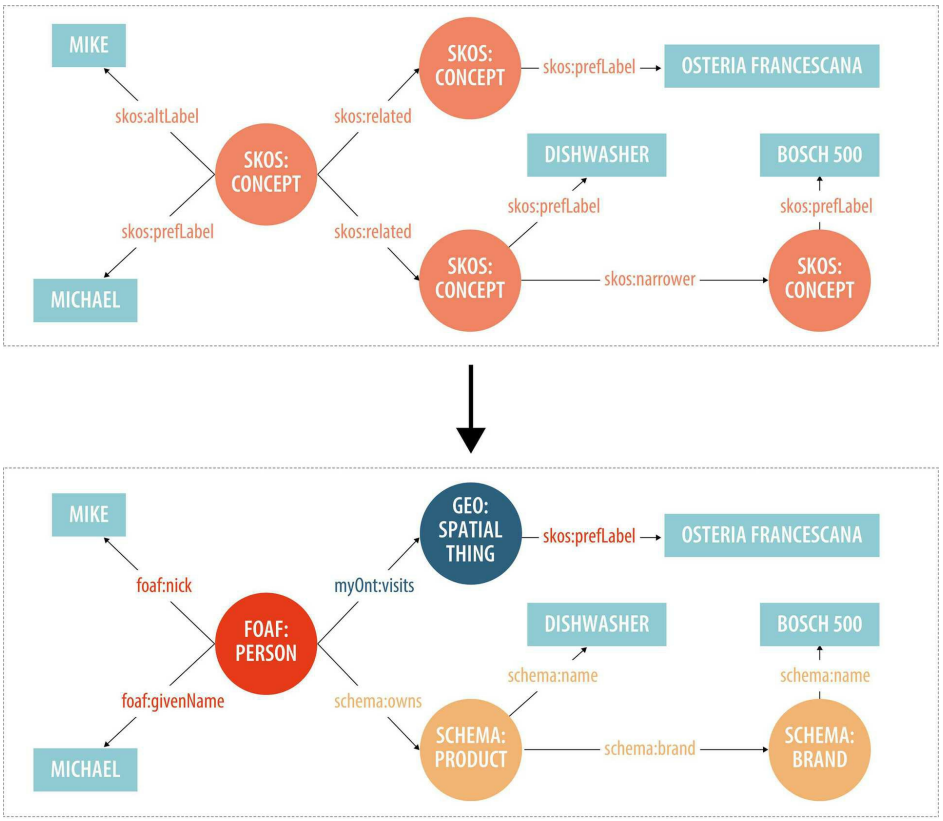
In many of our projects we have seen how organizations have started with taxonomies based on SKOS to construct the first pillar of their knowledge

graph. The elegance of SKOS lies in the fact that it is relatively easy to create and maintain, but it is not too simply woven to already cover some important use cases for KGs. In each further developmental phase, SKOS taxonomies can be extended or can be integrated into more comprehensive knowledge graphs, since SKOS is part of the Semantic Web standard stack.

> ***"Everything should be made as simple as possible, but not simpler."***
>
> —ALBERT EINSTEIN

Most knowledge engineers have a very good understanding of what can be achieved with taxonomies and when ontologies should come into play. In SKOS all things are instances of a single class, namely skos:Concept. This also allows us to relate everything to everything else, using the unspecific relationship types 'related', 'broader' and 'narrower'. All this is defined by the OWL specification of SKOS.[65] This means that you already have knowledge about ontologies as soon as you know the basic features of SKOS.

This simplicity, on the other hand, has several shortcomings. Illogical relations cannot be avoided or automatically detected. For instance, if you relate an ingredient to a dish, that makes perfect sense, but what exactly could an unspecific `skos:related` between two things (or concepts) mean, which are actually two ingredients? Ontologies are used to give more dimensionality to a knowledge graph: Ontologies classify things and define more specific relations and attributes.




*How SKOS is extended with ontologies*

With the help of ontologies, entities can be specifically classified and thus become an instance of one or more classes. In conjunction with this, restrictions can now be expressed in order to develop a more consistent knowledge model. For example, OWL 2[66] can be used to express that all instances of the class Recipe should have at least one Ingredient, or that instances of the class Soup cannot be members of the class Dessert.

Ontologies are also used to take advantage of inference mechanisms. This is essential for data integration tasks: ontologies are not only a perfect data structure to map relational data models into the graph world, they are also important to detect possible inconsistencies in the (integrated) data. Furthermore, ontologies also allow the discovery of new relationships, for example, if, within the ontology, “vegan diet” is defined as a subclass of “vegetarian diet”, and "Rainbow Spring Rolls" are classified as “vegan”, then these are automatically also allowed as “vegetarian diet.”

With OWL 2 and RDF Schema and their various levels of expressivity, the Semantic Web Developer has various ontology languages at their disposal,

which should probably be used in a dosed manner in order to be combined with SHACL in a further step: OWL is used to specify constraints to prevent inconsistent data from being added to an RDF graph database. Often, however, data from new sources is structurally inconsistent with the constraints specified via OWL. Consequently, this new data would have to be modified before it could be integrated with the data already loaded into the triplestore. In contrast to OWL, SHACL can also be used to validate data that already exists in the triplestore.[67]

To sum up: SKOS Taxonomies offer a great basis for all types of text mining tasks, although with ontologies in place, more complex rules for entity extraction can be applied. As soon as a knowledge graph initiative seeks for ways to deal with structured and unstructured data at the same time, ontologies and constraints will become mandatory.

# Reusing Existing Knowledge Models and Graphs

So one of the key questions at the beginning of creating your enterprise knowledge graph will be: "Do I have to start from scratch?" And we could start now with a nice metaphor, which is true in the end: "making soup from the bone will always be better than making it from a bouillon cube, but of course, it will also require more effort." So already existing knowledge models and graphs can speed up your process, but you should at least take the time to adapt them to your use case and needs, and the better the pre-built knowledge graph fits your use case or domain, the less you have to adapt. See our section about good practices for further considerations on reusing existing knowledge graphs.

We distinguish between two main types of knowledge graphs:

- world knowledge graphs, and
- domain knowledge graphs.

In the following sections we will provide an overview of both, and it will cover taxonomies, ontologies and full-fledged knowledge graphs.

## World Knowledge Graphs

So what do we mean by "world knowledge graphs"? These are knowledge graphs that do not focus on a single field of knowledge, but try to collect and structure all the knowledge of the world. There are closed world knowledge graphs that cannot be fully reused, such as the Google Knowledge Graph or Microsoft Satori, both of which are used as the basis for their respective search services, Google Search and Microsoft Bing. Interestingly both of them state that they are based on pre-existing knowledge graphs like Freebase[68]—acquired by Google in 2010—and Wikipedia or, in their semantic and publicly accessible forms, Wikidata[69] and DBpedia. Both have been, of course, vastly extended by editorial work and machine learning, basically following the principle proposed here for the development of enterprise knowledge graphs.

So how can we make use of reusable world knowledge graphs like Wikidata,

KBpedia[70] and DBpedia, or upper ontologies like the Basic Formal Ontology (BFO)[71] or Schema.org—all based on linked open data principles,[72]—to get our enterprise knowledge graph going?

- They could contain relevant subsets for our area, which we can cut out and use as a starting point.
- They offer generic ontologies, which in turn can be reused and further refined for different areas.
- Finally, they can provide relevant information on many general topics that you would like to include in your enterprise knowledge graphs, such as geographic information, information about places, events, brands and organizations, etc.

However, we suggest that the usefulness of such graphs should be carefully evaluated before reuse, as the content is often too generic and the quality varies. Wikipedia and its semantic derivatives have become more and more extensive and often of higher quality in many areas over time, but of course, there are still incorrect or contradictory information or structural problems. So in some cases it will be more work to curate the data you get from there to achieve the required quality than to create the same data from scratch.

## Domain Knowledge Graphs

In principle, you will find knowledge graphs or at least ontologies and taxonomies for any field of interest from which you can start your work. Search engines like Linked Open Vocabularies (LOV)[73] or the Basel Register of Thesauri, Ontologies & Classifications (BARTOC)[74] help you to find such starting points. In any case, you should not blindly reuse your findings, but first check whether reuse is reasonable for your application. We have collected some more detailed information on the following areas:

- business and finance,
- pharma and medicine,
- cultural heritage,
- sustainable development, and
- geographic information.

### *Business and Finance*

In financial services organizations, “data is often disconnected and stored in different formats, creating isolated repositories of information that are not available to an entire organization. This makes bank-wide research ineffective and prevents artificial intelligence applications from discovering insights from data.”[75]

The central resource for the business and finance domain is the Financial Industry Business Ontology (FIBO). It “defines the sets of things that are of interest in financial business applications and the ways that those things can relate to one another.”[76] FIBO provides a SKOS vocabulary as well as an extensive OWL ontology. It is maintained and updated on a regular basis by the EDM council.

Another valuable resource in this domain is the Standard Thesaurus for Economics (STW)[77] that is provided by the Leibnitz institute in Germany providing a multilingual vocabulary for the economic domain in German and English.

XBRL[78] is an international standard for digital reporting of financial, performance, risk and compliance information. It defines authoritative taxonomies for reporting terms and allows synchronizing reporting information between different departments or organizations. Different taxonomies for business reporting purposes are available.

The Currencies Name Authority List[79] is a controlled vocabulary listing currencies and sub-units with their authority code and labels in the 24 official languages of the EU provided by the Publications Office of the European Union.

World Bank Topical & World Bank Business Taxonomy[80] are two vocabularies describing the organizational structure, subject fields and activities and the business concept of the World Bank.

EuroVoc[81] is a multilingual SKOS thesaurus maintained by the Publications Office of the European Union, which covers areas such as economics, trade, business and competition, or employment.

UNBIS Thesaurus[82] is a multilingual database of the controlled vocabulary used to describe UN documents and other materials in the Library's collection.

For Human Resources-related tasks, ESCO, the European multilingual classification of skills, competences, qualifications and occupations, serves as a valuable resource. ESCO helps to describe, identify and classify professional occupations, skills, and qualifications relevant for the labour, education and training market. It is used by semantic search engines like Monster[83] or by recommender systems like the PoolParty HR Recommender.

Thomson Reuters Permanent Identifier (PermID)[84] offers business and especially the financial industry a comprehensive way to uniquely identify or reference entities of different classes, such as organizations, financial instruments, funds, issuers and persons. Thomson Reuters has been using PermID in the center of their own information model and knowledge graph for many years.

### *Pharma and Medicine*

The pharmaceutical and medical sector has always been one of the pioneers in the field of knowledge graphs. A starting point to find ontologies and taxonomies in this domain is the BioPortal.[85] Their vision is that “all biomedical knowledge and data are disseminated on the Internet using principled ontologies in such a way that the knowledge and data are semantically interoperable and useful for furthering biomedical science and clinical care.”

BioPortal gives access to most of the main sources in this domain like:

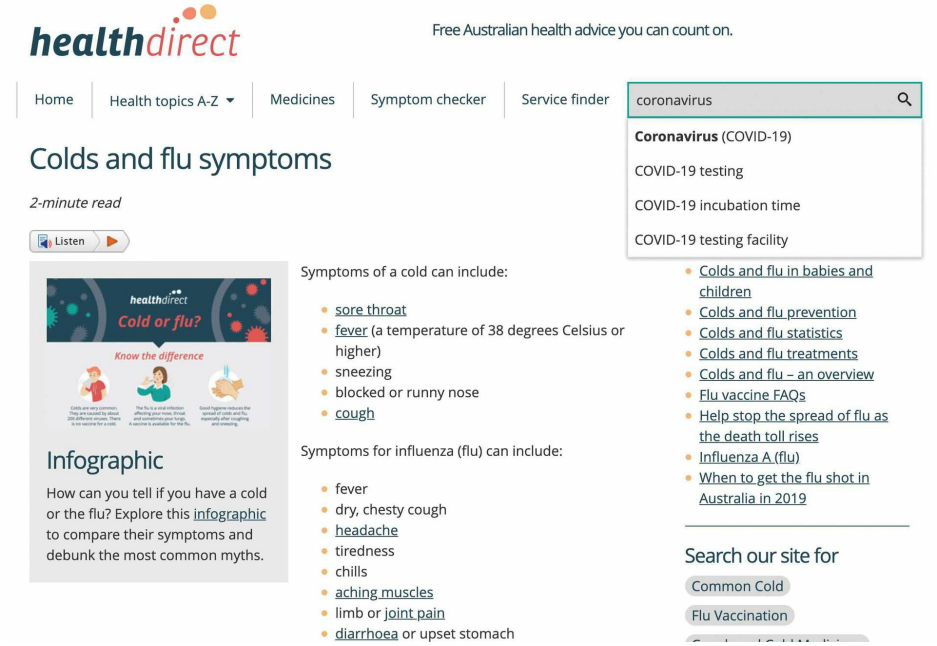
- Medical Subject Headings (MeSH), a controlled vocabulary, created and maintained by the National Library of Medicine (NLM) that provides a consistent way of retrieving information on medical subjects.
- Chemical Entities of Biological Interest Ontology (ChEBI), a vocabulary of molecular entities focused on ‘small’ chemical compounds.

- International Classification of Diseases (ICD), which is the international classification of diseases and related health problems published by the World Health Organization (WHO).
- SNOMED Clinical Terms (SNOMED CT) as one of the most comprehensive, multilingual clinical healthcare terminologies in the world.
- Gene Ontology provides structured controlled vocabularies for the annotation of gene products with respect to their molecular function, cellular component, and biological role.
- And many more.

Another well-known source is the Open Biological and Biomedical Ontology (OBO) Foundry.[86] The OBO Foundry's mission is "to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate." Most resources provided by the OBO Foundry can also be found via the BioPortal.

Additionally, the European Bioinformatics Institute (EMBL-EBI) maintains "the world's most comprehensive range of freely available and up-to-date molecular data resources."[87]

An example of how one of these resources has been used as a starting point for developing an existing knowledge graph into something more specific is the Australian Health Thesaurus (AHT).[88] AHT serves as the backbone of Healthdirect,[89] Australia's largest citizen health portal, and is based on MeSH, but has since been adapted to the specific Australian health system.




*Screenshot: Healthdirect.com*

## *Cultural Heritage*

A fundamental challenge in relation to cultural heritage data, which are usually provided by different cultural heritage stakeholders in different languages and in various formats, is to make them available in a way that is interoperable with each other, so that they can be searched, linked and presented in a more harmonized way across data sets and data silos. Let us look at some examples of how the GLAM sector (galleries, libraries, archives and museums) uses knowledge graphs:

Using data from over 3500 European museums, libraries and archives, Europeana[90] provides access to millions of books, music, artworks and more with sophisticated search and filter tools. One way to access the data of Europeana and to use it with other applications is via their SPARQL endpoint which allows to explore connections between Europeana data and outside data sources like VIAF,[91] Getty Vocabularies,[92] Geonames, Wikidata, and DBPedia. By that, more sophisticated queries can be executed, as for example, to find objects in Europeana linked to concepts from the Getty vocabulary.

Initiated by the Library of Congress, BIBFRAME[93] provides a foundation

for bibliographic description that is grounded in Linked Data techniques. It was developed as an alternative to the commonly used MARC 21 formats. A lightweight approach that goes in a similar direction is OCLC's WorldCat Linked Data Vocabulary,[94] which uses a subset of terms from Schema.org as its core vocabulary.

Three starting points to explore options to use standardized classification systems and controlled vocabularies based on the Semantic Web in the GLAM sector are, firstly, the "Library of Congress Subject Headings,"[95] secondly, "The Nomenclature for Museum Cataloging"[96] provided by the Canadian Heritage Information Network (CHIN) and some other North American organizations, and thirdly, the SPARQL endpoint of the Getty Vocabularies.[97] A frequently used ontology in the GLAM sector is CIDOC CRM, which "allows the integration of data from multiple sources in a software and schema agnostic fashion."[98]

Also on a national level, there are various data platforms that offer new ways of approaching their collections and resources by providing linked open data. Some examples are the bibliographic data portals of the National Library of Spain,[99] UK,[100] or of Germany,[101] or ArCo,[102] which is the knowledge graph of the Italian cultural heritage.

### *Sustainable Development*

Sustainable development is an organizational principle that must relate and link different goals and thus measures, methods and ultimately data and knowledge models with each other. It is therefore an excellent field of application for linked data and knowledge graphs. Accordingly, one can build on numerous well developed and established sources in this field, e.g., SKOS-based taxonomies like the Sustainable Development Goals Taxonomy[103] and UNBIS Thesaurus[104] as components of the United Nations' platform for linked data services, which is hosted by the Dag Hammarskjöld Library, or other sources like Clean Energy Thesaurus,[105] GEMET,[106] Agrovoc,[107] or KG services like Climate Tagger[108] or Semantic Data Services of the European Environment Agency.[109]

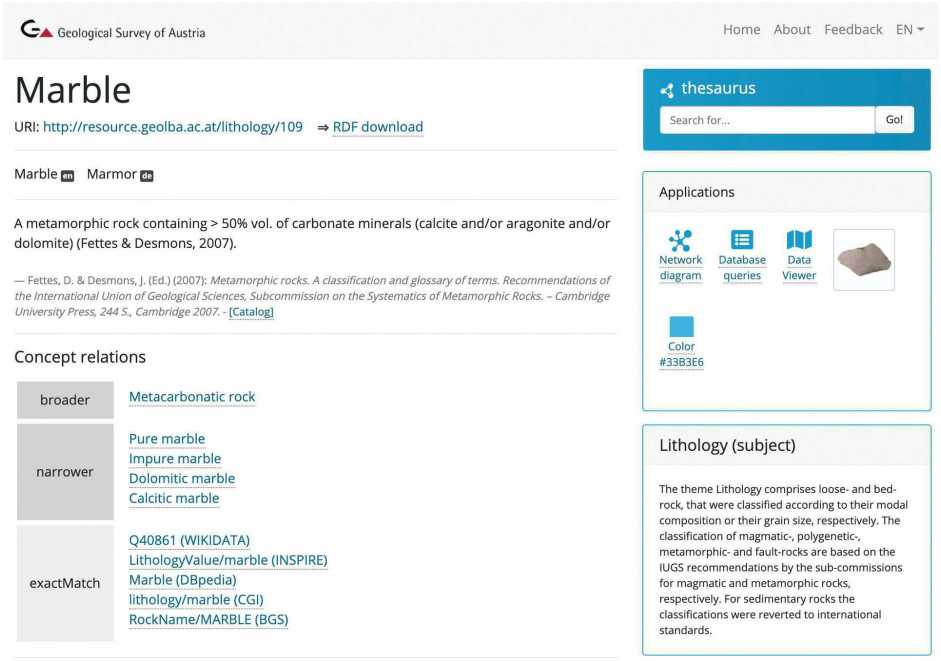
### *Geographic Information*

With Geonames,[110] one of the most complete knowledge graphs of geographical information is available, based on its own ontology[111] and integrable via API or as a data dump, with a choice of free or premium versions. Geonames is a great source to link your own enterprise knowledge graphs with location information and enrich them with additional data.

In addition, the Library of Congress Linked Data Services[112] as well as the EU Open Data Portal[113] provide authority lists for geographic entities like countries and regions.

The INSPIRE Geoportal[114] as the central European access point to the data provided by EU Member States and several EFTA countries under the INSPIRE Directive. INSPIRE is an EU initiative to help make spatial and geographical information more accessible and interoperable for a wide range of purposes in support of sustainable development.

GBA Thesaurus[115] is a controlled vocabulary based on SKOS for geosciences as used in geoscientific text publications and geological maps of the Geological Survey of Austria. Their datasets are coded with thesaurus terms, while the thesaurus is linked to INSPIRE terminology at the same time.




*Screenshot of GBA Thesaurus*

Furthermore, many World Knowledge Graphs contain rich geo-information, e.g., about 2 million entities in DBpedia are geographical things.

# Methodologies

## Card Sorting

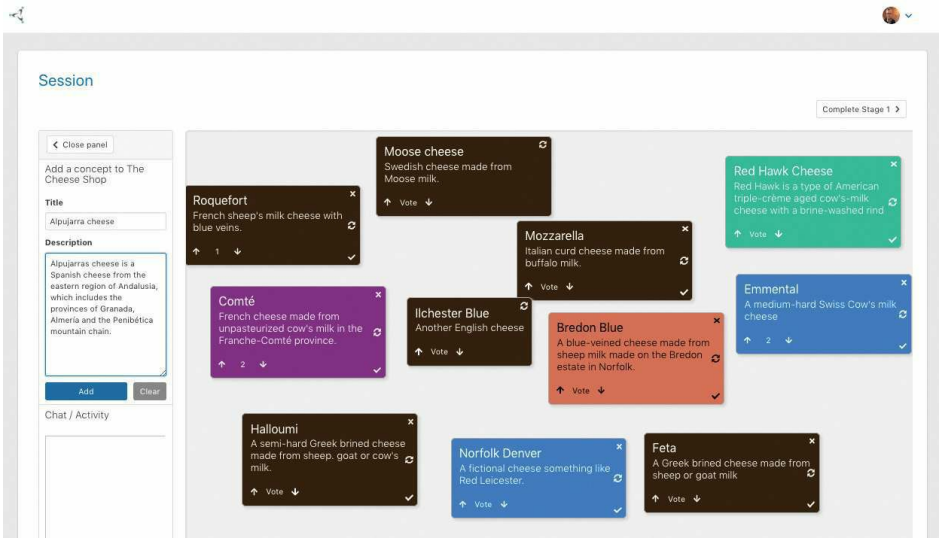
Card sorting is a method of identifying topics, naming them and putting them into categories that make sense for a group of people. It is a commonly used method to outline a domain or inventory a dataset in order to create a business glossary, which is later extended to taxonomies and ontologies and finally to an enterprise knowledge graph.

To set a scope for a card-sorting session, so-called "key questions" must first be formulated by business users. These questions thus define the knowledge that potential applications such as chatbots, search or analysis tools must be able to access later.

Actual cards or a card sorting software tool are often used to perform card sorting. Card sorting can be performed by any subject matter expert who has some knowledge of a particular domain of knowledge. There is no need to have a background in knowledge modeling, ontology engineering, or any related discipline.

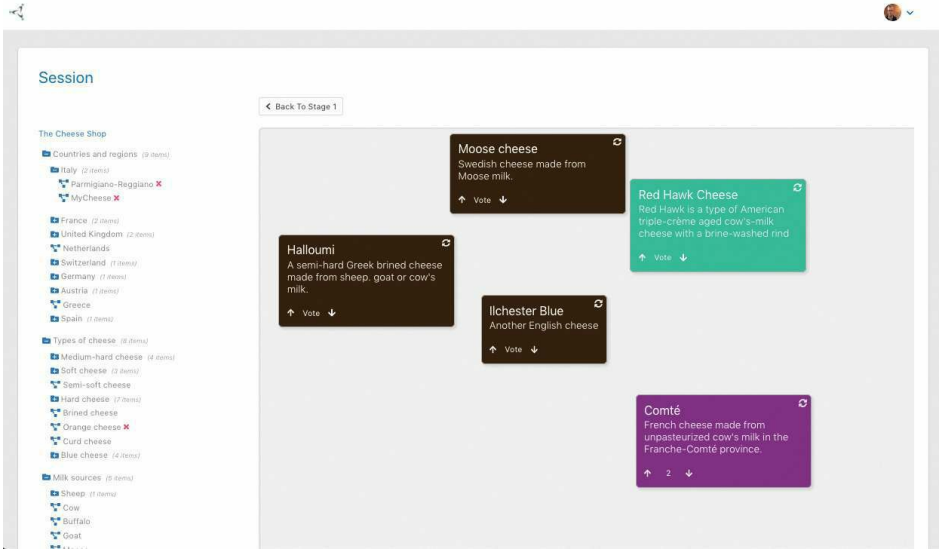
This aspect of card sorting makes it the perfect entry point into a broader process of knowledge graph development. It serves as the simplest version of semantic knowledge modelling, preferably in collaborative working environments, which can already produce a basic structure of a knowledge graph.

Below find an example screenshot of an online card sorting tool, which is part of the PoolParty Semantic Suite platform. This tool allows collaborators to suggest, confirm or reject new topics in a very intuitive way. Each card represents a thing (or topic of interest), and their colors indicate who the author was.




*Screenshot of a card sorting tool - brainstorming phase*

All accepted cards can be inserted into an already existing taxonomy using drag & drop. In this way, both activities, card sorting and taxonomy management, are seamlessly integrated. Typically, the creation of taxonomies, which later form the backbone of a knowledge graph, is initiated by some card sorting activities.




*Screenshot of a card sorting tool - structure phase*

This allows you already in early stages of your knowledge graph project to involve subject matter experts who have no or only little knowledge of knowledge engineering.

Card Sorting was originally developed to be used by information architects as a technique in user experience design. Jacob Nielsen first reported on a successful application of the Card Sorting method in 1995.[116]

## Taxonomy Management

The process of developing, continuously improving and embedding taxonomies or thesauri in business processes is often referred to as Enterprise Taxonomy Management. Usually one or more taxonomists and ontologists, SMEs, and often business users or data engineers are involved.

Successful implementation depends on whether a governance model appropriate to the organization is developed and whether the applied process model and software tools can generate the desired ROI within the defined time frame.

As we will see, taxonomies or taxonomy management are one of several "dishes" or "recipes" that have to be embedded in a broad menu, i.e., in a broader process model, in accordance with the Knowledge Graph Life Cycle.

### *Taxonomy Governance*

In contrast to the rather static, often monolithic and usually hardly agile process of maintaining classification schemes (e.g., Dewey decimal system[117]), the development of taxonomies and thesauri, especially when they are later used as part of a larger enterprise knowledge graph, is highly collaborative, networked, and agile.

The purpose of taxonomies, especially in enterprises, is mainly to tag and retrieve content and data later, but not to model a domain of knowledge or establish a strict regime for the later classification of digital assets.

In many cases there are several taxonomies per organization. These are managed by different departments according to their custom governance model, and in many cases these taxonomies are linked together to form the backbone of a larger enterprise knowledge graph. This can be achieved through a central vocabulary hub or through a more decentralized approach (similar to peer-to-peer networks) and requires a different way of thinking

than that often developed by traditional librarians or catalogers.

Managing taxonomies also means establishing a continuous process to ensure that new developments in the market or in the organization are well reflected and incorporated. This requirement deserves a balanced process in which automatic and manual work support each other.

In short, a corporate taxonomy governance model[118] for the sustainable management, linking and provision of corporate taxonomies throughout the organization must be well thought through and strictly aligned with the objectives of a larger KG initiative and its underlying governance model. A set of roles, responsibilities and processes must be defined to manage the development and application of a taxonomy so that it remains consistent and coherent over time.

### *Process model*

Developing taxonomies in organizations also means bringing together different stakeholders to agree on the scope, structure and content of a knowledge or business area. In addition, a common understanding of the business objectives to be pursued through the development of taxonomies and the higher-level enterprise knowledge graph must also be established.

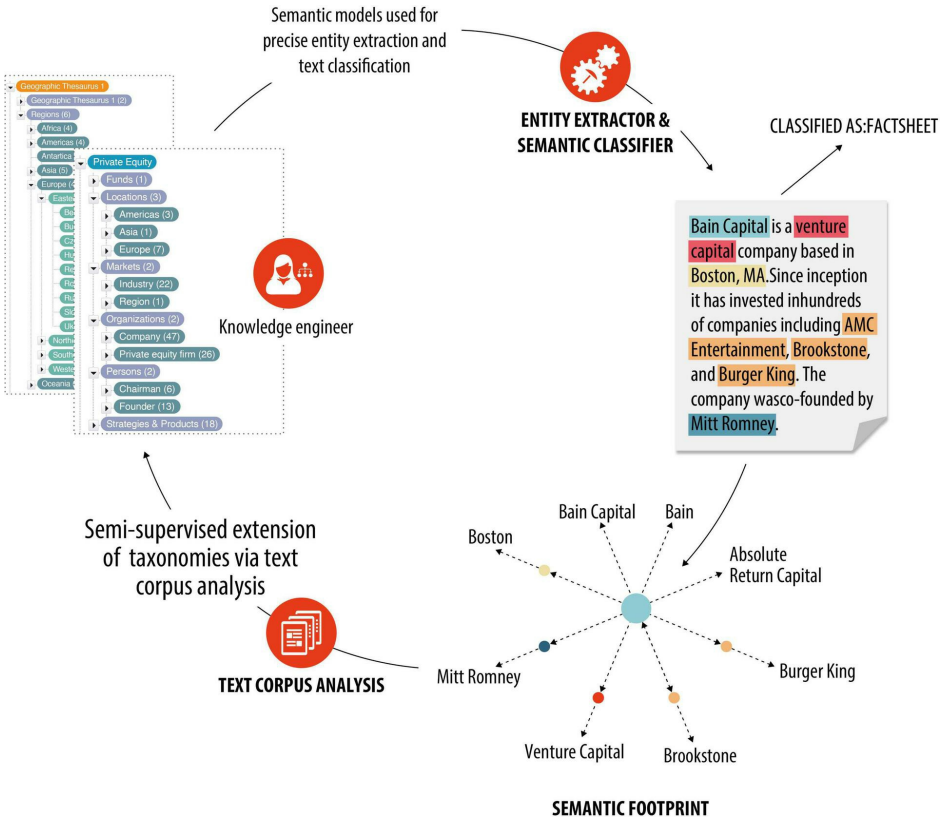
This agreement process rarely starts on a greenfield site, but in this case the method of card sorting can often provide a good starting point.

In many cases, however, it is already possible to extract taxonomy-like structures from existing category systems, tables or databases, which then serve as the basis for further modelling steps.

Furthermore, in many cases, thanks to the open standards of the Semantic Web, it is possible to fall back on already well developed taxonomies, which often provide a solid starting point for further steps in various industries.

In addition, suitable software tools can be used to extract subgraphs from larger knowledge graphs such as DBpedia, which can then serve as base taxonomies.[119] Furthermore, with the help of a reference text corpus and the corresponding corpus analyses,[120] it can be determined which topics within

the defined area should definitely be represented in a taxonomy. Text corpus analyses can also play an important role later on in the ongoing development and enhancement process of the taxonomy.




*Interplay between taxonomies, text corpus analysis and semantic profiling*

The process model can also make greater use of crowdsourcing methods. For example, if a suitable user interface is provided that allows each user to suggest missing concepts or labels (for example, embedded in a tagging or search dialog), or if the search behavior of users is simply analyzed using search log analysis, then, in conjunction with a suitable approval workflow, this can lead to a taxonomy that grows with user requirements and can quickly identify missing components.

How these different steps can be merged into a good recipe in a specific case is of course up to an experienced taxonomist. However, in the end the right success always depends on the existing will (also of the sponsors), the corresponding knowledge, and last but not least, on the organizational culture and its maturity with regard to more advanced methods of data management. But if you don't have taxonomists at hand it is probably a good time to

develop this role, starting with external consultants who are familiar with this profession.

## Ontology Management

Available methods for ontology management differ more than the approaches for taxonomy management. There are several reasons for this:

- The range of semantic expressivity and complexity of ontologies is much wider than is usually the case with taxonomies. In many cases, ontologies as well as taxonomies are concentrated on hierarchical or is-a relations.
- In some cases, however, the development of ontologies also has a strong focus on axioms, which goes far beyond the expressiveness of SKOS taxonomies. Axioms are statements that say what is true in the domain. For example, “nothing can be a soup and a dessert at the same time,” or “recipes with meat are not good for vegans,” or “each recipe can have at most one calorie value.”
- Some ontology management approaches bake all building blocks of the semantic knowledge model into one ontology, i.e., classes, instances, mappings, everything goes into the ontology. Other approaches are focussed only on the creation of the schema (called the ‘TBox’[121]), but not on the facts or instances (‘ABox’).
- Some ontology engineers still stick to the idea of building expert systems in the classical sense instead of supporting the Semantic AI approach (aka ‘Knowledge-based artificial intelligence’[122]), which has a fundamental impact on the design process since basic concepts of the Semantic Web like the open world assumption are not applied in this case.
- Many ontologies do not have any project goals in mind or requirements of applications that should be based on them. In order to develop universally valid ontologies (sometimes also called ‘upper ontologies’), different design principles and management methods must of course be applied than for specific ontologies that are often only relevant for a single subdomain.
- This leads to confusion, and some people believe that *the ontology **is** already the knowledge graph*.

Therefore, it is practically impossible to present a recipe for a successful ontology management that is applicable to all possible ontology projects. Nevertheless, ontology management should be based on a set of best practices[123] and design patterns, especially in a business context as a collaborative effort involving non-technical people and as part of a broader knowledge graph initiative:

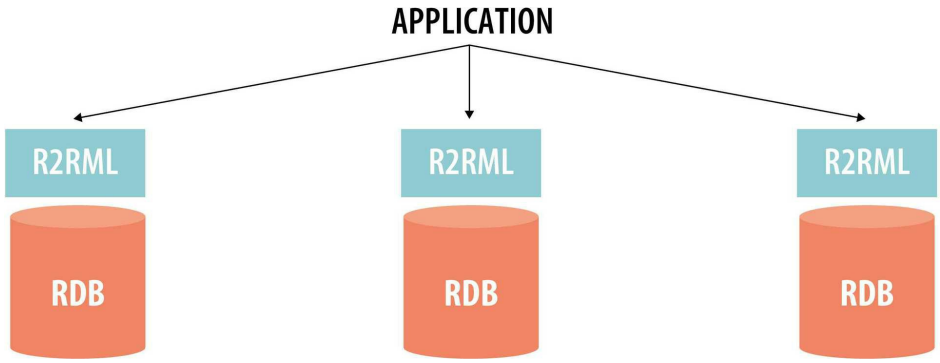
- Use non-technical terms: replace highly technical terminology with terms that are more accessible to your stakeholders.
- Define your domain: identify the subject area that the ontology describes and try to reuse public ontologies[124] with similar domains.
- Formulate measurable goals: define personas, use cases and identify exemplary content types and topics.
- Stay focused: prioritize the classes, entities and relations through the use cases and goals of the project.
- Think and develop in the form of onion rings: Start with a core ontology and first create a "minimal viable corporate ontology." Let the team celebrate its first success!
- Validate your design: show how the ontology relates to the content and information of the project stakeholders and how it helps them to achieve the goals defined at the beginning of the project.
- Stay agile: don’t boil the ocean and don’t try to come up with the ultimate ontology that will be the single source of truth. Ontology management, and the development of knowledge graphs in particular will remain an ongoing process iterating and evolving along the learning curves of involved stakeholders.

## RDFization: Transforming Structured Data into RDF

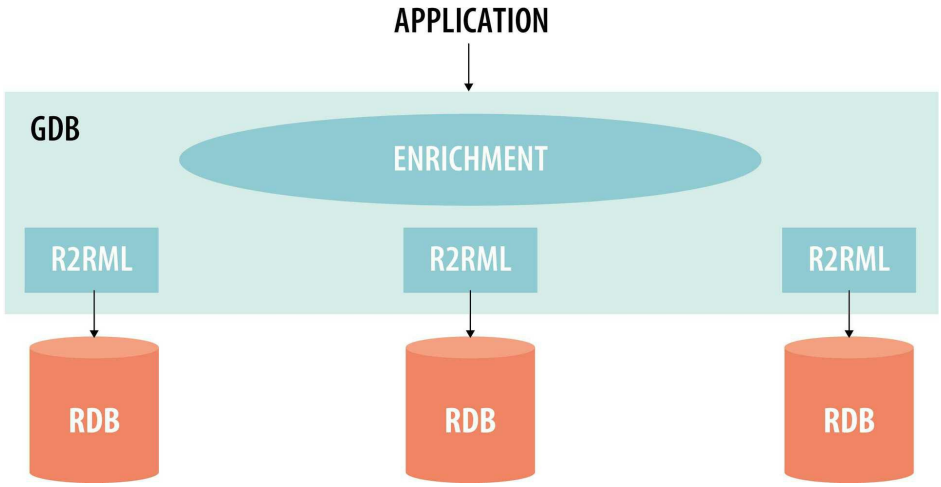
Once the groundwork is laid and we have built ontologies to provide the schema for mapping structured data as RDF, and as soon as we have taxonomies to provide controlled metadata to standardize and link entities and metadata values in our various data sources, we can start to make structured data available to the knowledge graph. In the end, there are different integration scenarios supported by different technologies.

One is the federated approach where a translation layer (R2RML) is put on

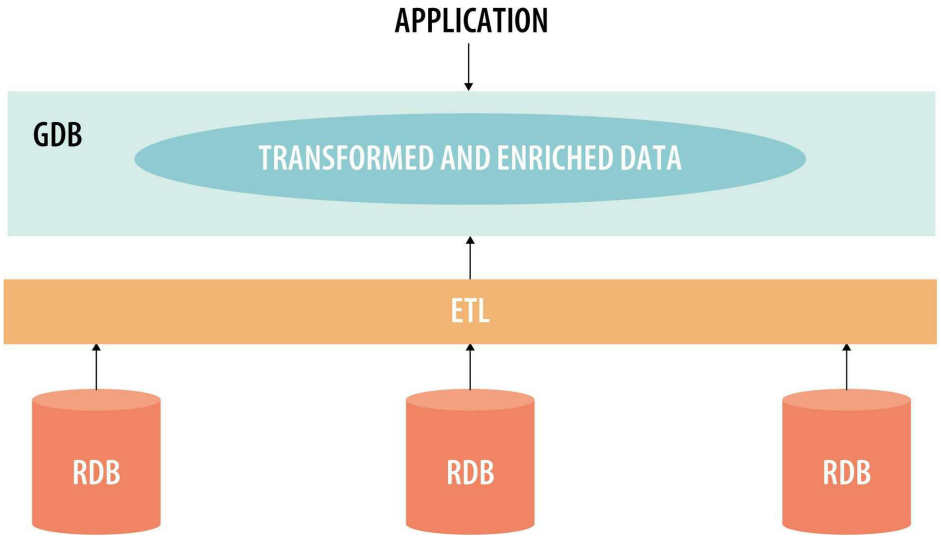
top of each structured or relational data source (RDB) to do the mapping to the ontology. R2RML the "RDB to RDF Mapping Language"[125] as a W3C recommendation is the standard to be used for this mapping. This allows us to basically access the data in real-time without the need for transformation or synchronization. On the other hand, it does not allow for mapping or linking entities to the controlled metadata layer as this additional information cannot be written back to the source. In that sense it only allows very shallow semantics and querying. It basically only translates the relational structure to RDF. Federation will also always have an impact on performance as multiple queries to different sources have to be made and combined. Nevertheless, it allows the integration and query of different sources as if they were one, which might be an option at least for very volatile data.




A second approach is the semi-federated or virtualization approach where a graph database (GDB) offers an integrated virtualization layer that allows to directly query the data in the underlying relational data source. In this case additional enrichment e.g., mapping to controlled metadata can be stored and queried altogether. This allows for more advanced semantics and more complex querying as all data can be queried as if it were in one place. Still, changes will require at least updates on the enriched information and the translation of the queries to the underlying systems will impact performance.




The third approach is the centralized or transformation approach where all data (needed) is transformed (ETL) and enriched based on the ontology and taxonomy and stored in the graph database. This is of course the most performant approach, also allowing the most complex semantics and querying. However, it is also the approach that comes with the highest costs as it requires full synchronization on all changes in the underlying sources.




The decision does not have to be made for one approach. In reality, a mixture of different approaches will be necessary depending on the intended application. The most important element of an RDFization setup for structured data is the establishment of an intelligent data management infrastructure that allows the best approach to be implemented in a timely manner and with the appropriate tools.

# Text Mining: Transforming Unstructured Data into RDF

*"THE AMOUNT OF UNSTRUCTURED DATA IN ENTERPRISES IS GROWING SIGNIFICANTLY"*

We can assume that more than 80 percent of the data in any organization is unstructured. The amount of unstructured data in enterprises is growing significantly —often many times faster than structured databases are growing.

When searching for documents and sifting through piles of unstructured data, business users and data analysts are not interested in the documents and texts themselves, but rather in finding the relevant facts and figures about the business objects in which they are interested at any given time in a particular workflow.

Therefore, users need support in extracting those passages from large volumes of text that are relevant in a particular business context. Methods of automatic text mining are an essential support, especially when embedded in knowledge graphs.

### *Entity Extraction*

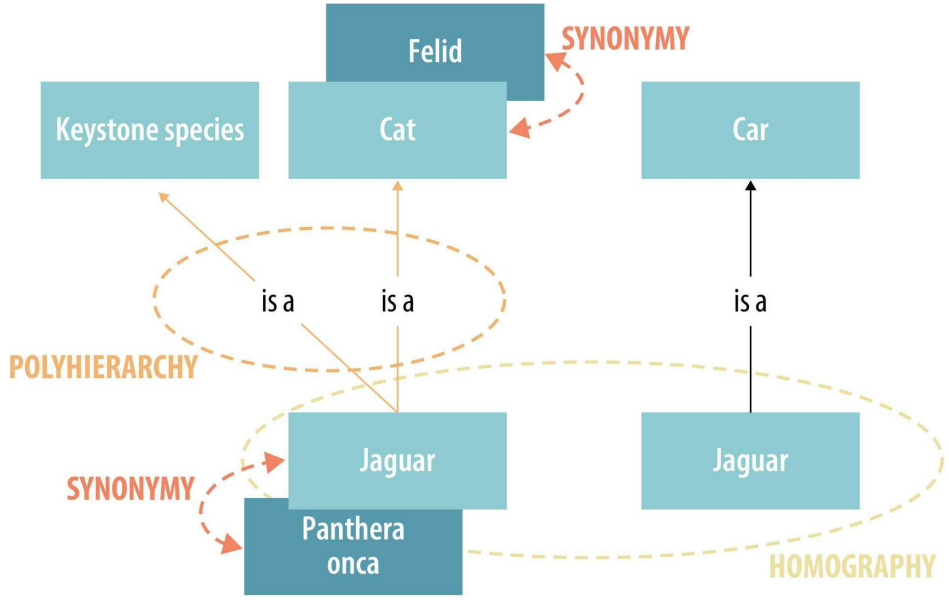
Text mining based on RDF technologies does not simply extract terms or groups of words, but rather entities from texts that refer to resources in a defined knowledge graph. A link between a text passage and a node in a knowledge graph is automatically created. This process is called a “tag event” and can be expressed and stored as a set of RDF triples.

An obvious advantage of this method compared to purely statistical text mining methods is the possibility to consider and summarize different terms that actually mean the same thing as synonyms, e.g., the sentence “the aircraft landed safely in the evening in NYC” is processed and indexed semantically equivalent to the sentence “the plane touched down for safe landing at 7 p.m. in New York.” The fact that both sentences will have the same semantic footprint can then later be exploited by recommender systems based on content similarity.

Furthermore, a knowledge graph offers the possibility to disambiguate homographs with high precision by providing additional contexts.[126] Accordingly, apples would no longer be confused with pears if, for example, the supposedly same thing appears in sentences like “an apple contains vitamin A” or “Apple has its HQ in Cupertino.”

Knowledge graphs thus help to solve common challenges of language processing (be it by humans or machines), which are often summarized under 'Babylonian confusion of language.'[127] These include

- synonymy,
- homography, and
- polyhierarchy.




*Linguistic features of a knowledge graph*

Methods of automatic extraction and linking of entities are more reliable and precise the more developed the underlying knowledge graph is. Obviously, there is a need for algorithms that can be used without knowledge graphs, whereas named-entity recognition (NER) methods based on machine learning or text corpus analysis are suitable.

This allows on the one hand, to identify missing elements in the knowledge graph automatically or to use them for supervised learning if the HITL design principle is to be applied. On the other hand, graph-based and ML-based

extraction can be combined to achieve a better F-score.

### *Text Classification*

Text or document classification is a typical machine learning task. As with most ML tasks, there is supervised, unsupervised and semi-supervised learning.

To classify text unsupervised, clustering algorithms or self-organizing maps (SOM) are often used. This approach is virtually at the other end of the AI spectrum when trying to develop a graph-based AI framework. Nevertheless, this method could be helpful in a first step, select the right documents for further steps, e.g., for corpus analysis.

Supervised and semi-supervised methods are based on a set of pre-classified documents that are used to train classifiers that normally use algorithms such as support vector machines (SVM) or linear regression.

In principle, this can be done without any semantic knowledge model, but the knowledge models are a valuable resource for training classifiers when little training data is available or for pre-qualifying documents for further use in the corresponding classifier training.

In any case, when the classification is embedded in a larger knowledge graph environment, the resulting classifier data is not just another data set, but is linked to the semantic footprint of a business object to enrich it with additional active metadata.

### *Fact Extraction*

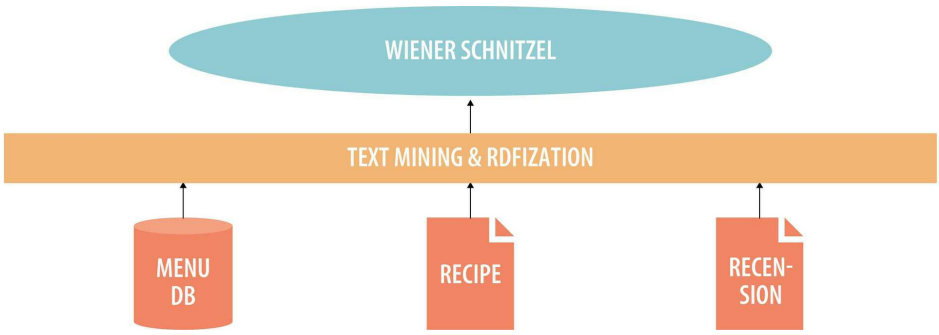
Another text mining task is the extraction of facts (in contrast to single entities) from unstructured text or also from tables that could be embedded in a document. To automate fact extraction, typically a set of fact patterns are predefined, for example, "PERSON hasPosition COMPANY" or "COMPANY ownsBrand BRAND". The goal of fact extraction (often called ‘relation extraction’) is to allow computation to be done on the previously unstructured data. This sounds like a great recipe and like a good fit for a graph-based semantic AI approach!

In essence, with fact extraction algorithms in place, sets of triples can be extracted from any given chunk of unstructured data, is it from documents or from database fields containing such. To train fact extraction algorithms, ontologies and knowledge graphs can play a central role. In return, this technology can also be used to enrich existing knowledge graphs, so-called 'link prediction'.[128]

Typical application scenarios for fact extraction are to analyze research papers in life sciences (e.g., gene-disease relationships or protein-protein interaction), to enable secondary use of electronic health records (EHRs) for clinical research (e.g., mining disease cases from narrative clinical notes), or to run automatic fact checks over news articles.

## Entity Linking and Data Fusion

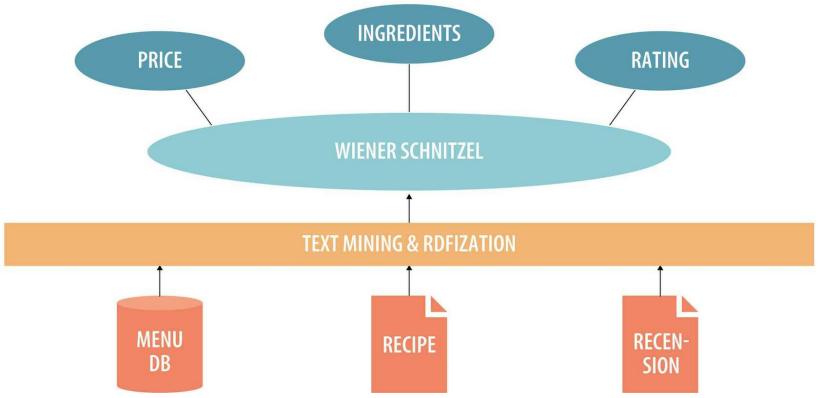
Entity linking and data fusion is the last step in mapping your structured and unstructured to your EKG. Like the crumble topping on an apple pie, it is the final kick needed for the perfect taste, but when done wrong, it can ruin all the work that you did. Let's start with entity linking. So I might have the price and variants of my "Wiener schnitzel" in the menu database. In addition, I have my secret special "Wiener schnitzel" recipe that, according to the latest newspaper reviews, is the best "Wiener schnitzel" in town.




However, this information is available from a variety of sources and cannot be accessed in its entirety. Only after I can see that all three sources refer to the same entity can I merge the information and create a value added. But I should be sure that my "Wiener schnitzel" is not mixed up with the "Apple pie." The knowledge graph provides the gold standard for training machine learning algorithms that can help to suggest the right connections with a high degree of certainty.

The better the original data quality is and the better the data is structured, the more precise the results will be. In a scenario with purely structured data, it might be possible to achieve a fully automated entity linking. Once unstructured data is in play, the situation will be different and a semi-automated approach should be adopted, with at least the approval of a subject matter expert as to whether the linking proposals are correct in case a certain threshold is not reached. In addition, regular quality checks embedded in the expert loop can be performed to detect incorrect linkings.

Once we have determined that we are talking about the same thing in different sources, we can take the next step and use different information about that thing in different sources. Data fusion is defined as the “process of fusing multiple records representing the same real-world object into a single, consistent, and clean representation."[129]




The ontology that defines the schema of the data helps here as well, when mapping structured data and also when extracting facts from unstructured data. Together with machine learning approaches, this will even enable the establishment of automated mapping of structured information and quality checks on the data itself. In addition, it will enable us to recognize more facts and information in unstructured data and make them more valuable.

## Querying Knowledge Graphs

*"SPARQL IS THE ULTIMATE MASHUP TOOL"*

So the table is set: we have ontologies based on standards that represent our various data models in a self-describing way. We have taxonomies that

model our language in its various forms and meanings. Both are used to make structured data accessible in a unified way and to get structure into our unstructured data. We have identified (real world) entities in our different sources using various entity extraction methodologies. And we have linked the entities from the different sources, which are actually the same entity, and we have fused the data of these entities to be able to consider them as one thing.

As a result, we can now retrieve all kinds of data from different systems in a uniform way. A knowledge graph of your data sources supports the access to and exploration of sometimes unpredictable and initially unknown information, thus creating new insights and values.

The SPARQL Protocol and RDF Query Language (SPARQL) is the query language for RDF based knowledge graphs and it's designed to support accessing and exploring unpredictable and sometimes unknown information across data sources. SPARQL allows you to query both the data and its description at the same time. Furthermore, queries can federate data in many silos across an intranet or across the Web. While traditional SQL databases act as barriers to data integration and Web services architectures restrict the ways in which one can interact with information, SPARQL is the ultimate mashup tool: with SPARQL one can explore unknown data and mix together whatever information is needed on the fly.[130]

In addition, GraphQL has established "a query language for your API" as another standard for graph data retrieval that is simple and intuitive. It provides a simple declarative lens for large knowledge graphs and provides developers with tools to bootstrap knowledge graph APIs.[131] Again, the ontologies can provide the schema that GraphQL needs to describe the data. The combination of SPARQL, which enables more complex queries and analyses, and GraphQL, which can set up an easy-to-use API layer for applications on top of the knowledge graph, offers completely new ways of accessing and exploiting data.

## Validating Data based on Constraints

We have seen how knowledge graphs allow us to access data from different systems in a unified way, but does this also mean that it automatically

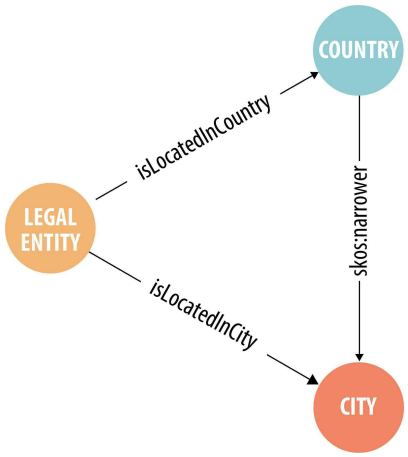
overcomes the problem of data inconsistency in a heterogeneous data landscape that has grown over the years? No, I'm sorry, it does not. The good news is that the implementation of an enterprise knowledge graph is always a means to improve data quality, as it is an initiative based on defined standards for describing data models, their structure (ontology) and the metadata used (taxonomy).

The most basic approach to validate data is using SPARQL queries. It is very expressive and can handle most validation needs for knowledge graphs. It is also available in all applications supporting the curation and validation of RDF based knowledge graphs. Downside is that writing and maintaining queries can become difficult and requires experience and expertise in SPARQL.[132]

Standard inference approaches available in most applications that support RDF-based knowledge graphs are not applicable. They are built on the open-world assumption (OWA) which limits the validation possibilities as validations of constraints are most often tailored to closed world use cases we typically have in enterprises. In this scenario, we do not want to look for things that *might* be available, but rather make sure that the data in place is consistent and complies with the defined data structure and metadata standards.

For this reason, several standards have been developed to formulate restrictions for knowledge graphs based on RDF. The latest approach, which eventually became a W3C recommendation, is the Shapes Constraint Language (SHACL).[133] A SHACL validation engine receives as inputs a data graph and a graph with shapes declarations and produces a validation report that can be consumed by other tools. All these graphs can be represented in any RDF serialization format.

Example: Consistent geographic information




**Constraint:** If a `Legal Entity` has a `Country` and a `City` assigned, then both places must be related with a `skos:narrower` path, so that the geographical information is consistent.

More and more software tools are becoming available that can translate these shapes into queries that can be used to validate data. These so-called SHACL processors improve the maintainability of constraint definitions and enable their use in a variety of scenarios:

- Validation of data consistency, which allows repair mechanisms to be built upon it
- Rule definitions for deep text analytics that allow the execution of complex analytics tasks (for example: compliance checks), e.g., in contract intelligence scenarios
- Validation rules for performing quality assurance or sanity checks, so that the quality or completeness of an (automatically) generated graph can be assessed

## Reasoning over Graphs

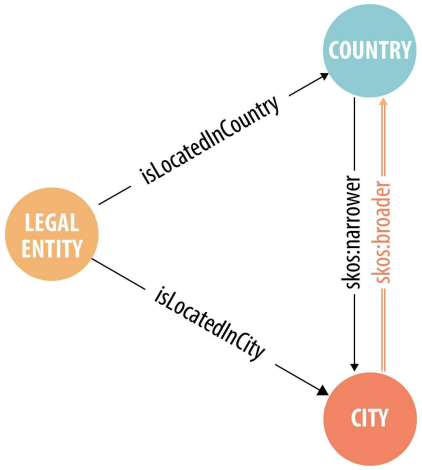
And now one could ask what else is possible. We can query our graphs and even validate consistency. But if we want to get the data in order, we need to perform other (automatable) actions to enhance our data, based on the problems found by the quality checks. Also, the SHACL standard mentioned above only includes the creation of queries to automatically check consistency from the SHACL forms, but it does not automatically create the queries to automatically update the data when a check fails. In some cases

this is not possible at all.

This is where reasoners or inference engines come in, ideally integrated with the graph database used in your company's knowledge graph infrastructure. And this is where new questions arise, because not all graph databases contain them or offer the same functionality and are thus not 1:1 comparable. Furthermore, reasoning engines are not always sufficiently performant for larger data sets.

After all, you can do two things with reasoning. First, you can add missing elements based on your ontology, which is called “forward chaining.”[134] The ontology provides the axioms or rules for the reasoning engine, which completes your data accordingly by automatically deriving the missing information.

Example: Provide completeness

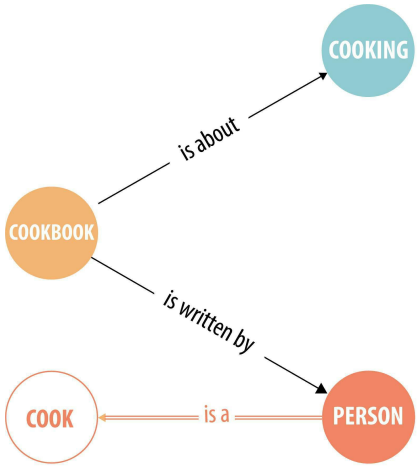



**Rule:** `skos:narrower` is `owl:inverseOf` the property `skos:broader` → `skos:broader` is automatically added as a new triple

It should be added here that applications in an enterprise scenario ideally add all information at the time of data creation, as completeness is a sign of data quality. This would also be the most performant way, since no additional measures are required. Many inference engines materialize the missing information anyway, that is, they write it to the graph database, since inferencing during query time is otherwise often a performance bottleneck.

And secondly, reasoning engines can infer new "knowledge" on the basis of existing information and given goals, which is also called "backward chaining." In this case, the given goal must be verified by the information available in the knowledge graph. This approach is not yet widely used in graph databases.

Example: Infer new knowledge




**Rule:** If a `CookBook` is written by `Person` and `CookBook` is about `cooking` then the `Person` is a `Cook`.

# How to Measure the Quality of an Enterprise Knowledge Graph

The quality of some elements, especially those of the ontologies and taxonomies used in a knowledge graph, determine the quality of the entire graph, especially the automatically generated parts of the graph, which make up a large part of the data graph. Quality has to be measured in order to make the following decisions:

- Which of the available ontologies should be used for the planned application?
- Should I improve my taxonomies, and in what respect?
- If I use this knowledge graph for my application, will I get satisfactory results?

Quality is a central factor in answering these questions. But what we want is not a “good” ontology in an abstract sense, but one that is well suited for our purposes. In other words, our goal is to measure fitness for purpose, not some abstract concept of "quality."

The following table gives an overview of the different possible aspects that can be evaluated.

| Category | Description | Remediation |
|---|---|---|
| Encoding | Does the structure and content follow general formal rules defined in the respective recommendations (e.g., RDF, XML etc.)? | Can be automated for the most part. |
| Labeling | Label issues like misspellings, inconsistent capitalization, etc. | Can be automated for the most part. |
| Design | Is the design of the ontology or taxonomy well formed according to existing standards (e.g., ISO-25964, OWL etc.) and the specifics of the domain? | Manually only, by knowledge engineer or metadata specialist |
| Correctness | Do the labels and hierarchical structure correctly model the knowledge domain? | Manually only, by knowledge engineer or metadata specialist with the help of subject matter |

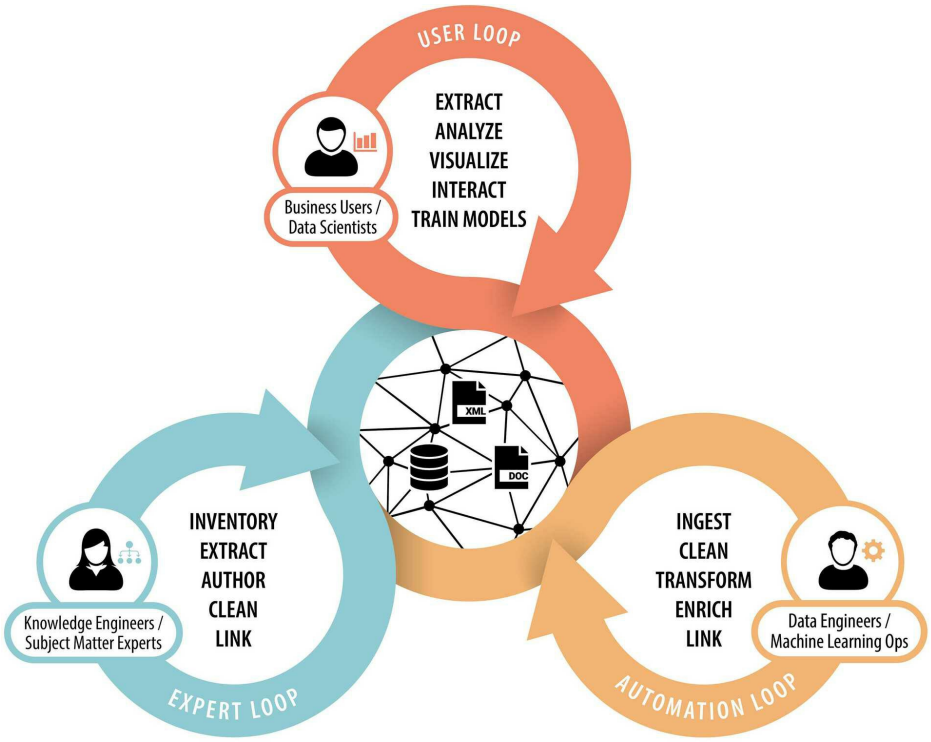
| | | |
|---|---|---|
| | | experts |
| Coverage | In the portion of the domain that has been modelled, to what degree is the model complete? | Data entry (supported by gold standards, corpus analysis, etc.) |
| Performance | Is the ontology or taxonomy fit for purpose? This depends on the particular purpose or purposes considered. For an ontology that can be expected to be formally correct (no structural errors or missing labels), this is the important question. | Remediation will depend on the property being measured. |

We would like to draw particular attention to the measurement of coverage by gold standards, which can become an important continuous quality benchmark. This is rarely done because gold standards usually have to be set manually within the expert loop by subject matter experts who manually add the knowledge graph to unstructured information or manually transform data to provide the basis for quality measurement.

However, once this has been done, the assessment can be done automatically, and if implemented as a guiding principle of governance, it can become an important benchmark for quality, but also for value creation. In many cases, the implementation of knowledge graphs replaces either the manual tagging approach or the manual transformation approach, and this makes the economic impact clear.

# Knowledge Graph Life Cycle

The enterprise knowledge graph life cycle provides an overview of the actors and agents involved during the most important operational steps for the (ongoing) development of the graph. This ranges from data inventory, extraction and curation, modeling (authoring), various transformation steps, to linking and enrichment (e.g., inferred data), and analysis or feedback of the newly acquired data into existing database systems. In reality, there are three cycles that are intertwined: the expert loop (see HITL), the automation loop, and the user loop.




*The KG Life Cycle*

A solid foundation for the creation of high quality data graphs can only be established if sufficient time is invested in the creation and maintenance of curated taxonomies and ontologies, but even these steps can be partially automated. Within the loops, agile and iterative working methods are predominant, whereby individual process steps can interact with each other.

In summary, the knowledge graph life cycle points out the following aspect:

1. The development of knowledge graphs is an endeavour involving

*several stakeholders***.**
2. Developing knowledge graphs means to proceed *iteratively and agilely*, not linearly.
3. Humans *and* machines should be equally involved in building an enterprise knowledge graph.
4. The knowledge graph is constantly being developed further in *three loops that are linked together*.
5. The aim is always to *balance the three most important perspectives* on the knowledge graph: representing domain knowledge, linking company data, and enriching it with user contexts.

## Expert Loop

The Expert Loop involves predominantly knowledge engineers and subject matter experts working on ontologies and taxonomies to be further used by the other loops. Here are the main tasks:

- ***Inventory***: run scoping sessions with business users and SMEs using card-sorting and taxonomy tools combined with automated analysis of selected content and data sources to determine which areas of interest, in combination with which data sets, are important for getting started.
- ***Extract***: extract relevant types of business objects, entities, and topics from identified data sets and put them into the individual enterprise context and link them to specific application scenarios.
- ***Author***: in several iteration steps, develop a viable ontology and taxonomy architecture, which can, for example, consist of several core ontologies and department-specific taxonomies. At the same time, harmonize the associated governance model with the organizational culture and the overall KG governance model.
- ***Clean***: curate suggestions from ML-based tools like corpus analysis. Clean up and adapt taxonomies and ontologies that are reused in the specific organizational setting.
- ***Link***: using ML algorithms, links between entities and concepts from different graphs, mainly between taxonomies, are curated and created.

## Automation Loop

Data Engineers and MLOps are responsible for all matters within the Automation Loop.

- ***Ingest***: retrieve data from defined sources and ingest data generated within the user loop for further processing, track provenance and provide data lineage information including technical metadata involving data transformations.
- ***Clean***: clean data from various sources with help from ontologies and corresponding consistency checks automatically.
- ***Transform***: with knowledge graphs in place, most of the ingested data and metadata can be transformed into RDF-based data graphs. Transformation steps follow the rules expressed by domain-specific taxonomies and ontologies.
- ***Enrich***: automatic entity extraction and lookup in knowledge graphs for context information help to enrich data points automatically. Additionally, powerful inferencing mechanisms by using ontologies and constraint languages like SHACL enrich enterprise data sets.
- ***Link***: linking on entity level, not only schema mapping, will generate a rich enterprise knowledge graph. Machine learning and algorithms such as spreading activation can automatically generate links between several graphs and data sets automatically with high precision.

## User Loop

As beneficiaries of the knowledge graph, mainly business users and data scientists interact with the data within the User Loop, but not only as passive users but also as data producers:

- ***Extract***: using digital assistants or more basic filtering methods such as faceted browsing, business users can extract even small chunks of information or single data points from large data sets precisely and efficiently. Graphs are the key to unlocking the value of large data sets by helping users to narrow down the search space based on individual information needs.
- ***Analyze***: graphs and query languages as SPARQL provide additional means for powerful data analytics and also help to lower the barrier

for user self-servicing complementing traditional data warehouses and their rather rigid reporting systems.

- ***Visualize***: business users benefit from linked data, especially when visualizing relationships between business objects and topics. This can be used to analyze causalities or risks in complex systems, to identify hubs in social or IT networks, or just to better understand how things relate in a knowledge domain, etc. But enterprise data modeled as graphs do not necessarily have to be visualized as graphs, but rather serve as a flexible model to present and interpret data in a more individual way than would be possible with rigid data models.
- ***Interact***: users in such systems are also data producers when they interact with the knowledge graph. While they benefit from comprehensive guidance through extensive data landscapes, users also provide feedback on the overall system and their behavior can be used to further enrich the knowledge graph.
- ***Train models***: data scientists can better filter and reuse data through semantically enriched metadata. Relevant data sets can thus be quickly extracted from data catalogs and used specifically for training ML algorithms. Data enriched and linked with knowledge graphs also have a higher expressiveness and are suitable, for example, for the training of classifiers even if only smaller volumes of training data are available.

The majority of technology platforms used in the development and implementation of enterprise knowledge graphs are specialized in one of the three loops. As a result, only special applications based on graphs can be implemented. Only the right mix and a balanced interaction of the three loops can support a long-term knowledge graph vision and strategy of a company. With the expert loop in the game, which interfaces with the automation loop, every AI system based on knowledge graphs automatically becomes an explainable AI.

# Good Practices Based on Real-World Use Cases

Finally, we want to share some experiences from cooking different dishes with different clients across different domains. A general observation, even when the domains differ, the problems and the solutions are the same. Another important fact is that a knowledge graph is never finished or perfect. It is a living thing that grows and changes over time, as knowledge (hopefully) grows and changes over time.

## Start small and grow

Or even better, “start small and grow” based on concrete use cases and examples to show the value the knowledge graph can bring to your organisation early. “Effective business applications and use cases are those that are driven by strategic goals, have defined business value either for a particular function or cross-functional team, and make processes or services more efficient and intelligent for the enterprise. Prioritization and selection of use cases should be driven by the foundational value proposition of the use-case for future implementations, technical and infrastructure complexity, stakeholder interest, and availability to support implementation.”[135]

Do not start with defining the perfect final data model (ontology) for all your data. You will find yourself 10 years later when you think you are finally done and realize that the data model does not fit your use cases. Do not try to define the perfect taxonomy or controlled vocabulary. When you are done many new things are there and need to be included and you might find that while you did model a nice taxonomy, it does not fit your data.

Personas can be used to define the users of your knowledge graph and the application based on it, to specify their expectations, requirements and needs. If you have already done this, you can build your knowledge graph from the beginning to meet the needs of your users. Don't forget to involve them in the process as soon as possible, either actively or to review the results. Develop prototypes and make them ready for production step by step, but do not hesitate to throw them away if they do not work. Learn your lessons and be agile, because knowledge graph development is best done in an agile mode of data management.

## Get to know your data

Before you start to develop your ontology you should have a good overview of your data landscape. “There are a few approaches for inventorying and organizing enterprise content and data. If you are faced with the challenging task of inventorying millions of content items, consider using tools to automate the process. A great starting place we recommend here would be to conduct user or Subject Matter Expert (SME) focused design sessions, coupled with bottom-up analysis of selected content, to determine which facets of content are important to your use case.”[136]

You will have structured as well as unstructured data in various forms and sources. There are differences in working with structured and unstructured data and the final goal is of course to bring both together. So think early on of setting up a data catalog as one access point that describes your different data sets and of course, use your knowledge graph to describe data sets in your data catalog.

Next, wisely choose some first data sources for your prototypes that:

- come from both sides (structured/unstructured) so you learn to work with different kinds of data.
- are not too volatile so you do not have to begin dealing with synchronization.
- are not too big so you do not have to begin dealing with performance.
- and last but not least, show the benefit by choosing data sources that when connected can do/show something that was not possible before.

## “Not invented here!” is not a good practice

So if you now start to develop your taxonomy and ontology for your well-defined small prototype, knowing very well what data will be used, what would be the first thing to do? See what is already there! There are already a lot of great taxonomies and ontologies out there for different domains, commercial and non-commercial. In the first step you should evaluate what you can reuse before you build something new.

On one hand that of course saves time. You do not have to build things on

your own. That might save a lot of effort and money. Second, if you reuse what others already use, it will be easy to connect your data with other data out there. The high art of knowledge graphing will be to connect your enterprise knowledge graph to others out there to bring in additional knowledge and by that additional value. And finally, re-use in semantic knowledge graphs is not static because semantic knowledge graphs are built to be extendable. Even though the taxonomy you have found may not be perfect, it is a good start to build on it and extend it or tailor it to your needs. You will have to do this anyway because nothing you find will “perfectly” fit your needs. The same goes for ontologies, where you can just pick the parts that are relevant and combine or extend based on your own needs. This applies both to domain-specific ontologies and to domain-agnostic, so-called "upper ontologies".[137]

Of course, there are limits to this, especially if you can't and don't need to reuse and combine the whole ontology without changes. In that case you should think about some basic limitations:

- **Reuse of existing ontologies**
  Reuse is, of course, the easiest option, and standards-compliant, so that your data is well connected to all other data using the same ontology. The biggest drawback is that you have limited control over the ontology.
- **Create your own (independent) ontology**
  Of course, this offers full control over your ontology, but you have to model it yourself, which requires the right skills and takes time. Also, your data is not connected to other data right away.
- **Create subclasses/properties for an existing ontology**
  Prerequisite is that reused ontologies also allow reuse, with that, you inherit the structure of the existing ontology and you are integrated with all other data based on this ontology. However, here you have to accept the structure of an existing ontology, which again could be restrictive.
- **Apply both, existing and own, ontologies**
  That way you can make use of the benefits of both sides. However, you also have to take care not to run into conflicts and querying might become complex.

So you can choose the right option based on your use case. In general, a good piece of advice is to follow one principle and set up a governance process around your ontology and taxonomy management process. The design of your ontology especially affects the retrieval of data in your smart applications and it will also affect the use of external data, because other people may use existing ontologies in a different way.

## URI Patterns: Put your Knowledge Graph on a Solid Foundation

No, this is not about the system architecture and the tools. It's about something much more fundamental, and yet you will forget it as soon as you read it. It's about URI patterns. When we talk about knowledge graphs, we are talking about knowledge graphs of the Semantic Web, and that means that URIs and triples are the basic elements of our knowledge graph. And so, from the beginning, we should make sure that these URIs are well constructed and meaningful so that we don't end up in chaos. A good URI scheme guarantees that URIs are maintainable, consistent and simple.[138]

That conflicts with the intention to have expressive URIs that tell us already what things are about. But what do you do when names change? Here are some basic guidelines that proved to be meaningful in practice. You should distinguish between different types of URI schemes.

All types can start with the same domain, but it would be good to be able to determine the respective type e.g., from the subdomain. That would result in different "baseURIs." For example:

- https://data.domain.org (e.g., for data sets, named graphs)
- https://resource.domain.org (e.g., for documents, data sets, persons)
- https://vocabulary.domain.org (e.g., for controlled vocabularies, taxonomies)
- https://schema.domain.org (e.g., for data models, schemes/ontologies, defining the structure of resources or extensions to vocabularies)

Add an individual name for the data, resource type, vocabulary, ontology, for example:

- https://data.domain.org/hr-records
- https://resource.domain.org/document
- https://vocabulary.domain.org/skills
- https://schema.domain.org/geo

And add an identifier for each entity at the end wherever the identifier can be provided, for example:

- from existing IDs in a system the resource is created from.
- from predefined patterns, e.g., incremental, UUID etc.
- from label or name of resource, but be careful here since labels change and URIs should stay stable. This is only recommended for things that are very unlikely to change.

These are the most basic patterns. Of course, you can add things in between to provide additional information, but as always, less is better. What you should never do is to include things that will definitely change, for example:

- version numbers
- dates and times
- prices
- etc.

Why all the fuss about that topic? As soon as you start to use your knowledge graph to enrich your information, URIs will be everywhere. So when you have to change them, you have to change them everywhere, and that will come with cost. In addition, URIs should be resolvable in an ideal "Semantic Web" world. That means when you look up the URI in a browser or with a software agent, you retrieve a description of the resource that is identified by it.[139] Many people do not initially see this as a valuable feature. But it is a fact that your knowledge graph becomes self-referential and thus self-explanatory. This will support reuse within your organization or even across organizations. Again, this will work only if you have implemented a meaningful URI schema from the beginning.

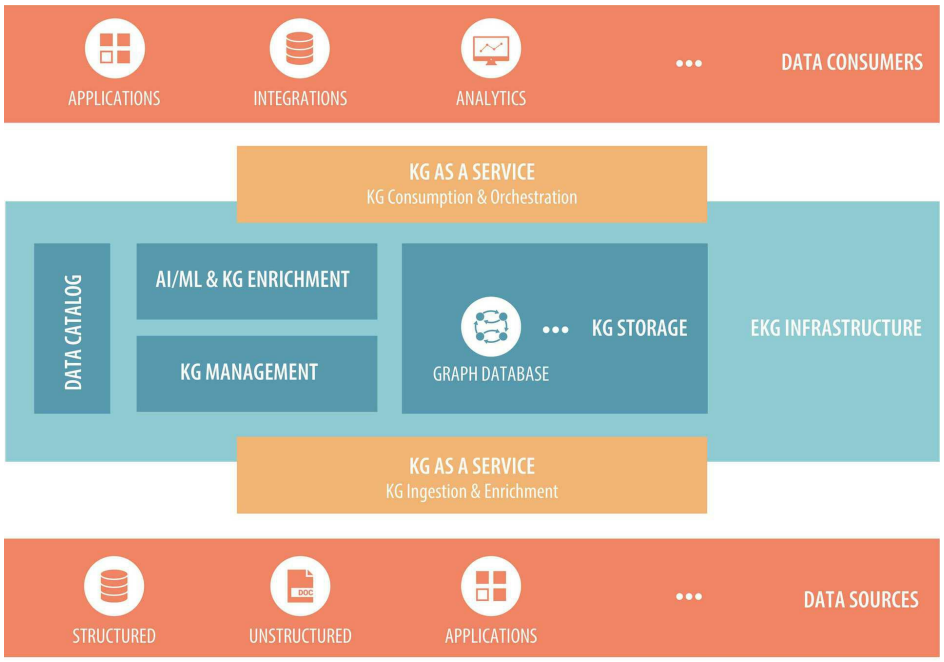
Now that we have said all of this—you will immediately forget (or not) about this recipe.

# PART 4: SYSTEM ARCHITECTURE AND TECHNOLOGIES

## A GREAT CHEF IS FIRST A GREAT TECHNICIAN

# Elements of an Enterprise Knowledge Graph Architecture

Is there such a thing as the “one” Enterprise Knowledge Graph architecture? Hmm, not really, but certain components or elements are always present in the core. More components will be added iteratively based on experience. As a non-intrusive technology, an Enterprise Knowledge Graph architecture must be able to fit into existing enterprise system architectures. The following diagram shows which building blocks are important and how they fit together.




*Enterprise Knowledge Graph Architecture*

As you can see, we make a distinction between the infrastructure of the enterprise knowledge graph and the service layer, which allows us to add data to the infrastructure of the knowledge graph and integrate the knowledge graph into your existing system architecture. As already mentioned, this architecture can be adapted or extended on a case-by-case basis, we will outline some typical scenarios. Based on this, we will talk about the knowledge graph as a service and what can typically be expected from this service layer.

In the previous chapter we described in detail methods for developing and

managing knowledge graphs in organizations. We have outlined how AI/ML and knowledge graphs interact to support different application scenarios, and how this finally leads to explainable AI systems. Now it is time to talk about technologies and infrastructure.

EKGs are not like a traditional data warehouse approach, where you put everything in one place and just make it available from there. It is a multimodal approach, where the goal is to combine data according to the situation and make it available in the best possible way. Since knowledge graphs are the key to agile data management in companies, the knowledge graph architecture implemented in the company must support this scenario. It must also offer the possibility to deliver the right data in the right format in a timely and high-performance manner. The knowledge graph architecture must therefore provide support in the following situations:

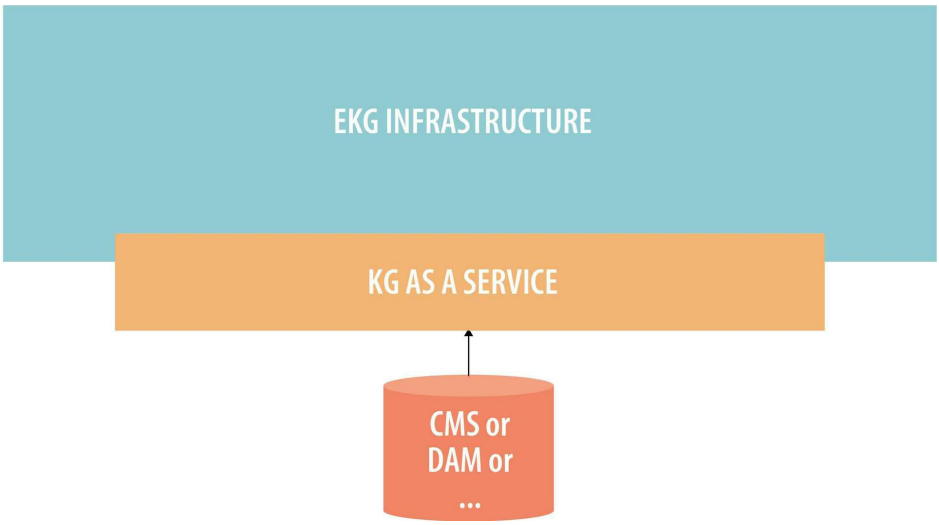
- knowing what data you have and where it is
- knowing how to bring your data into the right format for your application
- using the right database model (graph database, indexing engine, relational database, etc.) and methodology for your use case in terms of efficiency and performance
- providing easy access to your data for users and developers
- supporting the combination of different environments, on-premise, cloud, and hybrid

# Integration Scenarios in an Enterprise Systems Architecture

As soon as your enterprise knowledge graph is made available as a service in your organization, the integration into your existing Enterprise System Architecture (ESA) should be largely standardizable. However, let us first outline the typical integration scenarios.

## Single source integration

The first option is to integrate directly within the existing ESA, e.g., a tagging integration directly into your CMS or DAM, whereby the annotations and semantic metadata are then also stored within these systems. In this scenario, therefore, usually only one existing system is involved.



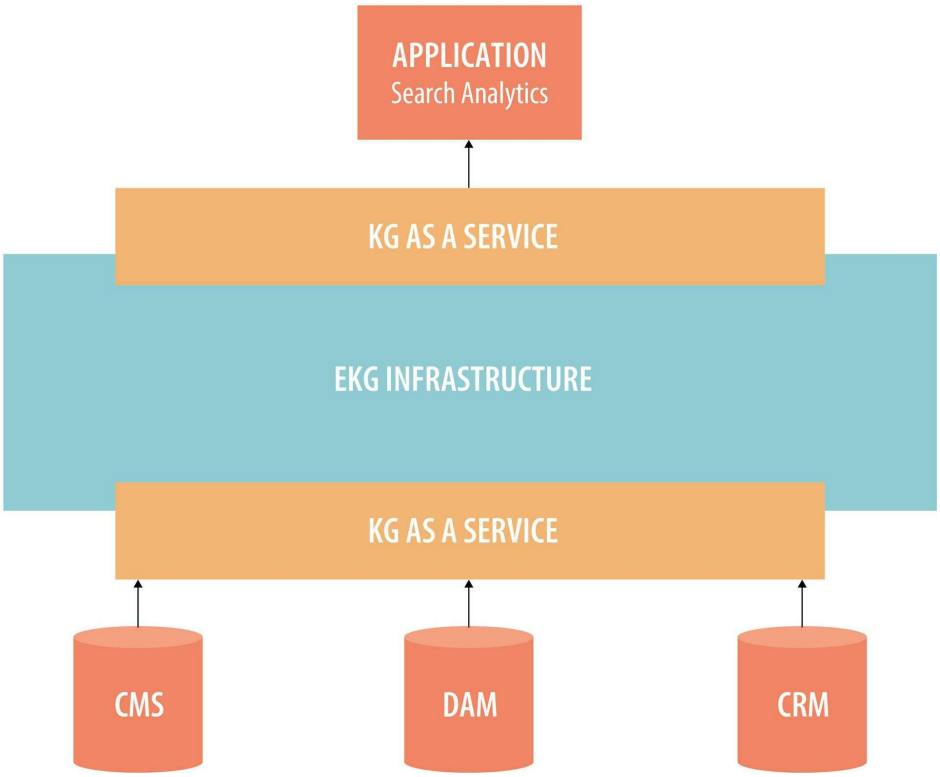

On the one hand, enrichment takes place close to the source in this scenario, which requires the least synchronization and data lineage effort in the event changes are necessary. Furthermore, existing security and access systems of the integrated systems are used and do not need to be further adapted. The semantic enrichment is therefore stored directly in the integrated systems, which are mostly based on relational DB systems.

On the other hand, this scenario therefore supports the use of the advantages of a knowledge graph only to a limited extent. In addition, all enriched metadata is in turn locked into one system and you still have to make

additional efforts to connect to other systems in your infrastructure. Since this scenario involves integration into an existing system, this can even lead to far-reaching organizational issues, since the existing infrastructure must be changed and adapted.

## Multi-source integration

The second option is to integrate with various systems in your ESA and store the results in your company-wide knowledge graph infrastructure. This of course, means that you have to think about synchronization and consider the security and access policies of all integrated systems. It also means that you cannot simply write the results (e.g., inferred data) back to the original systems without taking further integration steps.

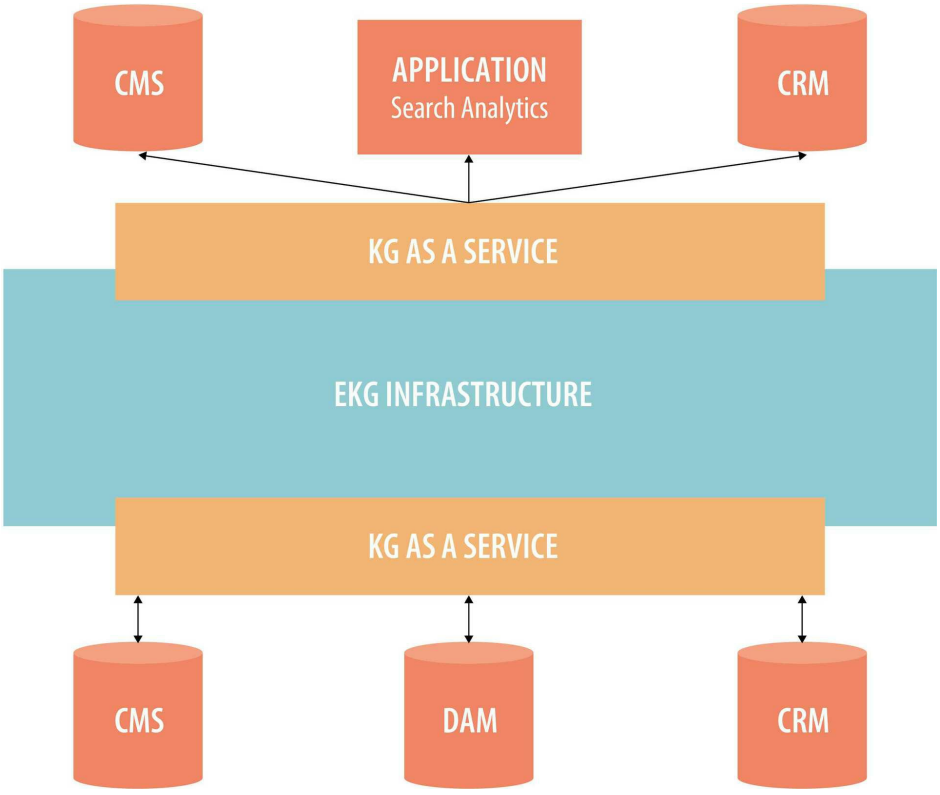



On the other hand, you can now take full advantage of the capabilities of your knowledge graph in the applications you provide and aggregate information from different sources and combine it into a unified view of the data. So while the integration effort in the different systems is lower in this case, the effort will mainly flow into synchronization and access management on the side of the knowledge graph infrastructure.

## Full integration with an ESA

What you ultimately want to achieve is, of course, full integration into your ESA, which can combine the advantages and disadvantages of the two approaches outlined above.

Do integrations directly at the source to avoid complex synchronization processes wherever possible. If you perform these integrations correctly, you will keep the information in the systems and that part of the information you include in the knowledge graph up to date at the same time. In addition, you include data in your knowledge graph so that you can make full use of the knowledge graph for search and analysis and even for federated searches across these integrated systems. And finally, you bring the results back into these systems and build any application, analysis dashboard or semantic AI application on your enterprise knowledge graph infrastructure. The fine art of cooking with knowledge graphs!




But here the same also applies: Just start and expand your cooking skills. It is therefore strongly recommended that you first try out the first two approaches, e.g., in the context of a PoC, when you start implementing your

knowledge graph in your enterprise system architecture and learn from this experience.

# Knowledge Graph as a Service

As you have probably already noticed, in order to outline the architecture of the knowledge graph we do not start with the description of the technical infrastructure for an enterprise knowledge graph. Why? That would be like arranging the ingredients for your delicious meal, but without planning in advance what different dishes you would like to cook for your menu.

The infrastructure for the knowledge graph must be provided as a semantic middleware for the system architecture of your company; therefore, the planning and conception of the services to be provided is crucial to ensure that the knowledge graph flavor is cooked to the liking of all stakeholders. If your knowledge graph initiative is to be successful, the knowledge graph must be easily accessible, and integrations should be done via standard service interfaces so that all your developers and data engineers can understand and easily work with it within the automation loop of the knowledge graph life cycle.


So what are the typical services needed? Let's group them into the following categories:

- KG ingestion services
- KG enrichment services
- KG consumption services
- KG orchestration services

And let's not forget that the knowledge graph consists of ontologies, taxonomies and the data graph. All of those components have to be made

available by different services and will play different roles in your integrations.

## Knowledge Graph Ingestion Services

In this section we sum up all services that are related to getting data into the knowledge graph or connect data to the knowledge graph. Let us begin with the services that allow you to connect to the different data sources available in your ESA:

- Structured data like relational databases, Excel and other spreadsheets, XML, etc. can be transformed into RDF as outlined in the chapter "RDFization: Transforming Structured Data into RDF". The key is to use standards like R2RML to connect relational databases, but traditional methodologies like XSLT are also of use here. Again, the key is to make it easy to set up those connections and provide services that allow us to do so.
- Unstructured data in file systems or CMS etc., has to be made available in the simplest case to be sent for tagging (enrichment) or to be broken down into structured data by making unstructured data structured using the document structure as an outline.
- In addition, connectivity to APIs of existing applications in your ESA or external services to fetch or link data must be made available.

For all those ingestion services, access to the ontology providing the conceptual model to map the data to is crucial. Services that expose the ontologies to be used for mapping data manually and ML algorithms that help to automate the mapping are needed to make this task as efficient as possible.

## Knowledge Graph Enrichment Services

Once the data has been made available it has to be enriched by and linked to the knowledge graph. Therefore, extensive enrichment services have to be put in place to sufficiently support the following enrichment and linking tasks for structured and unstructured information. These services include the following:

- term extraction
- concept-based tagging
- named entity extraction
- content classification
- relation and fact extraction
- sense extraction
- rules-based extraction
- entity linking

In addition, enrichment and linking will reveal problems in your data, so cleaning services should be included that allow you to indicate or even fix those problems. Some of those services are based on the knowledge graph, some of them will use ML algorithms. So it will be important to make taxonomies and ontologies available via service endpoints to support the enrichment process, but also to feed back into the knowledge graph information that is gathered during the enrichment phase (e.g., suggestion of new concepts to extend taxonomies or new entity types for the extension of ontologies).

In return, the enriched content can then be used as a gold standard to validate taxonomies and ontologies, to train ML algorithms, and to provide statistical models that allow to improve the enrichment. If everything is set up correctly, you have a supervised learning system that will continuously improve over time and with the right services in place it is fully integrated into your ESA.

## Knowledge Graph Consumption Services

Once ingestion, enrichment and linking is done, we can make use of our enterprise knowledge graph for integration into the EAS by making the knowledge graph available for the following reasons:

- data integration
- data virtualization
- data services
- graph analytics
- semantic AI

The key concept here is to make taxonomies, ontologies and the data graph available via API, as for example via SPARQL endpoint, to expose them as glossaries, for navigation, or to build analytics dashboards. Data access for integration, virtualization or services can also be made easier by exposing the knowledge graph via GraphQL, for example, to make it available for all systems in the ESA in the formats needed, which will again include structured formats, e.g., SQL, or unstructured formats like Word or PDF. RDF as a data model for knowledge graphs can very easily be transformed from and into any format needed.

In addition, complex graph analytics and conversational AI applications like semantic chatbots, etc., can be built on top of graphs, which require services such as the following:

- distance calculation
- similarity & recommendation services
- query expansion
- search suggestions
- faceted search

All of those services should allow existing applications to integrate into and complement the ESA or to build new search and analytics applications.

## Knowledge Graph Orchestration Services

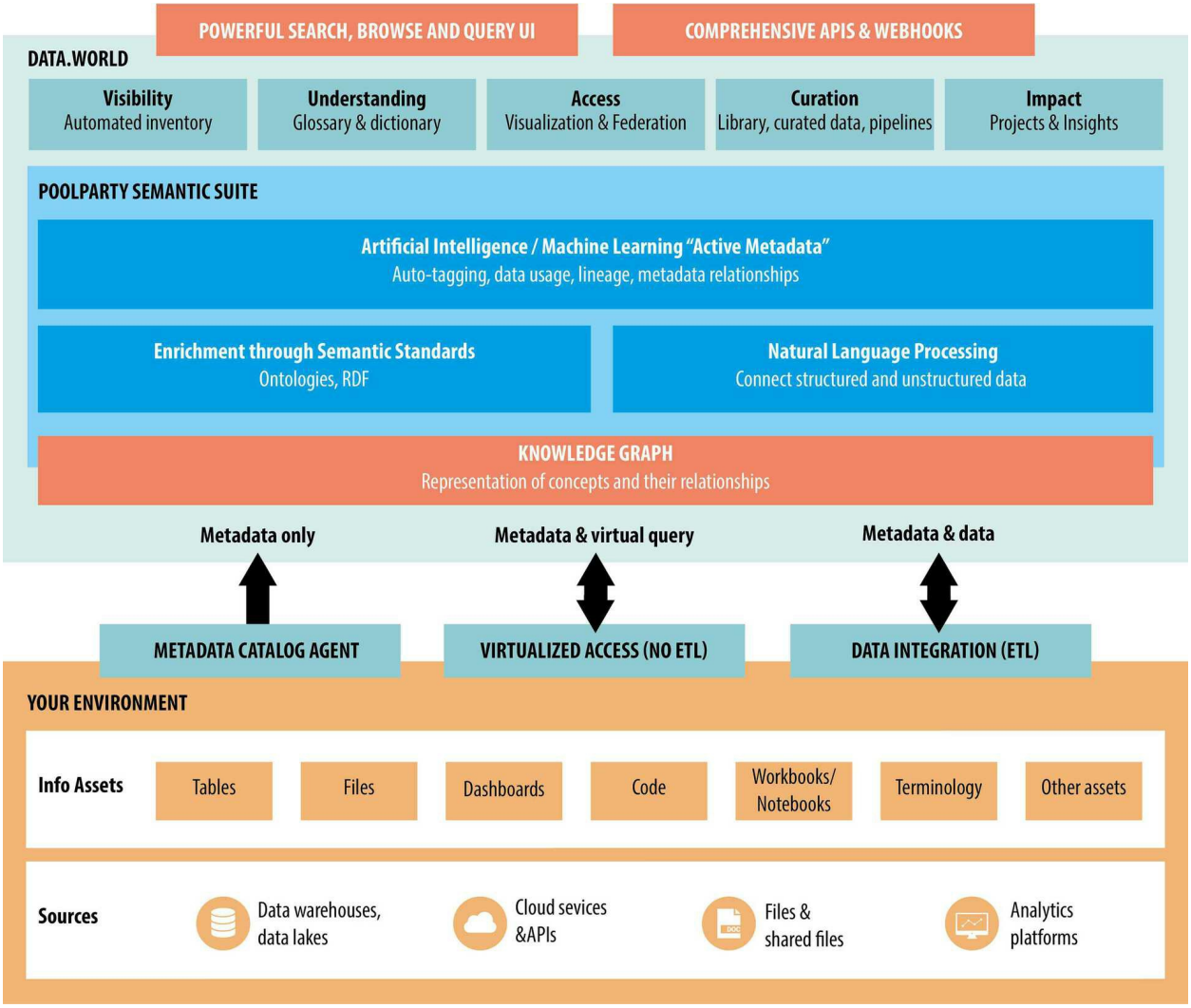
Orchestration services support automation of most of the tasks and services made available by the previous categories, so a flexible and adaptable automation platform for graph data should be introduced into the service layer that allows large amounts of data to be processed. Such orchestration services or a platform like that should support the following:

- typical features of classical ETL frameworks, including
    - connectivity to different sources: file systems, structured data, APIs, databases, and so on;
    - a plugin-system to easily develop plugins using services from the other categories of knowledge graph services;
    - the modeling of complex workflows including constraints

(criteria), looping (repeating), branching, and grouping of processes; and
    - the logging details of execution of processes including scheduling, re-running when failed, and limiting execution duration.

- native support of graph data (RDF) and ontologies
- user interface for the administration, debugging, scheduling and monitoring of the running processes
- API to integrate orchestration services
- notification system that notifies on status, and outcome of scheduled processes
- operational resilience, high scalability and performance

# A Semantic Data Catalog Architecture




*Architecture of a Semantic Data Catalog*

A data catalog can be described as "a metadata management tool designed to help organizations find and manage large amounts of data—including tables, files and databases—stored in their ERP, human resources, finance and e-commerce systems."[140] Data catalogs should also improve users' understanding of available sources and support them with collaborative workflows around data quality management to ultimately get maximum value from the organization's data.

A semantic middleware complements a data catalog with its ability to create and manage knowledge graphs and use them for sophisticated metadata

enrichment and classification tasks. These are based on text mining (entity extraction) and inference mechanisms. Seamless integration is enabled when both components are based on graph standards such as RDF. The diagram uses an example architecture to illustrate how the interaction between enterprise data, metadata, data catalog, semantic middleware and the knowledge graph works.

Moving beyond the concepts, an organization can address their data needs through the following prescriptive approach to data maturity:

1. Build an understanding and inventory of your data assets. This mapping of your data landscape is an essential first step to understanding, and the primary function of a data catalog.
2. Get everyone speaking the same language. Understanding the concepts and lexicon for your data landscape is essential for communication and effective decision making. This is the purpose of your business glossary.
3. Create a similar lexicon for your machines. Once you have the participants speaking the same language, you also need to make this understanding available to your tech. The ability to translate understanding between humans and tools (and between tools) is achieved through the use of taxonomies and ontologies.
4. Mine your assets. Now that humans and tools are all able to grasp the concepts within your data, you'll want to use technology to enrich this knowledge and fill in the gaps. This can be achieved through natural language processing (NLP).

Once you have followed these steps you will be converging toward a robust and scalable enterprise knowledge graph.

# Graph Databases

*"HUMANS SELDOM CONSTRUCT THEIR PERCEPTION OF THE WORLD WITH TABLES AND FOREIGN KEYS THAT LINK THEM"*

Why do we need graph databases for our enterprise knowledge graph architecture? I hope this has already been answered in the previous sections, but let us recap: we want to build knowledge graphs and knowledge is mainly about mapping and processing relations between entities that represent real world objects. Graph databases are closer to the functioning of the human brain and the ways in which human thinking generates meaning from the data. So the need for graph databases in an enterprise knowledge graph architecture is given.

Why can't we just stick to our relational databases? Well, on one hand, relational databases are not entity oriented but table oriented and consequently, they are not designed for a graph data model that mainly consists of nodes representing entities and their relations. “A relational model is a poor fit for real-world domains, where relationships between entities are numerous and semantically rich.”[141] Put in another way: humans seldom construct their perception of the world with tables and foreign keys that link them.

A relational data model reduces the flexibility and agility of data modeling and it does not support an agile data management approach very well as a result. Still, there will remain many application scenarios where a relational model is a good fit. And relational data can be easily used as if it were RDF using methods like R2RML that can also be implemented in data virtualization tools to provide direct access to a relational database. In addition, some triple stores sit on a relational data model and use it as the foundation of their graph architecture, such as Oracle, which provides an enterprise-ready RDF module.[142]

There are different types of graph models and graph databases,[143] but only RDF-based graph databases (also called “triple stores”) and labeled property graph (LPG) databases are widely used to develop knowledge graphs. The

main differentiation here is that RDF based graph databases sit on top of W3C recommendations and are standards-based as a result.

“Being a graph database, triplestores store data as a network of objects with materialized links between them. This makes RDF triplestores the preferred choice for managing highly interconnected data. Triplestores are more flexible and less costly than a relational database, for example.”[144]

In contrast to RDF triplestores, labeled property graph databases have been developed by various companies that have all implemented their own schemas and query languages and are not standards-based in so doing. Things are changing here since the Graph Query Language (GQL) initiative started out to create a standard for property graph databases, but that work is not yet final.

Another differentiator is that property graph databases allow us to add metadata to triples in a straight-forward manner. In RDF, this can be done via so-called “reification”[145] or partly using named graphs. In both cases one may argue that this might be complicated. Also, that aspect is changing, as there is ongoing work to accommodate this issue with RDF* and SPARQL*. [146] But, more importantly, what can we learn and expect from these initiatives? Both types of graph models have their pros and cons and both sides now have ongoing initiatives to work towards each other to remediate their cons.

So what should you use for the development of your knowledge graph in your enterprise architecture? Gartner summarizes this as follows: “A knowledge graph is unified information across an organization, enriched with contextual and semantic relevance across the silos. It combines capabilities of graph data stores with a knowledge toolkit for data unification and provides a holistic view of the organization’s data through relationships. Knowledge graphs are built on a graph data store with an RDF-based data model.”[147]

For the development and maintenance of enterprise knowledge graphs, an RDF-based graph database is the better option. RDF and the Semantic Web were designed to tackle data integration efforts at scale based on a fully-fledged interoperability framework based on standards. If this can work on the web, it can most likely also be implemented in enterprises. Still a property graph database might be a good choice for various analytics use cases, while transformation and interfacing between both types of graph data

could also be part of an agile data management approach to be implemented in your enterprise knowledge graph architecture.

As one of the core elements of an enterprise knowledge graph architecture, operational ACID-compliant graph databases typically consist of (some of) the following components:

- **SPARQL engine**: full SPARQL 1.1 support, typically including support for GeoSPARQL[148]
- **Reasoner**: typically forward-chaining reasoning for RDFS and OWL 2 profiles such as RL and QL
- **SHACL processor**: Shapes Constraint Language (SHACL) validation
- **Built-in machine learning**: predictive analytics, automated recommendation, etc.
- **RDF API**: support of either RDF4J[149] or Apache Jena[150]
- **Security model**: triple level security
- **Administration interface**: manage repositories, user accounts and access roles
- **Connectors**: connectors to SQL, NoSQL databases and indexing engines
- **Scalability**: automatic failover, synchronization and load balancing to maximize cluster utilization

# PART 5: EXPERT'S OPINIONS

## VARIETY IS THE SPICE OF LIFE

# Interviews

**The creation of knowledge graphs is interdisciplinary. Good chefs regularly visit other restaurants for inspiration. We have asked experts working in the field of knowledge graphs and semantic data modelling to comment on their experience in this area. They have worked with various stakeholders in different industries, so that you, dear reader, may further develop your understanding of the topic.**

## Jans Aasman (Franz)

**Dr. Jans Aasman is CEO at Franz Inc., a leading provider of Knowledge Graph Technologies (AllegroGraph) and AI-based Enterprise solutions. Dr. Aasman is a noted speaker, author, and industry evangelist on all things graph.**

*"KNOWLEDGE GRAPHS AREN'T WORTH THEIR NAME IF THEY DON'T ALSO LEARN AND BECOME SMARTER DAY BY DAY"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

I'm a cognitive scientist at heart, even if I'm now running a company. My thesis work was about modeling car driver behavior with a cognitive modeling architecture called SOAR. Soar is a goal-based architecture based

entirely on psychological principles. Like our own human information processing architecture, Soar has a Long Term Memory consisting of rules and patterns and a Short Term Memory that consists of a symbolic, graph-based knowledge representation. Soar was used in many different domains including Natural Language Processing, automatic algorithm generation, and learning how to solve complex puzzles. It was even used in military game simulators.

I can easily see an equivalence of my research in modern intelligent knowledge graphs. In the knowledge graphs that we build we usually still have a body of rules in Prolog or SPARQL and a data layer that is obviously a graph-based representation of knowledge. But, with today's technologies we also have efficient statistical pattern recognition, visual object recognition, and amazing advances in natural language processing. So I have the feeling that I can help my customers create very cool systems and still be a cognitive scientist.

***Which concrete business problems can be solved with this approach?***

Almost any business problem needs a combination of rule-based and statistical processing of complex data. If you only want to analyze logging data or time series data, then you probably don't need a knowledge graph. If the answer to your question is hidden in hundreds to thousands of tables, then knowledge graphs are the only way to integrate and simplify the complexity into something that facilitates ad-hoc queries, rule-based processing, or predictive analytics.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Is bread more the result of grain or the result of baking processes? I could leave it at that, but maybe the following helps: knowledge graphs are the result of a series of processes where you take mostly raw data from silos and information streams and turn it into a simple, understandable model that can be easily communicated to business people, data scientists, and business analysts. Knowledge graphs aren't worth their name if they don't also learn and become smarter day by day. So a secondary process is to take the output of rules and analytics and put it back in the graph, thus enriching the content

for further queries and processing.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

It depends on what marketing material they've read first :-) Some people think if you just buy a Graph Database, you already almost have a knowledge graph. Others think it is just an application on top of a graph database. However, I've now sat in enough presentations about knowledge graphs to see that almost everyone has a mix of symbolic knowledge representation (the graph), NLP, machine learning, and predictive analytics. I also see that new customers we meet have absorbed this frame of mind.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

Many of the potential users we talk to already believe they need graph technology, they also think they may need NLP and machine learning. So we inspire them with a set of successful knowledge graph solutions and build their confidence around successfully implementing their own knowledge graph.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

The presence of an enlightened business user that understands that with AI, he can cut costs or increase sales. However, most higher level managers are paid to maintain the status quo, and think "why rock the boat?" Obviously, these enlightened business users are also willing to listen to their own engineers who would love to do AI, but are overwhelmed doing the old stuff.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

Currently, big organizations think that everything will be solved just by hiring data scientists specialized in statistical machine learning. But machine learning is only a tool that has a function within a knowledge graph that puts

all tools into context. So companies first have to realize that knowledge graphs provide the 'context' for all the data science they need to do and then have the willingness to invest not only in machine-learning-data-scientists, but also graph-database-and-rules-scientists (knowledge scientists?)

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I strongly believe in Data Driven Event knowledge graphs. I think in ten years a large number of companies will have transformed their silos, Data Lakes, and Data Warehouses into more coherent all-encompassing knowledge graphs. We are working with some large customers in healthcare and finance and we are already seeing results because of this knowledge graph approach.

## Aaron Bradley (Electronic Arts)

**Aaron Bradley is Knowledge Graph Strategist at Electronic Arts. Aaron helps to build EA's knowledge graph and to facilitate the growth of an intelligent content ecosystem. He is specialized in developing ontologies, taxonomies and content models using linked data standards**

*"SO IN AGGREGATE, MY INTEREST IN KNOWLEDGE GRAPHS ENDURES BECAUSE OF THEIR POTENTIAL TO ENRICH INFORMATION BY DINT OF BOTH CONTEXT AND CONNECTIVITY"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

My interest in what I now think of as 'graph structures' goes back decades: as a young man studying literary criticism, I read Barthes and Derrida on the nature of texts and how people interpret them. I didn't become a literary critic, but what always stuck with me is their notion of "play"; that is, that texts don't have singularly "correct" meanings but derive their meaning from context, including the many different contexts that each individual reader brings to any given text.

Is knowledge, from an epistemological perspective, the sum of semantically-meaningful connections? I couldn't begin to say, but from the perspective of the contemporary data-rich enterprise I think it that, in combination with information about the objects that have been connected, it might be.

So in aggregate, my interest in knowledge graphs endures because of their potential to enrich information by dint of both context and connectivity. And while these fundamental aspects of graphs fascinate me personally, they're bread and butter to me professionally: in the enterprise there's never a dull moment when your problem-solving approach rests on connected data.

***Which concrete business problems can be solved with this approach?***

Any business problem which requires making sense of data from two or more sources, or more broadly, by which business objects can be improved by semantic enrichment (including through interlinking data), is a good candidate for a knowledge graph-based solution.

For example, two very different domains, finance and pharmaceuticals, have perhaps longer than other industries, been employing knowledge graph technology because it provides a method of connecting disparate data points important in each of those businesses. In both of these domains regulatory compliance is critical, and knowledge graphs provide an approach by which data about business objects and, say, the data compliance required by regulatory bodies, can be managed holistically.

Another way of looking at this is that a knowledge graph-based approach allows organizations to transform large amounts of information into knowledge. Airbnb's knowledge graph[151] is a good example of this. The business problem they faced was bringing context to their customers' Airbnb experience so those customers could make better booking and travel decisions. They have detailed information about their rental properties, and there's buckets of information out there about, say, the characteristics of neighborhoods, or things to do in a particular city. Building a knowledge graph has enabled Airbnb to combine these sources of information so that their customers are then armed with the knowledge they need in order to–in the context of those examples–inform their choice of neighborhood for their stay, or to plan activities during their visit..

***Do you see knowledge graphs more as data or as a process that links business objects?***

Trick question, right? :) Because (to take these in reverse order), the ability

to meaningfully link objects from disparate sources, as epitomized by the ontologies that typically form part of a knowledge graph's scaffolding, is a fundamental capability of a knowledge graph. But to generate business value from a graph it must contain instances of the objects in question.

And as an aside, I think one of the reasons people haven't heard of knowledge graphs is because they're both a tangible thing, and a far less tangible process. A graph database is a "thing" for which examples can be provided, just as concepts like "artificial intelligence" or "machine learning" are readily understood as "processes" or "approaches" for which examples can be given. As a knowledge graph comprises both the data and its (semantic) organization, there's difficulties in providing readily understood examples.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

"Knowledge what?" Despite the relative prevalence of knowledge graphs in the enterprise now, most people, even those in mid- to high-tier tech jobs, haven't heard the term used before.

A little surprising, perhaps, given the Google Knowledge Graph has been around since 2012, but because Google's graph mostly succeeds in being something that blends seamlessly into their response to search queries, most people don't think of it as being an especially separate feature of search results, or a part of the technology that's used by smart speakers like the Amazon Echo or Google Home.

So what do people first think of when you raise the term? If they know the term, they probably know the broad strokes of what maketh the beast, although SEOs tend to see it through the lens of their professional interest in Google Knowledge Panels and similar features.

If, like most, they haven't heard the term, the Google Knowledge Graph is the go-to explainer for a guy like me with "Knowledge Graph" in his job title. But I have the luxury of being able to point to in-house examples for the benefit of colleagues, and that's definitely the first impression you want to make if you can.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

Obviously, the best way to highlight the benefits of a knowledge graph to the uninitiated is by example. The degree to which you're able to speak to successes that have made or are making a bottom line difference, even at small scale, will be your best ally.

But even further, (and here disclosing I'm an incurable linked data optimist) knowledge graphs–once you come to know what they are, how they work, and what their potential is–offer an obvious solution to a broad range of problems. So really what garners stakeholder interest is providing a solution to one of their problems using semantics.

Take, for example, the enduring challenge in analytics of combining data from disparate sources and have it make some sort of sense. Perhaps I'm a simpleton, but just framing it that way makes me immediately respond, "meaningfully combining data from heterogeneous sources is a knowledge graph's main value proposition." And while there's other means of combining those bits, the semantics allow you to both describe those connections, and to enduringly connect that data rather than endlessly transforming it.

Knowledge graphs are very much not the solution to every problem, but they're most useful in situations where you want to combine a bunch of data and make some sense of it—which happens to be a ubiquitous use case in computing.

Your success in engaging stakeholders also depends to what degree any solution is a realistic one, and this in turn depends to what degree you already have some semantics available. If some solution rests in part on referencing IRIs for common objects in the business domain, it really helps if these are already available. Whether it comes to tooling, or talent, or systems integration, this is why semantics projects typically start small, and knowledge graphs when initially built are of limited scope. But the utility of having a knowledge graph–that is, of having a bucket of well-described business objects and data about them, a bucket to which you can keep adding new things and new data–means the graph takes on a life of its own once

you've got a bunch of stuff in there some new stakeholder can use and can profit by when they bring their own data into the mix.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

I think the biggest challenge is the paucity of experienced technologists that are required to bring any AI project to fruition. While first and foremost that pertains to the relatively small numbers of experienced and capable knowledge engineers available to work on enterprise AI, at an organizational level the challenge extends to the skill sets and outlook of all of those necessary to build and successfully deploy an AI application.

That is, even with a capable team of knowledge engineers available for an AI initiative, it will never even get off the drawing board without the buy-in of executives that have enough understanding of the approach to support it. And at the opposite end of the spectrum, an AI project leveraging semantic technologies, even if led by competent knowledge engineers, will be plagued by missteps if the bulk of those working on it relentlessly bring relational database thinking to a NoSQL party.

This is changing as the knowledge graph space matures, tooling improves and graph technology looms larger in computer science education, but for the foreseeable future I think the demand for knowledge workers will continue to outpace the available talent pool.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

*"POSITIONING ENTERPRISE KNOWLEDGE GRAPHS FOR SUCCESS USUALLY ENTAILS A CHANGE IN MINDSET"*

Positioning enterprise knowledge graphs for success usually entails a change in mindset, in terms of both technological and business approaches to the organization and use of knowledge in the enterprise.

Because of the nature of knowledge graphs, part of that change implicates traditional organizational structures, as correctly-employed knowledge graphs are silo busters. Put another way, business units that weren't previously required to engage with one another are now in a position to benefit from increased engagement.

All of this is aside from the almost certainly inevitable retooling required, which at any sort of scale is expensive, time-consuming and as often as not taxes both the available knowledge and bandwidth of engineering teams. Unsurprisingly, given this, iterative rollouts rather than once-and-done grand efforts stand a better chance of success when it comes to this sort of change.

Finally, educating and engaging stakeholders, especially in terms of explaining the benefits of a knowledge graph approach for that specific business environment, is critical in maintaining forward momentum. If the people central to defining AI strategy aren't convinced that knowledge graphs have a substantial contribution to AI success they'll turn to other approaches.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I think we'll increasingly see semantic AI employed as an elegant solution to a whole range of complex data problems: everything from providing a means by which content can be personalized for and recommended to consumers in commerce environments, to fueling the generation of literally life-changing insights in the health sciences realm.

Enterprise search engines will continue to loom large as exemplars of applied semantic technologies, though the label "search engine" will become (even as we see today) increasingly ill-suited to the range of the functionality provided by the likes of Alexa, Siri, Cortona and the Google Assistant. And just as the rapid growth of the mobile web was a major motivator for the development of these semantic solutions, voice search and other human-machine interfaces that don't involve keyboards will propel further innovation in this space, such as much-improved conversational AI.

We'll also, I think, see knowledge graphs play a larger and larger role in the public realm, because they're exceptionally well-suited to making sense of the vast amounts of data produced by governments, research bodies and academia. Knowledge graphs have enormous potential to inform public policy development by providing lawmakers with contextually-relevant information freed from previously-inaccessible data silos.

Finally, AI will build on its own successes, and we'll see a more automated approach to the construction of knowledge graphs (see, for example, Diffbot), though I think there will always be a role for explicit semantics in semantic AI.

## Yanko Ivanov (Enterprise Knowledge)

**Yanko Ivanov is a Solution Architect at Enterprise Knowledge. Yanko specializes in Semantic Web technologies strategy and design, taxonomy and ontology design and implementation, content approval workflows, and systems analysis and integration.**

*"THE MOST EFFECTIVE ORGANIZATIONS WILL TAKE TIME UPFRONT TO DEFINE SPECIFICALLY WHAT IT IS FOR THEM AND, MORE IMPORTANTLY, WHAT THEY'LL GET OUT OF IT"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

Over my career in the various aspects of knowledge management, I have worked with a number of tools, platforms, and solutions that attempt to provide a consolidated view of an organization's content and information. What I found with the vast majority of these solutions is that they either need an incredible amount of effort to implement, or only solve the silo problem partially. Conversely, knowledge graphs address both the technical and business challenges many of the organizations for whom I consult are facing.

Once I was introduced to the Semantic Web and knowledge graphs concepts, I found this to be a very elegant way to allow organizations to actually produce a unified view of their information and knowledge. Add to that the

ability to integrate structured data and unstructured content, the ability to traverse information, and discover facts that we wouldn't know otherwise, and semantics quickly became one of my passions.

***Which concrete business problems can be solved with this approach?***

This is the beauty of the semantic approach—it is so flexible that it can be applied to a wide variety of use cases and business problems: this approach can answer "who did what when?" in a highly decentralized and matrixed project organization, organize all content and data that we have on a specific topic, person, or business asset, or integrate with advanced NLP and AI techniques to analyze the full set of information we have on a specific problem and produce an actionable result in business terms. Additionally, customer centricity, content auto-tagging, inventory tracking and management, and product information management are all use cases in which this technology shines. In short, this is an extremely exciting technology limited only by the potential use cases an organization might conjure.

One of the more common examples of solving business problems is implementing a smart, context-based recommendation engine to push relevant information at the point of need. For instance, one of my clients is leveraging knowledge graphs to recommend articles on a topic before a calendar meeting based on the topic of the meeting. Another example is running semantics-based text analytics and mining tools to collect and present all relevant information they have on a topic, person, or asset. This technique is incredibly valuable with the advancement of GDPR-type of laws and regulations or for legal purposes and risk mitigation.

It really is fascinating how this technology can be applied to address a vast number of business problems.

***Do you see knowledge graphs more as data or as a process that links business objects?***

In my mind, implementing and managing a knowledge graph is a process, without a doubt. As I've expressed to some of our clients in the past, implementing a knowledge graph is not the end result, it is rather a way to run your business. There are many variables to be considered in the strategy,

planning, implementation, and governance of a knowledge graph, and the majority of those variables are organizational, process factors. Designing and implementing the technology in itself, while not necessarily trivial, is relatively straight forward. But designing it in a way that it is infused with your day-to-day activities, that it supplements them rather than being just another system to maintain, that is where the challenge is.

If designed properly, the consideration is in fact moot. A well-designed knowledge graph will help an organization manage, find, and discover structured information, unstructured content, and for that matter, anything in between.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

Like many terms in our industry, the term knowledge graph is presently being used very differently by different organizations. The knowledge graph term is often vague and nonspecific from a business person's perspective. It is not like, say enterprise search, content management, or CRM. This is one of the reasons we often need to spend some time explaining what it is and really defining the business value for each specific organization. While the popularity of knowledge graphs has skyrocketed in recent years, the definition of the concept and, more importantly, the "how can this thing help me solve my business problem" question is still very much relevant and needs attention. The most effective organizations will take time upfront to define specifically what it is for them and, more importantly, what they'll get out of it.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

In my experience, the most productive ways to demonstrate the value of a knowledge graph-based solution include conducting a more in-depth demo of a working solution, or even better, conducting a short proof of concept that is focused on a specific business challenge and leverages a subset of the organization's content and data. At Enterprise Knowledge, we often conduct such PoCs that iteratively demonstrate the value of the technology to the organization and actually solve a real world problem for them. With that in

mind though, a key piece of implementing a successful knowledge graph is developing a long-term strategy and roadmap for it, including plans for the supporting organization, data, ecosystem, and measurable success criteria.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

Based on the work I've conducted with our clients, I see two challenges:

1. AI is not a silver bullet. There is still the notion that implementing an "AI tool" is a plug-and-play process, that it will do everything on its own, that it will define the taxonomy and ontology that the *actual end users* care about, that it will do the data transformation and unification on its own, and that it will know what that user is trying to do with minimal level of effort or user input. In most cases, we are simply not there and this is an area we work hard on to ensure organizations are well prepared for the long-term investment in their AI endeavors.
2. Training material for machine learning requires expertise and time to develop. And by training material I mean curated content or data, validated and verified by a subject matter expert, the gold standard if you will, that will be fed into the machine learning algorithm for it to learn the specific domain. Organizations are asking for machine learning, wanting to leverage the power that it can provide, but it requires training of the tool. Organizations on the path of implementing such technologies need to understand and plan for resources and time to develop the gold standard, the training material that can then be used to scale the solution through machine learning.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

First and foremost, education. Understanding the power behind the technology, its capabilities, how it can be plugged in the day-to-day business activities, and the roadmap for implementation is a critical step in the road of successful implementation. True AI can fundamentally benefit from the implementation of knowledge graphs, but it also requires thoughtful

integration, intuitive user experience, and clear reasoning on the decisions or actions of the AI.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I think that in 10 years no one will be talking about knowledge graphs anymore. Not because they'd be forgotten, but because knowledge graphs will be a key piece of the solution, a foregone conclusion that a knowledge graph is a fundamental component of the advanced AI solutions we are implementing.

What won't change is the need to understand the business and to define a tailored, actionable, and achievable roadmap for implementing and governing these AI solutions.

## Bryon Jacob (data.world)

**Bryon is the CTO and co-founder of data.world—on a mission to build the world's most meaningful, collaborative, and abundant data resource. Bryon is a recognized leader in building large-scale consumer internet systems and an expert in data integration solutions.**

*"A DATA CATALOG IS A KNOWLEDGE GRAPH—ONE WHOSE UNIVERSE OF DISCOURSE IS THE DATABASES, REPORTS, GLOSSARY TERMS, AND SO ON"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

My academic interests, pre-dating my professional career, centered on cognitive science and particularly in understanding and simulating the mechanisms of human thought in software. Knowledge graphs are a realization of how logical reasoning can be captured in a declarative fashion and applied to data structures, giving us a way to communicate with computers at a very deep level–to take some of those core structures of thought and make them machine-processable.

***Which concrete business problems can be solved with this approach?***

ETL, ELT, Data Prep—these are all forms of inference! You're taking raw

facts, applying some set of rules about what that data means or how it should be represented for some analysis. RDF is a fantastic format for representing data in an interoperable way, so that data integration becomes as simple as merging graphs. And when you represent your data as RDF in a triple store, so many of these common business operations reduce to a matter of representing the logical relationships between the entities that the data are about.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Of course they're both—but I see them more as data. There's a famous saying that "a little semantics goes a long way", and I think that's especially relevant in the business arena. Companies are being crushed under the weight of the data that they've stockpiled for the last couple of decades, and they're looking for a way to understand what data assets they have and how they're interconnected.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

For many, it's completely alien—they think in terms of graph visualizations, or maybe have some dim awareness that this is a special type of database that social networks use to manage interpersonal relationship data. Some do have a better understanding, there will often be an engineer or IT professional who has some previous exposure to semantics, RDF, or the like. By and large, though, it's their first deep exposure to the concept and they're eager to learn.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

Our approach is to start with the first problem most organizations face when they're trying to operationalize their data at scale—cataloging their data. A data catalog is a knowledge graph—one whose universe of discourse is the databases, reports, glossary terms, and so on. It contains an understanding of how all of these things are interconnected, how they are used, and what their lineage is. From that foundational knowledge graph, users can start to build

more highly articulated domain and application knowledge graphs, using the metadata from a data catalog both as references for ontology design, and as a roadmap to connect to the data points themselves.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

Understanding what data resources the organization has at its disposal, and what real world entities and relationships are represented by that data.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

For many, it's primarily an awareness problem. Knowledge graphs are not part of the mainstream data management toolkit, so education is the first step. Once folks have an understanding of how powerful knowledge graphs are, another challenge is the "all or nothing" mentality that has been ingrained by years of data warehouse and data lake solutions—many IT professionals hear the pitch for knowledge graphs and think "sure, but first I have to abandon all of my relational databases and big data solutions?" No! A knowledge graph is inherently a logical structure, so it lends itself very well to data virtualization—existing investments in infrastructure can remain, and can become a part of the knowledge graph.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

We'll have reached the point of scale in most organizations where data integration in order to support AI is going to essentially require data representation as RDF—any solution that works will be isomorphic to RDF, and sharing data with third parties will become increasingly important, so re-inventing it will become less common. As more data is shared, stored, and moved around in RDF, the incentives to create better shared domain, industry, and use-case driven ontologies are increased, and we will see more machine readable business logic shared and standardized as a result. That, in turn, will create more incentive for data representation that can take

advantage of that shared logic - and that flywheel effect will bring the Semantic Web to critical mass. The ability for AI solutions to leverage this network of human-encoded knowledge on essentially all data will greatly accelerate what can be accomplished with AI, and the AI solutions will start to meaningfully contribute back with machine-generated ontologies—addressing some of the issues with explainable AI, and capturing machine insights in a form that can be directly reasoned about and shared.

## Atanas Kiryakov (Ontotext)

**Atanas Kiryakov is Founder and CEO of Ontotext. Atanas is a leading expert in semantic databases and knowledge graphs, and author of scientific publications with 2500+ citations.**

*"WE CANNOT GET ACCURATE RESULTS FROM THE MACHINE IF WE CANNOT AGREE AMONGST OURSELVES WHAT THE CORRECT OUTPUT IS"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

Knowledge graphs are the most advanced knowledge representation paradigm. With over 25 experience in AI, I can tell humanity never had an instrument like this. They combine the best we had with taxonomies (380 BC), semantic networks (1956), network model databases (1971), knowledge bases (1980s), ontologies (early 1990s), semantic dictionaries (late 1990s) and linked data (2000s). And all this at a scale which reveals new qualities. As Marx said, “quantitative changes result in qualitative changes”:

1. A KG of millions of entities and concepts exceeds the knowledge of every single expert in any imaginable domain.
2. A network of such size allows for the efficient and concrete use of cognitive paradigms like priming, analogies, etc., not just in a toy example for someone's PhD. By simply running PageRank on a big knowledge graph you get already a very meaningful relevancy

ranking.

***Which concrete business problems can be solved with this approach?***

No one directly. Which concrete business problem does Java solve? This is a tool that allows humans and computers to complement each other in knowledge management, information extraction and information retrieval. You automate the rudimentary part of the work of librarians, editors and countless entry-level knowledge workers. By making more knowledge explicit and making it easier to discover and interpret, it is no longer locked up in the minds of a few experts, but much easier to share and use.

Human experts can model knowledge as ontologies and produce high-quality metadata to bootstrap a KG. Computers use this KG to interpret and interlink data from different sources and gather a critical mass of domain awareness. This allows computers to analyze unstructured data, extract new facts and generate vast volumes of new metadata and enrich the graphs. This way we arrive at systems that in many aspects surpass the analytical capabilities a graduate student has in their area of study. A lot of the value of knowledge graphs is in the fact that they are human readable and explainable—unlike neural network models. One can explore it, use it as reference data structure, correct it, govern it, publish it, etc. Knowledge graphs make knowledge management and AI work together.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Knowledge graphs represent information structures, which can be used in processes or managed via processes. They are similar to taxonomies and databases in their nature—partially explicit, simplified models of the world and representations of human knowledge. One needs engines, to store it, query it and search it, and methodologies and tools to manage it and use it. But KG graphs are not software. A KG may or may not contain process knowledge.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

It depends on their background. Many think of a KG as "taxonomy on

steroids". Others consider it a next generation data warehouse. Quite a few think of them as reference/master data.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

For the last 10 years it has been tough to inspire users which were not already enthusiastic about knowledge graphs. It has been an early adapter's market. We had the chance to work with enterprise customers, which have spent millions and tried everything money can buy in mainstream data management and content management. They came to us, because they had problems they cannot solve without semantics. But how can you inspire customers who have not had this experience themselves? How would you inspire a schoolchild to go to a university and get a degree? If she doesn't already believe this makes sense, if her family and social environment haven't educated her to believe this will pay off in the long run, you don't have great chances. You can provide examples of successful organizations which did it —sometimes it works.

In 2019, we finally got to a point where knowledge graphs are recognized as the next big thing in metadata management and master data management. You can now inspire customers with simpler arguments: semantic data schemata allow more automation in data management; explicit semantics brings better continuity in business; connecting data helps you put data in context and gain deeper insights.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

Often clients do not understand the importance of properly defining the tasks that we want the machine to solve. Let's take for instance the task of extracting parent-subsidiary relationships from text. First, the relationships need to be properly specified and get answered questions like: What counts as subsidiary? Should we count the owner of 60% of the shares as a parent? Then there is a need for a good quality golden corpus of texts annotated by a human with the types of metadata we expect the computer to produce from it. To get this right, one should have good annotation guidelines so that human experts following them can reach a high level of inter-annotator agreement.

We cannot get accurate results from the machine if we cannot agree amongst ourselves what the correct output is. In such situations there will always be people who judge it as stupid AI and blame the developers.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

Organizations should understand how knowledge graphs can lower the costs, speed up and improve the results of AI projects. It's mostly about repurposing data preparation efforts across projects. Instead of wasting data which is already integrated, cleaned up and unified, enterprises can use knowledge graph platforms to manage such datasets in a form that keeps them connected, up to date and easy to discover. This way, knowledge graphs lower the preparation efforts needed for AI projects and enable deeper analytics based on richer data with better context information.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I expect steady growth of the market. In 10 years KG platforms will replace today's taxonomy management systems and will become the most popular metadata management paradigm. KG technology will also become an intrinsic part of solutions for master data management, data cataloging, data warehousing, content management system and the so-called ‘Insight engines’.

Gartner positions knowledge graphs in the first part of their Hype Cycle for AI in 2019 —the Innovation trigger phase. They expect that soon we’ll arrive at the peak of inflated expectations and disillusionment. I disagree: we passed the disillusionment phase in 2014–2015, when the same vision and tools were considered semantic technology. Now we see mature demand from enterprises, which already got burnt with badly shaped semantic projects and immature technology in the past and now have much more realistic expectations, better defined applications and better evaluation criteria for such technology. We don't see the hockey-stick growth typical for the first phases of hype on the market; rather, we see normal demand growth from leading vendors who are around for more than 10 years and have learned

their lessons too.

## Mark Kitson (Capco)

**Mark Kitson is a Management Consultant with Capco's UK Data Practice—leading on the application of graph and semantic technologies across Capco's clients in banking, wealth management, insurance and other sectors of financial services.**

*"IN FINANCIAL SERVICES MANY POTENTIAL AI APPLICATIONS ARE NOT SUITABLE AS REGULATIONS AND ETHICS DEMAND EXPLAINABILITY"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

Graphs' ability to express knowledge and allow that knowledge to be used at any scale is simply awesome. This is knowledge—smart data that supports complex reasoning and inference. Seeing this smart data working for people, rather than smart people working on data gives me hope for humanity and our ability to untangle and understand complex issues. Knowledge graphs' superpowers: flexibility, self-assembly, knowledge sharing, reasoning-at-scale across complexity—are game changers. I love the ability to weave together data from different sources, to ask simple questions, and get meaningful answers. As more of us work in agile ways, graphs' iterative schema-late or schema-less data models are a revelation.

***Which concrete business problems can be solved with this approach?***

- Point solutions—identity fraud, network management—but also new solutions like helping large complex organizations manage risk by turning the controls and obligations buried in pages and pages of internal policies, regulations and contracts into risk and control networks—as a graph.
- Enterprise Knowledge Graphs and enterprise-wide knowledge are broad but still concrete problems. Helping businesses tame and shed complexity as they transform and grow. Silos, data fragments, and the resulting ambiguity make businesses more opaque and complex than necessary. Reducing the negative effects of these silos is an enormous opportunity.

***Two powerful approaches that we are helping clients with:***

- Using graphs and semantics to provide self-service, cohesive consistent data—absorbing the cost and confusion of legacy data by aligning identifiers and defining hidden links with semantics. This also simplifies and facilitates exploitation of external data.
- Complex enterprises share core taxonomies of people, process and technology to reveal a rich Digital Twin. This creates a mirror of themselves that helps them to transform, be leaner, more in control and agile, yet able to enjoy the strength of scale and confidence that comes from mastering complexity.

One opportunity that we are using internally and with our clients is to use graphs to better understand customers and respond to their needs. Our clients have rich and complex relationships with products, their customers and other stakeholders. With semantics their complexities can be captured and understood—allowing even the largest firms to offer personalised services and products.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Knowledge graphs are data with a few simple components, but with the potential to manage huge complexity—the things or nodes, the links or edge

between them. There are also simple processes that allow that data to be shaped and curated into knowledge, but these processes require a different mindset to traditional data modelling—a challenging paradigm shift.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

In one word: confusion. Awareness of knowledge graphs is patchy in the financial services industry, even among data professionals. Confusion between knowledge graphs and the other graphs (bar charts, etc.) is a regular tripping point. But the issues that knowledge graphs can resolve—fragmented, inconsistent silo's data—are all too well known.

Financial services is already heavy in hype and jargon, particularly when it comes to data. So whiteboarding examples of how simple graphs might address issues in a familiar domain is often an easier path to that "aha moment" than explaining confusing terms. Customers understand the challenges of complexity and the need to connect-the-dots. Customers get genuinely excited about the prospect of clarity—like a breath of fresh air.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

When users are less familiar with the topic, getting hands on with large team-sized problems works well.

Our most effective approach is to bring people together at Capco's meeting spaces between the financial heart of the City of London and the start-ups of Silicon Roundabout. We bring data and business leaders from one firm together with selected vendors to get hands on with graph data, and graph thinking—quickly moving from theory to real world opportunities.

Most clients will have small but complex data sets that can be quickly transformed into a simple knowledge graph. A dynamic visualisation is often enough to make the connection needed and inspire potential users to explore the potential further.

***What is the biggest challenge in developing organizations to bring AI***

***applications into production?***

Data Quality is a major challenge for many organizations. This slows down AI delivery and limits the scale of datasets that can fuel AI applications.

In financial services many potential AI applications are not suitable as regulations and ethics demand explainability. Graph based AI solutions offer the opportunity to expose and explore reasoned answers in ways that machine learning models cannot.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

There is a lot of excitement about the potential for AI—and rightly so in many cases. Beyond the hype, organizations will need to find the right sandwich of people, process, technology and data. Relational data, graphs and semantics, robotics, machine learning, and in almost every case—the human in the loop.

Capco's focus on financial services and multi-disciplinary teams allow us to bring domain experts, engineers, designers, communicators and vendors together to design powerful "people-tech-data sandwiches" with our clients. Ultimately, AI is fuelled by cohesive connected data—graph is the fastest, cheapest way to unlock that data.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

In financial services, point solutions will continue to pop up as proven use cases gain traction and become the norm. Forward thinking firms will have recognised the need for a "semantic strategy" that optimises the reusability of data and capabilities —ensuring that point solutions and isolated graphs can self-assemble to form the firm's aggregated knowledge graph. We are helping a few clients who have recognised this opportunity to shape their approach—incrementally building an organization's brain—true business intelligence.

## Lutz Krueger (formerly DXC Technologies)

**Lutz worked as Knowledge Manager for DXC Technology till February 2020. He has been working in knowledge management, semantic technologies, AI-based modelling and in the implementation of knowledge graphs for years. Lutz introduced a federated knowledge graph service approach in DXC as the central foundation for semantic enterprise applications.**

*"CUSTOMERS WHO ALREADY INITIATED THE DIGITAL TRANSFORMATION HAVE USUALLY VERY QUICKLY REALIZED THE POWER AND THE NEED OF KNOWLEDGE GRAPHS TO BUILD CUTTING-EDGE SEMANTIC APPLICATIONS"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

Knowledge graphs are my preferred approach to model and to represent real-world knowledge and domain-specific enterprise knowledge. Knowledge graphs ensure a seamless collaboration between humans and machines.

***Which concrete business problems can be solved with this approach?***

In various domains of knowledge and data management, analysis, search and knowledge discovery, I see valuable opportunities. A concrete differentiator

might be to transform a sophisticated enterprise keyword search solution into a semantic, knowledge graph and NLP/NLU-based enterprise knowledge discovery solution combined with a seamless harmonization and alignment of structured and unstructured data.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Knowledge graphs can combine both capabilities, to represent semantic data models and to carry out logical reasoning to trigger related transactional processes.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

Customers who already initiated the Digital Transformation have usually very quickly realized the power and the need of knowledge graphs to build cutting-edge semantic applications.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

We have seen true user inspirations while demonstrating various semantic capabilities of a rapidly developed search solution prototype which is using unstructured data and based on a knowledge graph built using real user data and taxonomies.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

The challenge is to develop and to execute a strategic and combined deployment plan for all in-scope emerging technologies and to follow a holistic maturity model for all building blocks.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

Beside the strategic reorganization, a primary element is to establish a governance to build, to curate and to control federated knowledge models based on domain specific taxonomies, ontologies and metadata models for structured and unstructured data.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I think we will see an evolution towards adaptive knowledge forests based on recent AI/ML methods and knowledge graphs. Various technologies in this field will be used as integrated building blocks within an open semantic framework.

## Joe Pairman (SDL)

**Joe Pairman is a Senior Product Manager at SDL. Before joining SDL, Joe led a consulting practice, helping clients from a wide range of fields and industries get the most out of intelligent content, semantic technology and taxonomy.**

*"WE HAVE TO MAKE THESE MENTAL CONNECTIONS AND DROP OUR OLD, APPLICATION-EXCLUSIVE THINKING"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

My interest in this area stems from when I had established myself in the field of structured content, and yet saw a gap. In my particular niche—slightly insular as are so many groups of tools and technologies—we described formal rhetorical structures in machine-readable form, and yet had no such precision to describe the meaning of the text itself.

Linked Data filled that gap for me, representing each real-world idea or object with a simple, globally unique identifier. No matter the different names it had, one URI let machines and people alike refer to this “thing” without confusion. I absorbed and evangelized this idea, and could never again be content with a content tool that couldn’t accommodate unique taxonomical IDs.

Lately, though, my mind has sat more in the space between entities—their relationships. It's easy to picture hard-edged objects, but less so to conceive the dependencies between them. For example, project management tools revel in tasks, time-blocks, and roles, but always struggle to represent the connections; the fact that person A needs to do task B before she really knows what delivery C will look like.

So it's the "edges" in knowledge graphs that get me thinking about the future of human-machine interaction. A graph models relations, we believe, in a similar way to that of the brain. How then can we represent and manipulate those connections in a way that speaks more directly to our perceptions and those of all users? Many people in the field are working on this challenge, and it's certainly one that keeps my brain turning over!

***Which concrete business problems can be solved with this approach?***

The "table stakes" of knowledge graph applications are to describe real-world ideas and objects unambiguously. There is already a lot of value in this, for example to use a common vocabulary across and beyond an enterprise, avoiding the Babel of crosswalks.

But there are problems of even higher value that knowledge graphs address. If an engineer changes one part of a complex medical device, what other parts are affected? Do the safety requirements change, and should the documentation be updated accordingly? Or what about externally authored legal texts where each one of a thousand paragraphs has multiple versions, of which only one is currently in effect, but the upcoming versions require detailed changes to internal guidelines? Through fast and flexible mapping of relationships, knowledge graphs provide a more powerful, cost-effective way to model and manage critical real-world domains.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Knowledge graphs are data, of course, but the key benefits come from the links to and between very concrete business objects.

***What do customers usually think of first when they are introduced to the***

***term 'knowledge graph'?***

Many customers recognize the term since Google used it to describe a specific application. Those boxes of biographical information next to the “ten blue links” are at least an introduction to the idea of unambiguous entities, although they do little to illustrate the underlying connections between those entities. A graph isn’t a graph without edges! Other customers—at least their more technical people—identify the term with graph databases in general. This is closer to the mark, although can still be taken as simply a siloed, application-centric data store instead of the powerful, pan- and inter-enterprise enabler that is an actual knowledge graph.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

It starts the same as any selling of ideas: acknowledging pain points, focusing on potential benefits, and then gradually settling on the applications that bring those benefits. Where things get interesting, though, is to persuade people of the advantages of the Linked Data approach compared to other approaches driving similar applications. This can be done by example; pointing out the cost savings or the superior user experience. But most effective (and hardest) is to persuade the user of the fundamental advantage of knowledge graphs for certain classes of problems.

The popularity of AI has helped here. Many people have had the potential applications pointed out to them loudly and persistently by advocates of a pure machine learning, hands-off approach. But people are starting to distrust such an approach on its own, through a combination of disappointing implementations and high-profile mistakes—voice assistants developing sociopathic responses, for example. So the world is ready now for the idea of an explainable AI; one that operates not solely through mechanisms that not even its creators fully understand, but rather one that bases its decisions on the kind of knowledge model that we all have of the world; a web of people, objects, and organizations linked by dynamic interactions and relationships.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

To put an underperforming, unscalable AI application into production is not that hard! But to do AI sustainably is a big challenge. Not long ago, I would have said that knowledge and culture were the biggest bottlenecks. Organizations, including development teams, lacked the background and affinity with semantic AI that would let them even start to put the pieces of an effective solution into place.

Now, interest and knowledge is spreading gradually. The next bottleneck may be one of engineering. Code is not so different from concrete and steel; architectural edifices built in one way cannot simply bend into a different shape, or have the foundations replaced. Certainly, solutions that already have a decoupled approach to metadata, such as the product I manage, are at a significant advantage. But in any case, to build a sustainable semantic infrastructure on which to base AI applications takes planning and time. So better start now!

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

The most essential change is to stop seeing AI as a collection of applications based on a pool of amorphous data, and start joining things up. For example, a taxonomy can improve search, but that same taxonomy can drive reporting and insights, and help connect systems across the enterprise. An external-facing recommendation engine for content could, without fundamental modification, highlight dependencies and risks for internal administrative users. We have to make these mental connections and drop our old, application-exclusive thinking.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

Perhaps real success is when the technology itself disappears from view and becomes an assumed part of the plumbing. As I wouldn’t go to a conference and talk about basic Javascript now, it may become redundant to talk about the fundamentals of knowledge graphs (except in high school classes). I will no longer have to reach out carefully in conversations with technical peers and see whether they agree that we should not rely on strings, but represent

“things”. Yet the benefits of semantic AI will be available to many more people; in Bret Victor style, high-level managers without a line of code in them will be able to drag and drop entities directly onto data visualizations to understand deeply their surrounding business contexts, dependencies and risks.

But this only happens if certain things come together. To succeed, we’ll have to focus on the “good enough” and the “commercially viable” rather than aiming at pristine elegance. To retain credibility and influence, we’ll have to do our part in connecting the siloed clusters of tools and tribes; conceding graciously when other approaches are objectively better for certain classes of problem. As much as visionaries and engineers, we’ll need to be diplomats and even politicians to bring the full benefits of knowledge graphs to a mass market.

## Ian Piper (Tellura Information Services)

**Dr Ian Piper is Director of Tellura Information Services. Ian specialises in taxonomy design, semantic technologies, content architecture and modelling, usability testing, content management and web application development.**

*"THERE IS NO NEED TO PUT ALL INFORMATION INTO A SINGLE SYSTEM; ALL THAT IS NEEDED IS TO HAVE AN UNAMBIGUOUS WAY TO GET TO THAT INFORMATION"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

My principal interest in knowledge graphs is the opportunity that this approach to information management offers, that is unmatched by conventional approaches to information. From a very simple structure—the triple—it is easy to build out massive networks of connected information. This structure then allows sophisticated exploration across this network, and offers new insights into the organisation's information.

For me, the possibility of novel information exploration tools is very attractive. The traditional search box is outdated and limited in value. I want to see new exploratory environments that enable organisations to get the most out of knowledge graphs by revealing wide-ranging and unpredicted links between information across the entire information landscape.

***Which concrete business problems can be solved with this approach?***

The key problems that are addressed by knowledge graphs all come down to information connectivity. Information marooned in silos is common in organisations of all sizes and types.

Traditionally the "solution" offered to get useful actionable information out of silos is to break them down—to pull all information together into one common system, such as Oracle. I have never seen this approach work, and the reason is understandable; people and departments within organisations have a feeling of stewardship of "their" information, don't trust others to look after it properly, and want to keep control of it. It is very difficult to argue against this principle.

Another common approach has been to use a search engine to index all of the information. This is equally prone to failure, mainly because of the inherent shortcomings of search technologies. Search engines don't know what you are looking for, they only know what you've typed into a search box. You will probably get some of what you are looking for, but it will be buried within a mass of other things that you are not looking for. The piece that is lacking in traditional search is context. The results from a search query can only address the words used in the query, not the meaning in the mind of the person doing the search. Also, since search indexes the occurrence of words within text and not the meaning of those words, we have an even greater absence of context. Search engines take a literally meaning-less query and send it to meaning-less data. How can you expect to get high quality meaning-full results?

Building a knowledge graph enables you to address both of these approaches. There is no need to put all information into a single system; all that is needed is to have an unambiguous way to get to that information. This is a Uniform Resource Identifier (URI); an identifier and locator for a piece of information. So you can leave your information where it is, and you know that you can get to it in future via its URI.

Just as important, using a knowledge graph helps you to semantically describe the information objects in your system (they represent Persons, or Roles, or Projects, or Digital Assets) and, crucially, the nature of the

relationships between those information objects (a Digital Asset has a hasAuthor relationship to a Person, a Person has a hasRole relationship to a Business Role, a Project has a requiresRole relationship to a Role). We now have information objects that can be related together contextually, a relationship that enables meaning-full information discovery processes for the business.

***Do you see knowledge graphs more as data or as a process that links business objects?***

To me, a knowledge graph is not just data. It is a flexible, infinitely extensible network of semantically linked business objects. As the question implies, it's also a way to build up and explore that network, so it is indeed also a process. That's quite a mouthful, so I'll just take a moment to unpack it.

First, a business object is a piece of information in the most general sense possible. It might be a piece of narrative content that may eventually appear on a website or in a customer user guide or an image, or a piece of interactive content to be used in an online learning management system. A business object will have information—what you might call a payload—which is the real information that a user is interested in and will need to get to. It will have descriptive metadata, which provides additional cues to help discover and contextualise the information. Some of this metadata, crucially for the purposes of the current discussion, provides semantic relationships between the current business object and other business objects. A business object will often model a real-life object within the business; a person, an information asset, a product or service. This latter design feature is important in cementing the value, the relevance of a business object to the work of the organisation.

Turning to the semantic links; this simply means that not only can one business object have a relationship with another business object, but the link between the two objects itself has a meaning. Building semantic links between things not only joins things together, it also provides the context in which things relate one to another. You know that two things are related (say, Person A and Person B) but you also know exactly how they are related (Person A is the mother of Person B and Person B is the son of Person A). Adding semantics to what was a simple relationship now provides a massive

amount of contextually rich value.

So how does this become a network? The network arises from the fact that the simple, first-order relationship between two business objects is not the only relationship that either of those objects have.

To put this in concrete terms, a business object representing an Article would have a link to the business object representing a Person, and the link itself (hasAuthor) would be meaningful.

```
[Article] hasAuthor [Person]
```

But that Person might have written many Articles, so would have many hasAuthor relations pointing to other Article objects. An Article will also have date information describing when it was written, it will have links to taxonomy concepts representing the aboutness of the Article, it may have links representing a larger structure in which it exists (Article A isPartOf InformationAsset B) and possibly many other such links.

There is more to be seen here too. The nature of this kind of semantic relation is that it can be explored in more than one direction. The fact the Article A has an author Person B means that Person B has written Article A. With a network of linked objects, you can explore in either direction—"who wrote Article A?" and "what Articles has Person B written?". Since every information object may have semantic links to many others, it is clear how an extensive and rich network of information objects can emerge.

An obvious risk is that the resultant network is chaotic—with so much interlinked information, how can you hope to get valuable insights about that information? This is where the underlying structural principles of a graph help. The entire graph can be reduced to a collection of three components: a subject (a business object), a predicate (the semantically defined link) and the object (the other business object). However, a subject in one relation can also be an object in another relation. So when exploring a graph you can define where you want to start, what relation you want to explore, where you want to end, or any combination of these. It is more complex, but then businesses are complex, and the graph approach helps navigate that complexity.

***What do customers usually think of first when they are introduced to the***

***term 'knowledge graph'?***

Clients in my experience run the gamut of responses from enthusiasm to disinterest. Since there is usually at least some existing interest in the ideas, clients are usually receptive, so I don't encounter much disinterest.

Explaining the basic ideas of linked data and semantics usually elicits quite positive responses. Most organisations struggle with how to maximise the value and actionability of their information, and the advantages of being able to link their own content together in meaningful ways are usually clear.

Much more crucial is the degree to which the principles of knowledge graphs are directly relevant to their information needs. If there is some existing desire to make better use of the organisation's information, this usually forms the basis for a productive conversation about taking on knowledge graphs.

But there is inevitably some reluctance to take on any new technology, especially one as fundamentally new as knowledge graphs. So early responses will often take the form of:

- Do we have to get rid of Oracle (answer: no; graph data and relation data can co-exist quite happily)?
- Do we have to get rid of our search engine (answer: no, it will enhance your search tools)?
- What is this going to cost (answer: tactically very little; strategically an amount dependent on how deeply you invest in the ideas)?
- How much human effort will we have to commit (answer: probably less than you think; there is excellent technology support for designing and building knowledge graphs)?

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

Nothing works better in my experience than demonstrating a working example, and particularly in the form of a visual graph. This is why I developed the Content Graph Explorer.[152] This application shows how to build a graph based on real-life content linked to a controlled vocabulary of business concepts. Starting from an item of content it is possible to see all of

the related concepts. For any selected concept it is possible to see all of the content that has been classified with that concept. And from any of those linked content objects we can explore their concepts. With just two types of business objects—a content object and a taxonomy concept—we can quickly explore a network of linked information. When I demonstrate this to clients they get it immediately—this is a way of exploring their content that is simply not possible by other methods.

Without a doubt, it is the rarity of intuitive and accessible tools to build knowledge graphs that holds back users from engagement. I'll have more to say on that in the final question below.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

I'm an enthusiast optimist. There is an increasing appreciation amongst business users that the conventional tools for exploring information—old-school search engines, relational databases—are not fit for purpose in the age of linked data. We are presently in a good position, with a changing business mindset coupled with the availability of good supporting technologies.

However, there is more to be done. While graph development products are available, they are largely the province of specialists like me. As I mentioned briefly above, I believe that the key to significant uptake of knowledge graph technologies will be the emergence of effective and easy to use tools aimed at general business users. By this I mean several types of tools:

- Business-focused tools for building graphs. This includes intuitive tools for both building taxonomies (such as the Cardsort application) and for linking things together (my Content Graph Explorer and TurboTagger applications are early technology exemplifiers).
- A move away from the conventional search box towards more sophisticated exploratory information discovery mechanisms.
- New tools and APIs for building simple end user applications—possibly microservices—for building semantic information discovery features into other applications.

I believe that such new tools will begin to appear in the very near future—my company has already built simple technology demonstrators that explore these areas of interest.

Another crucial component to bringing knowledge graphs into the mainstream of business will be the appearance of tools with a low-cost entrypoint. Many online business tools—Slack, BaseCamp and even Google—have become hugely successful by using the free-to-premium cost model. This model offers users a free edition with a low level of capability, but with clear, tiered, value-added services at different cost levels. There is at present no commercial graph development tool that offers such a model, but it is certain, not to mention essential for the wider uptake of knowledge graphs, that such tools will appear.

## Boris Shalumov (Deloitte)

**As a Senior Consultant in the Strategy & Operations department of Deloitte, Boris Shalumov is managing the service offering around semantic technologies with focus on industrial applications of knowledge graphs, NLP and semantic AI for different areas such as business transformation, product development and supply chain analysis.**

*"KNOWLEDGE GRAPHS ARE BASICALLY BUSINESS DIGITAL TWINS OF A COMPANY"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

A semantic manifestation of domain knowledge into a graph allows human users and machines to accurately represent and communicate very complex, dynamic, highly interdependent and ambiguous information. The knowledge over a domain becomes transparent, easily accessible and exists as a part of something bigger rather than a lonely data island.

***Which concrete business problems can be solved with this approach?***

One of the most valuable and interesting applications of a knowledge graph within an enterprise is the impact of management decisions and business-related changes. Even though the well-known butterfly effect has been scientifically disproved, the far-reaching impact of business decisions is undisputed. Thus, knowledge graphs serve as a basis for a decision support or

recommendation engine for management decisions by accessing structured and unstructured information.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Knowledge graphs are basically Business Digital Twins of a company that represent processes, data models structures and business rules of the organization. One might describe it as a dense cloud of connected business objects.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

Many customers assume knowledge graphs to be “just another fancy database” at first sight.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

We provide so-called “incubation workshops” to demonstrate the power of knowledge graphs for different parts of a company. Different approaches are required depending on the position of the audience.

Sometimes we call it a “google-like” search engine for the enterprise which often helps to get started with potential users, even if this is only one of many features of a semantic knowledge graph. But I think the biggest benefit is making knowledge and even AI-related knowledge processing easy, accessible, and understandable for everyone.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

Understanding and trust. Is this just a new hype or a disruptive technology?

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

An organization has to redefine knowledge engineering and management roles as well as business roles and adjust them to working with an enterprise-wide, schema-free data model.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

In my opinion, semantic AI and knowledge graphs will have a huge impact on and probably become the basis for:

1. **Business models:** knowledge sharing might become a service for some companies and organizations
2. **AI applications**, due to the ability of tracing back recommendations of these applications throughout the knowledge graph
3. **Interacting with knowledge DBs**: storage of real world knowledge will be enhanced by easy access through visual interaction (e.g., VR navigation) and Free speech (Voice recognition and NLP)

## Michael J. Sullivan (Oracle)

**Mr. Sullivan has twenty+ years of experience in professional services as a senior architect and tech lead responsible for designing and implementing custom integrated Customer Experience (CX) and Digital Experience (DX) solutions to Fortune 1000 companies using the Oracle stack.**

*"THE PROMISE OF KNOWLEDGE GRAPHS IS TO BREAK DOWN THESE ARTIFICIAL BARRIERS TURNING THE INDIVIDUAL SILOS INTO SHARDS OF A GREATER WHOLE"*

***What interests you personally about knowledge graphs, what is the fascination with them?***

I have always been interested in connections. In fact, one of my favorite TV shows as a young adult was the BBC series *Connections* with James Burke. Burke's premise for the show is that one cannot consider the development of any particular piece of the modern world in isolation. For me, knowledge graphs are the only technology we have that gets close to that ideal.

***Which concrete business problems can be solved with this approach?***

Any of the typical left-hand vs. right-hand problems that plague all enterprises. And all of these endemic issues have one thing in common:

siloed applications—more often than not with duplicated functions and data. Currently, the only way to get Marketing, Sales, Commerce, Social, Product, Service, and HR on the same page is to orchestrate multiple meetings between the various groups! The promise of knowledge graphs is to break down these artificial barriers turning the individual silos into shards of a greater whole.

***Do you see knowledge graphs more as data or as a process that links business objects?***

Definitely more of an iterative process of synchronizing and harmonizing information over time. Increasingly, I am viewing knowledge graphs as a sort of registry/data-catalog/data-dictionary (take your pick) of all relationships within the enterprise. The data will remain in its native form (e.g., relational, NoSQL, Hadoop, whatever) but the need for mastering the silo's schema would be greatly diminished or even eliminated. However, without harmonized taxonomies and ontologies, metadata—particularly domain-specific metadata derived from silos—by itself is of limited value.

***What do customers usually think of first when they are introduced to the term 'knowledge graph'?***

Unfamiliarity = Perceived Risk. Few middle-managers are willing to take the initiative to embark on such a project given that (to them) the downsides are obvious while the potential upsides appear fuzzy and unclear, with no discerned ROI.

***How have you been able to inspire potential users to take a closer look at knowledge graphs so far?***

Been a slow go, but it is getting easier.

***What is the biggest challenge in developing organizations to bring AI applications into production?***

Frankly, not being able to come up with legitimate use cases. And I credit that to approaching the problem in a waterfall manner. The issue is that we

don’t know what we don’t know. As such, I feel the best approach is to create a framework where collaboration and sharing of knowledge is facilitated. We can’t predict what the result of such collaboration will be, but we need only look to the birth of the Internet (an earlier example of exponential sharing and collaboration) to see the potential for explosive growth and opportunity.

***To position knowledge graphs as a central building block of an AI strategy, what are the essential changes an organization has to cope with?***

Being able to understand and discover the various serendipitous connections and relationships between all your data prior to implementing an AI strategy is going to be a safe bet—one that will reduce risk and increase the likelihood of success. Further, traditional graph analytics such as PageRank, PathFinding, CommunityDetection, and PatternMatching, might be all that is necessary to implement rather than a full-scale AI project (depending on your use cases of course). As such, it behooves us to put the data and metadata into a graph first—not only to better understand what we are trying to achieve but also provide a more flexible and agile architecture for performing graph analytics together with machine learning and traditional business intelligence.

***What is your personal opinion about the future of Semantic AI and Knowledge Graphs, where do we stand in 10 years and what developments have we seen until then?***

Cloud infrastructures are going to facilitate an explosion of citizen-developer and citizen-data-analyst self-served analytics, sharing of data, and collaboration. This in turn will be a huge strategic advantage to those enterprises able to take advantage of such benefits. A practical requirement will be to maintain the data where it is—moving terabytes of data is simply a non-starter for a variety of reasons. Thus, in most cases we will be limited to extracting just the metadata. But that metadata can be aggregated and enriched over time into a virtual enterprise-wide semantic view of all data—a true single source of truth. However, a huge blocker to achieving that vision is privacy. Currently most organizations have little leeway with regard to how they are able to use customer data—in effect their hands are tied. Yet having those insights will ultimately be beneficial to both the customer and the enterprise. This needs to be resolved if we are to make any progress in

this arena.

# PART 6: THE FUTURE OF KNOWLEDGE GRAPHS

## READ THE TEA LEAVES

# AI and Knowledge Technologies in a Post-Corona Society

As of this writing, we've entered the fourth week of quarantine and are probably only at the beginning of what has become the world's largest crisis since World War II. In a few months, the fog will lift and we will be able to see more clearly not only the destruction caused by the coronavirus, but perhaps also the ways in which it has changed things for the better. One thing is certain, the outbreak of the pandemic will change all of our lives forever: our patterns of social behavior, the way we work together—now and in the future—how we research and search for solutions as a global community, how we reorganize our supply chains, and how we will think about big data, surveillance and privacy.

A key observation right at the beginning: What we're seeing right now is how central an infrastructure called the Internet has become to ensuring the continued existence of many of our vital systems around the world, and how crucial it is to have data, information, news, and facts that can be trusted, accessed, processed, and networked at lightning speed. Many people, even entire industries, did not see it that way until very recently, but now it has probably become clear to everyone.

"As humans have spread across the world, so have infectious diseases. Even in this modern era, outbreaks are nearly constant, though not every outbreak reaches pandemic level as the Novel Coronavirus (COVID-19) has."[153] Virus outbreaks are inevitable, but next time we should be better prepared, and for that we should build systems and societies based on trust.

The post-corona era will divide the world in two: into countries where the acceleration of digital transformation is based on recognizing the importance of evidence-based decision-making, the need for data quality, and the crucial importance of linking people and organizations across borders to benefit from explainable AI—and into another half, which uses Big Data and AI to build societies that are centrally governed by a few, using pandemics as a pretext to increasingly instrumentalize people as data points.

In which environment do smart networking technologies unfold—where the

benefits of people and citizens are at the center, where the diversity of ideas, knowledge, and research is stimulated in such a way that sustainable and countable results are achieved? Where are resilient societies[154] emerging in the post-corona era, developing strategies that will be effective in the next —possibly even more catastrophic—pandemic? Let's take a look at some of the possible building blocks of a post-corona society and at upcoming trends that we should pay attention to in order to shape our new future in a humane way.

### *Self-servicing Based on Explainable AI*

The economy and public administration are now in turmoil and under enormous pressure to cut costs, and at the same time, a door has opened that is pushing the use of AI to provide cost-saving self services.

Digital self-service services will be ubiquitous, they will support many more interactions between citizens and public administration than today, they will complement existing e-learning services (for teachers and students), they will serve younger and older people, in health care, to acquire financial literacy or even to plan the next trip to be economically and ecologically balanced, in short: conversational AI will help to make the "right" decisions.

As described above, however, this is happening in different countries under diametrically different circumstances. While in some regions of the world explainable AI (XAI) and Big Data are being developed for peoples' benefit , in other regions this is happening under very different auspices: by using knowledge graphs, complete digital twins of citizens are being generated and ultimately used against the individual in order to prevent individual behaviour, to destroy diversity, to make the future allegedly "predictable".

Gartner recommends that Government CIOs must "leverage the urgency created by the virus outbreak to accelerate the development of data-centric transformation initiatives", and further on they state that "the increased need for transparency and improved decision making is putting greater emphasis on data centricity, while exacerbating ethical issues."[155]

### *Fight Fake News and Hate Speech*

To a large extent, the degree of the pandemic is due to the fact that even before the outbreak of the crisis, but primarily during it, false news and opinions were constantly spread via fake news spinners like Facebook and other social networks, but also via so-called 'established' media. As mentioned above, the foundation of a resilient society and its organizations is built on trust. Every wrong message and every hate posting undermines this foundation a little bit more. And it was during the pandemic that the vulnerability of digital systems in this respect became apparent, with Facebook having to send home thousands of content moderators while at the same time relying on AI algorithms to ensure that false messages like medical hoaxes could not spread virally across the platform. Facebook's CEO Mark Zuckerberg acknowledged the decision could result in "false positives," including the removal of content that should not be taken down.[156]

Considering that even big data technology giants have to employ thousands of people who have to manually classify their content, one can easily deduce how impossible it will be—at least in the near future—to rely on any AI without the human-in-the-loop (HITL). The approaches to combat fake news and hate speech will be a mixture of AI, HITL, and stricter policies and regulations. Let's stop trusting tech giants who have told us over and over again how resilient their AI algorithms are. The virus revealed their limitations within days.

### *HR at the Heart of Learning Organizations*

Qualified employees and human resources will become increasingly important in a post-corona society and its organizations that want to base their values and business models not only on data, but above all on knowledge, in response to increasingly dynamic environments. Many organizations will have learned at least one thing from the Corona pandemic: self-motivated, self-determined, networkable and knowledgeable employees form the foundation of every company, one which can remain resilient and capable of action even in times of crisis. While some have closed their borders and put up their blinders, others have sought out collaborators and have intensified global networking, especially within the pharmaceutical industry. "While political leaders have locked their borders, scientists have been shattering theirs, creating a global collaboration unlike any in

history."[157]

Paradoxically, where networking is becoming more important, the human being is again at the centre, and on a level above this, the "learning organisation"[158] now comes into play.

> ***"It is not the strongest of the species who survive, nor the most intelligent; rather it is those most responsive to change."***
>
> —CHARLES DARWIN

HR management in a learning organization can benefit from semantic AI and knowledge graphs in many ways: semi-automated and more accurate recruitment, more precise identification of skills gaps, semi-automatic orchestration of knowledge communities within an organization, working law intelligence based on deep text analytics, e-learning systems based on semantics,[159] job seekers identify opportunities that match their skill sets, etc.

Like all other management tasks, HR Management needs good data to support good decisions. Good data also means that they follow the FAIR principles, i.e., that they are based on a data model that can always adapt to new realities. Graph-based data models are agile and therefore a good fit.

### *Rebirth of Linked Open (Government) Data*

"Linked Open Data" experienced its first heyday around 2010, when organizations around the world and government bodies in particular—at least in the long term and in terms of society—recognized and invested in the added value of open data. It has since become clearer that added value is created when data is based on interoperable standards and is therefore machine-readable across borders. For example, even in 2015 the European Commission still looked optimistically into the future and announced in their study on the impact of re-use of public data resources that "The total market value of Open Data is estimated between €193B and €209B for 2016 with an

estimated projection of €265B to €286B for 2020, including inflation corrections."[160]

Expectations were probably very high and since then, the Open Data movement in general has stagnated and what the 'Global Open Data Index' stated in its last report in 2017[161] continues to be the main obstacle to overcome before we can make use of open data on a large scale:

- Data findability is a major challenge and a prerequisite for open data to fulfill its potential. Currently, most data is very hard to find.
- A lot of 'data' *is* online, but the ways in which it is presented are limiting their openness. Governments publish data in many forms, not only as tabular datasets but also visualisations, maps, graphs, and texts. While this is a good effort to make data relatable, it sometimes makes the data very hard or even impossible for reuse.

The scientific community is already doing better, which has paid off during the pandemic. By applying the FAIR principles to their data, such as the open research data set COVID-19,[162] which contains the text of more than 24,000 research papers, or the COVID-19 image data collection,[163] which is supporting the joint development of a system for identifying COVID-19 in lung scans, a cohort of data scientists from around the world has been brought together to achieve a common goal.

Governments and public administrations would be well advised to finally learn from science and, after years of chaotic Open Data efforts, to finally bring their data strategies to a level that takes into account the FAIR principles, and thus Semantic Web standards.[164]

### *The Beginning of a New AI Era*

## *"DEEP LEARNING IS AI FOR THE GOOD WEATHER"*

Before the outbreak of the pandemic, AI had been heralded as a great promise of salvation, and its litmus test: the virus. So could AI pass this test? Yes and no. COVID-19 has turned reality and the future upside down, and with it all the models that were trained before the outbreak.[165]

The COVID-19 crisis has exposed some of the key shortfalls of the current state of AI. Machine learning always requires a large amount of historical data, and this data is not available at the beginning of a pandemic, or more generally, during times of change. By the time they are available, it is often too late. So Deep Learning is AI for the good weather, but what we need is an AI that can learn quicker and can produce answers to questions, not only predictions based on obsolete data.

This can only work when AI can make use of human knowledge and creativity, and is able to make abstractions. Thus, AI systems need support from machine readable knowledge models; additionally, collaboration is key! “Efforts to leverage AI tools in the time of COVID-19 will be most effective when they involve the input and collaboration of humans in several different roles.”[166]

This all requires a major reworking of our AI architectures, which should be based on the Semantic AI design principle.

## *"ONLY BY APPLYING THE FAIR AND HITL PRINCIPLES TO AI WE CAN BRING THIS INTO BALANCE"*

For everyone’s safety, the use of personal health data will experience an unprecedented proliferation and it is imperative that it is based on the HITL principles, otherwise we will either live in societies that are underperforming in combating pandemic outbreaks or other crises, or that are overperforming in surveillance.[167] Only by applying the FAIR and HITL principles to AI we can bring this into balance. This must be placed in an appropriate legal framework and should become the cornerstones of a new AI era.

# New Roles: The Rise of the Knowledge Scientist

The still young discipline of the management and governance of knowledge graphs is gradually beginning to consolidate on the basis of concrete project experience. It has been clearly recognized that the underlying methodology is multidisciplinary and that it cannot simply be covered by existing, often classical roles and skills in information management. Rather, there is a need for new roles in which the "Knowledge Scientist"[168] is to be given a central position because he is able to bring together the two archetypical, sometimes rivalling roles of the "Data Engineer" and the "Knowledge Modeler".

What an enterprise knowledge graph is and how it is created, there are (at least) two different answers to that in the current discourse. These two points of view are often understood as if they were mutually exclusive and incompatible; however, these are two approaches to semantic data modeling that should be combined in the concrete development of a knowledge graph.

For practitioners and potential users, these supposed opposites naturally cause confusion, because the two approaches are often understood as alternatives to each other, if presented in simplified form. Here are the two views in simple words:

**Approach 1—Principle 'Knowledge'**: A knowledge graph is a model of a knowledge domain that is curated by corresponding subject-matter experts (SMEs) with the support of knowledge modelers, e.g., taxonomists or ontologists, whereby partially automatable methods can be used. Knowledge domains can overlap and represent in most cases only a subdomain of the entire enterprise. Knowledge modelers tend to create specific, expressive and semantically rich knowledge models, but only for a limited scope of an enterprise. This approach is mainly focused on the expert loop within the entire knowledge graph lifecycle.

**Approach 2—Principle 'Data'**: A knowledge graph is a graph-based representation of already existing data sources, which is created by data engineers with the help of automatable transformation, enrichment and validation steps. Ontologies and rules play an essential role in this process, and data lineage is one of the most complex problems involved. In this approach, data engineers focus on the automation loop and aim to reuse and

integrate as many data sources as possible to create a data graph. The ontologies and taxonomies involved in this approach provide only the level of expressiveness needed to automate data transformation and integration.

With the principle 'Data', the graph-based representation of often heterogeneous data landscapes moves into the center so that it can roll out agile methods of data integration (e.g., 'Customer 360'), data quality management, and extended possibilities of data analysis. The 'Knowledge' principle, on the other hand, introduces to a greater extent the idea of linking and enriching existing data with additional knowledge as a means to, for example, support knowledge discovery and in-depth analyses in large and complex databases.

So, are these two approaches mutually exclusive? The acting protagonists and proponents of both scenarios look at the same corporate knowledge from two different perspectives. This sometimes seems as if they are pursuing different goals, especially when participants' mindsets can vary significantly.

**The view of 'Knowledge modelers'**: Approach 1 involves knowledge modelers/engineers, computer linguists and partly also data scientists who have a holistic view of data, i.e., they want to be able to link data and bring it into new contexts in order to be able to provide extended possibilities for data analysis, knowledge retrieval, or recommender systems. This is done without 'container thinking', no matter whether information or facts are locked up in relational databases or proprietary document structures, they should be extracted and made (re-)usable. Proponents of approach 1 often assume that the data quality—especially of so-called 'structured data'—is high enough for fully automated approaches, which is seldom the case in reality. Accordingly, the phase of data preparation and data transformation involving ontologies to build a robust nucleus for a knowledge graph at scale is underestimated, thus there is a risk of unnecessarily increasing the proportion of manual work in the long run.

**The view of 'Data engineers'**: Approach 2 mainly employs data engineers who want to solve various problems in enterprise data management, e.g., insufficient data quality, cumbersome data integration (keyword: data silos), etc. This is often done independently from concrete business use cases. Restrictions due to rigid database schemata are a central problem that should

be addressed by knowledge graphs. Data engineers see ontologies as central building blocks of an EKG, sometimes ontologies are even equated with a KG. Taxonomic relationships between entities and unstructured data (e.g., PDF documents) are often ignored and find no or merely a subordinate place in the design of a data engineer's KG, where the danger exists that one might waive existing data sources unnecessarily. Approach 2 therefore, creates a virtual data graph that mirrors existing data virtually 1:1. The focus is more on data integration and better accessibility rather than enriching the data with further knowledge models.

Obviously, both approaches and mindsets have good reasons to work with graph technologies, and they each involve different risks of having produced significant gaps and relying on inefficient methods at the end of the journey to develop a fully-fledged enterprise knowledge graph. The way out is therefore to network both directions of thought and to get the respective proponents out of their isolation. How can this be achieved? How can knowledge modelers, data engineers and their objectives be linked?

A relatively new role has been introduced recently, which is the so-called '**knowledge scientist**'. Knowledge scientists combine the more holistic and connected views of the knowledge modelers with the more pragmatic views of the data engineers. They interact with knowledge graphs, extract data from them to train new models and provide their insights as feedback for others to use. Knowledge scientists work closely together with businesses and understand their actual needs, which are typically centered around business objects and facts about them. Eventually, this results in a more complete and entity-centric view of knowledge graphs.

**Approach 3—Principle 'Entity'**: A knowledge graph is a multi-layered, multidimensional network of entities and introduces a fundamentally new perspective on enterprise data: the entity-centric view. Each layer of a KG represents a context in which a business object, represented by an entity, can occur (Named Graph). Each dimension represents a way to look at an entity that occurs in a particular data source, whether structured, semi-structured, or unstructured. KGs contain facts about entities that can be very concrete but also abstract, and are represented in the form of instance data, taxonomies, and ontologies. In this approach, the knowledge and data perspectives are consolidated and the business users' perspective is included.

**Conclusion:** While some work on linking existing data ("data graphs") and others mainly focus on the development of semantic knowledge models ("semantic graphs"), a third perspective on knowledge graphs, which includes the user perspective has become increasingly important: "entity graphs". The focus is on all relevant business objects including the users themselves, which in turn, should be linked to all facts from the other two layers. This clearly entity-centered view of the knowledge graph ultimately introduces the business view. All the questions that are linked to the respective business objects are formulated by the 'knowledge scientist' and partly answered with the help of machine learning methods, partly by SMEs and then returned to the knowledge graphs.

# Upcoming New Graph Standards

One of the main conflicts around knowledge graphs has always been the discussion which graph model (RDF versus LPG) works better. Both formats have their pros and cons and support different use cases in a better way. The main disadvantage of labeled property graphs has always been that they are not based on standards and therefore you always have lock-in effects no matter which provider you choose.

The W3C hosted a workshop on the standardization of Graph Data in 2019[169] as an attempt to bridge that gap between those different formats (also including SQL).

The development of the Graph Query Language (GQL) goes into the same direction and is a joint project of all major LPG vendors to develop an ISO standard for property graphs, which started in 2017. The most recent addition in this direction is the position paper for RDF*/SPARQL*[170] that proposes a way to overcome one of the main down sides of the RDF data model against LPG, which is the varying complexity to make statements on the edges of a graph or triple (so-called “meta triples”).

So there are initiatives that try to develop a standard for property graphs and on the other hand initiatives to bring the RDF and the LPG model closer together.

**Conclusion:** From our perspective, it is not the question anymore which of those approaches will win in the end, but how long will it take that both approaches will end up in a complete knowledge graph standard that offers the benefits of both approaches in a performant way and is implemented in all stores that are, at the moment, divided by different and partly proprietary data models.

# ADDENDUM: FAQS AND GLOSSARY

# FAQs

### *Why do you think I should be interested in knowledge graphs?*

Knowledge graphs are not just "the new kid on the block," they have matured over many years and are now ready to be used in enterprises on a large scale. The wide range of applications (e.g., improved user experience, efficient HR management, automated recommendation and advisory systems, etc.) that benefit from these technologies is an argument for at least considering them for any type of intelligent application. However, there is also a strategic perspective: the introduction of graphs helps to solve an age-old problem in data management that lies buried in the inability to correctly and automatically reuse and interpret information as soon as it leaves the defined boundaries of a data silo, and all organizations have them in abundance. Knowledge graphs are a game-changer at every level of our overall data and system architecture, and looking forward we can be certain that "the future of work will be augmented. The new work nucleus comes preloaded with artificial intelligence, which constantly improves with a combination of machine learning and knowledge graphs."[171]

Read more in our Why Knowledge Graphs? chapter.

### *How can I measure the business value of knowledge graphs?*

While knowledge graphs will gradually penetrate all levels of an enterprise system architecture, calculating the ROI as a whole will hardly be possible; however, for individual applications that are fed by the knowledge graph, this can in fact be determined and should be definable from the outset.

A smart advisor system that utilizes knowledge graphs can, for example, answer x-percent more customer inquiries to their satisfaction and in turn reduce customer fluctuation by y-percent, or increase a cross-selling rate by z-percent, can quickly pay off the investment.

Read more in our section on How to Measure the Economic Impact of a Enterprise Knowledge Graph.

***Are knowledge graphs created primarily for data visualization and analytics?***

Better data analysis is a field of application for semantic knowledge graphs, where large potential for improvement can be achieved by linking data across silos, rich metadata, additional background information derived from the knowledge graph, and highly structured data based on standards. Data as graphs also offers a fresh interface for analysts who can now approach and interact with data sets in a more intuitive fashion than would be possible with tables alone. Visualizing graphs is an easy win for any graph project and quickly sparks the interest of business users. However, visualization is only the tip of the iceberg, because knowledge graphs can do much more with your data than just visualizing it in a new and fun way.

Read more in our Knowledge Graphs are not just for Visualization chapter.

***Do I have to create a knowledge graph by hand or can this be automated?***

Neither. Both types of activities interact but most parts of a knowledge graph can be generated automatically. Most of the triples found in a graph are either the result of automatic entity extraction or linking, or transformation of (semi-)structured data into RDF. Nevertheless, a solid foundation for the creation of such high quality data graphs can only be established if sufficient time is invested in the creation and maintenance of curated taxonomies and ontologies. But even these steps can be partially automated, e.g., by using text corpus analysis, word embeddings or other language modelling and feature learning techniques derived from natural language processing (NLP).

Learn more in our Knowledge Graph Life Cycle chapter.

***Where can I download or purchase knowledge graphs?***

We distinguish between two types of knowledge graphs: open knowledge graphs and enterprise knowledge graphs (EKGs). Open knowledge graphs are open to the public, are often created and maintained by NGOs, government organizations, or research institutions, and in many cases serve as a basic element for the development of EKGs. A large collection of publicly accessible knowledge graphs is found in the so-called Linked Open Data

Cloud. However, the EKGs are always subject to customization and as you might expect, cannot be downloaded from the Web because they contain what is probably the most important asset of any organization—its knowledge.

However, the reuse of freely available ontologies or taxonomies to build up your own knowledge graph works in many cases and is even recommended, although it is also important to bear in mind that within every organization, there is always enough data that can serve as a basis for creating your own 'seed taxonomy'. Some professional publishers have started to develop strategies to sell their taxonomies and ontologies, but this kind of business is still in its infancy.

Read more in our Reusing Existing Knowledge Models and Graphs chapter.

### *Who in our organization will be working on knowledge graphs?*

The creation and maintenance of knowledge graphs requires a collaborative approach involving a variety of stakeholders. Some of them are established roles in enterprise data management, such as the data engineer, but some new skills and responsibilities typically need to be developed while knowledge graphs are being built. These include professions such as taxonomist, ontologist, MLOps, knowledge graph strategist, semantic web developer, or the knowledge scientist. But don't worry, all of this can be developed step by step as long as the underlying working methods maintain an agile style.

Read more in our section on Personas: too many cooks?

### *How are knowledge graphs related to artificial intelligence?*

Knowledge graphs are at the core of semantic artificial intelligence. Semantic AI fuses symbolic and statistical AI. It combines methods from machine learning, knowledge modeling, natural language processing, text mining and the Semantic Web. It combines the advantages of both AI strategies, mainly semantic reasoning and neural networks. Semantic AI is not an alternative, but an extension of what is mainly used to build explainable AI-based systems today. Knowledge graphs help to create high-quality data sets to be processed by ML algorithms and in return, ML is used to automate the

creation of KGs.

Read more: Machine Learning and Artificial intelligence: Make it explainable

***Which tools do I need to create and run a knowledge graph?***

The tools required for the creation and development of knowledge graphs include taxonomy and ontology management software, data transformation, entity linking and enrichment tools, reasoners, and graph databases. The tools are usually used by different stakeholders with different skills and backgrounds, some of them are more involved in manual graph creation and curation, others support graph development more as part of the automation loop. Low-threshold systems such as simple content editors, where tags can be created and proposed, or card sorting tools as entry points to knowledge graphs, could also be part of the toolbox and should be considered as possible elements of the knowledge graph lifecycle and enterprise architecture in order for knowledge graphs to be rolled out without a hitch.

Read more in our Enterprise Knowledge Graph Architecture chapter.

***What's the difference between a taxonomy and an ontology?***

Taxonomies are used to express linguistic elements of a knowledge graph, including names for concepts (including synonyms) and fundamental relationships between concepts. Ontologies, meanwhile, express the conceptual framework of a knowledge graph by grouping concepts or things into classes and subclasses, expressing possible relationships between them and making axiomatic statements such as "each child can have only one mother." All in all, taxonomies and ontologies express a knowledge domain in a model-like way and are fundamental to transforming existing data and texts into graphs in an automated way.

Read more in our section on Knowledge Organization Systems

***What's the difference between the Semantic Web, linked data and knowledge graphs?***

‘Semantic Web' is the precursor term to 'Knowledge Graph'. Since then, the largely identical concept behind it has also been called 'Linked Data', but essentially all three terms mean the same thing, namely the controlled linking of data. The Semantic Web is based on a multitude of standards and therefore offers the possibility to use interoperable data standards to link and reuse data across departments and organizations. Not all formats for developing knowledge graphs have this feature.

Learn more: Semantic Web

***Are graph databases the same as knowledge graphs?***

No, not at all. Knowledge graphs are data, and it is not just the underlying database that makes a difference. Knowledge graphs introduce a whole range of new methods and standards into an enterprise data management landscape, as well as new roles and tools, but not just a new database.

Read more in our Graph databases chapter.

# Glossary

### *AutoML*

AutoML aims to reduce the need for highly skilled data scientists to create models for machine learning. With an AutoML system, you can instead provide the labelled training data as input and get an optimized model as output. Knowledge models can play an important role in this process, since they contain the 'building instructions' for training models that can be advanced without the participation of data scientists.

Automated machine learning can target different phases of the machine learning process including data preparation, feature engineering, model selection, evaluation metric selection, and hyperparameter optimization.

### *Business Glossary*

A business glossary defines the meaning of business terms and can be made available, retrieved, and looked up within an entire organization or even for a whole industry. Such glossaries allow for a better understanding of key business concepts and terms and also show how vocabulary may differ across segments of an industry or across business functions. Unlike a data dictionary, which is a detailed definition and description of datasets and their fields, a business glossary thus defines business concepts for an organization or an entire industry and is therefore independent of a specific database or vendor.

Business glossaries improve data governance and can typically be used to increase confidence in the data of an organization. Business glossaries can be expressed as part of enterprise taxonomies and thesauri and can be made available as interoperable, machine-readable formats using standards such as SKOS. The Linked Data Glossary[172] or even this small glossary you're currently reading are examples of a business glossary.

### *Enterprise Knowledge Graph (EKG)*

An Enterprise Knowledge Graph (EKG) typically consists of three layers:

1. A domain model of a knowledge domain, created and maintained by knowledge engineers and subject matter experts using machine learning algorithms, providing a structure and common interface for all your data to enable the ‘data graph’ to be created automatically.
2. A data graph that consists of or represents intelligent multilateral relationships in your databases, content, and document repositories, all structured as an additional virtual data layer to link all of your data, even on a large scale, whether structured or unstructured.
3. A user graph, which contains semantic profiles of the users partly automatically derived from user behavior in order to link them with each other and with knowledge and data objects in a targeted manner.

### *Human-in-the-Loop (HITL)*

The human-in-the-loop design principle is a prerequisite to build trustworthy AI platforms providing explainable AI. Building well-formed enterprise knowledge graphs is heavily dependent on the efficiency of the expert loop and user loop being involved.

At present, the added value of AI for humans consists mainly of classifications and non-systemic, one-dimensional predictions based on correlation models. Thus, although the current AI generates short summaries from large amounts of data, it does not provide much evidence for a better understanding of systemic relationships and causalities. Finding a lingua franca (or a usable translation/UI) between humans and AI will take some time, while HITL solutions are a core piece of this puzzle.[173]

### *Inference and Reasoning*

Inference is the derivation of new knowledge (facts, triples) from existing knowledge and axioms.[174]. Based on a set of axioms (TBox), typically expressed by OWL-2 ontologies, and a set of explicit facts (ABox), usually stored in an RDF graph database, a reasoner is able to derive implicit, previously unknown facts.

Reasoning also refers to the ability to decide whether a propositional formula is satisfiable or not and is carried out via a search process involving multiple inferences.

### *Information Retrieval (IR)*

IR deals with computer-aided searches for complex content. Precision and Recall are decisive key figures for an information retrieval system. An ideal system would filter out all relevant records of a document collection after a search query, excluding documents that are not relevant. What is relevant and what is not relevant, however, often depends on the actual information needs of the users, which often can only be formulated vaguely, otherwise one would actually have to know what one does not know. An information retrieval system usually consists of two components: the indexing system and the query system.

### *Knowledge Domain*

Knowledge domains are a way of dividing the entire knowledge of an organization (or society) in such a way that only certain groups of users (typically the domain experts or subject-matter experts) have access to this knowledge. This often leads to the risk of losing connections to other domains and thus to the loss of valuable knowledge. Knowledge domains are usually characterized by their specific semantics. The creation of domain knowledge models, such as ontologies and taxonomies, especially when using interoperable standards like the Semantic Web Standards of the W3C, help to make the closed language and logic systems of knowledge domains more accessible and interpretable for other systems.

### *Know Your Customer (KYC)*

In the financial industry today, KYC is an important element in the fight against financial crime, fraud, and money laundering. KYC programmes obviously benefit from more holistic views on customers which can be created using knowledge graphs.

### *Named Graphs*

Named graphs help to divide large knowledge graphs into subsets that can only be used for specific purposes. This additional context could be also provenance information (to support data lineage) or other such metadata. For example, you might have a named graph that contains facts (triples) about

food from a nutritional perspective and another named graph that only contains sales statistics. However, a thing called "Emmentaler" could have the same URI in both named graphs and can therefore easily be used in analyses that require facts from both named graphs.

### *Natural Language Processing (NLP)*

Natural language processing or NLP, not to be confused with neuro-linguistic programming, is the application of computer-aided techniques for the analysis and synthesis of natural human language. As a branch of artificial intelligence, NLP is used for text mining (also known as text analysis) to transform the unstructured text in documents and databases into structured data suitable for further analysis or for training machine learning algorithms. NLP, embedded in knowledge graphs, unfolds its potential to better deal with the underlying or latent semantics and metadata of texts.

A typical NLP pipeline is a sequence of some of the following steps: sentence splitting, tokenization, regular expression extraction, stop word removal, lemmatization, entity extraction based on ontology/taxonomy, named entity recognition, word sense disambiguation, entity linking/mapping, and text classification.

### *Open-World Assumption (OWA)*

Semantic Web languages make the open-world assumption. OWA is used in systems that are known to contain incomplete information. In contrast to, for example, a booking system, where each booking is supposed to be correct and available, the WWW is an example of a system with typically incomplete information.[175] The lack of information on the Web may only mean that this information has not been made explicit. In essence, from the absence of a statement alone, a deductive reasoner cannot (and must not) infer that the statement is false. For this reason, the Semantic Web uses OWA. The essence of the Semantic Web is the ability to derive new information from existing information.

In enterprises, OWA is of course only partially useful, since in order to ensure data consistency, an at least partially closed world is assumed.

### *Precision and Recall (F1 score)*

The training of named entity extractors or document classifiers are typical machine learning tasks. When classifying between two cases ("positive" and "negative"), there are four possible results of prediction:

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted Positive** | True Positives | False Positives |
| **Predicted Negative** | False Negatives | True Negatives |

To measure the quality of a classifier, there are two important numbers: precision and recall.

- Precision answers the question, "what percentage of positive predictions is true?"
- Recall gives the answer to the question, "out of all the true positives, what fraction of them did we identify?"

The F1 score is a way to combine and balance precision and recall. To achieve a high F1 result, a classifier must have both high precision and high recall.

### *Semantic AI*

Semantic AI[176] fuses symbolic and statistical AI. It combines methods from machine learning, knowledge modeling, natural language processing, text mining and the Semantic Web. It combines the advantages of both AI strategies, mainly semantic reasoning and neural networks. In short, semantic AI is not an alternative, but an extension of what is mainly used to build AI-based systems today. This brings not only strategic options, but also an immediate advantage: faster learning from less training data, for example to overcome the so-called cold-start problem when developing chatbots while providing explainable AI. Gartner[177] states that, "Semantic AI (e.g., ontological models, rule-based systems and graphs) has the advantage of being explainable by design."

## *Semantic Footprint*

*"THE SEMANTIC FOOTPRINT CAN ALSO BE THOUGHT OF AS A DIGITAL ASSET'S 'IMMUNE SYSTEM'"*

The semantic footprint represents the semantics of a business object (e.g., a customer) or a digital asset (e.g., a document) in its entirety. As the sub-graph of a comprehensive Enterprise Knowledge Graph that refers to a specific digital asset, it can be used, for example, as a basis for semantic matchmaking, analysis tasks, or recommender systems.

The semantic footprint can also be thought of as a digital asset's 'immune system'. It helps to shield business objects from unnecessary relationships. The ontologies and taxonomies on which the footprint and a corresponding recommender system are based then serve as a kind of blueprint for the development of this protection mechanism, whereby the importance of explainable AI's use becomes even clearer.

## *Semantic Layer*

The semantic layer serves as the central hub and reference point where all the different metadata systems are mapped and where their meaning is described in a standards-based modelling language. This central data interface can be developed in organizations as an Enterprise Knowledge Graph.

As a common roof for all kinds of data, the semantic layer ensures that the semantics of the data do not remain buried in data silos. It helps to "harmonize" different data and metadata schemas, and different vocabularies. It makes the semantics (meaning) of metadata and data in general explicitly available and to a large extent machine-processable.

---

[1] What is a Knowledge Graph? Transforming Data into Knowledge (PoolParty.biz, 2020), https://www.poolparty.biz/what-is-a-knowledge-graph

[2] Artificial Intelligence Could Be a $14 Trillion Boon to the Global Economy (Fortune.com, 2019), https://fortune.com/2019/10/09/artificial-intelligence-14-trillion-boon-only-if-overcome-one-thing/

[3] LOD cloud diagram containing 1,239 datasets (as of March 2019), https://lod-cloud.net/

[4] Conceptual Graphs for a Data Base Interface (John F. Sowa. In: IBM Journal of Research and Development, 1976), http://www.jfsowa.com/pubs/cg1976.pdf

[5] The Semantic Web (Tim Berners-Lee, James Hendler and Ora Lassila. In: Scientific American, 2001), https://www.scientificamerican.com/article/the-semantic-web/

[6] DBpedia - Global and Unified Access to Knowledge, https://wiki.dbpedia.org/

[7] The GQL Manifesto - One Property Graph Query Language, https://gql.today/

[8] Gartner, Inc: 'Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders' (Ehtisham Zaidi and Guido De Simoni, 2019), https://www.gartner.com/en/documents/3957301

[9] Understand how structured data works (Google, 2020), https://developers.google.com/search/docs/guides/intro-structured-data

[10] Schema.org, https://schema.org/

[11] 2020 Database Predictions and Trends (Patrick McFadin, 2019), https://www.datastax.com/blog/2019/12/2020-database-predictions-and-trends

[12] Amazon Neptune, https://aws.amazon.com/neptune/

[13] SPARQL 1.1 Overview (W3C Recommendation, 2013), https://www.w3.org/TR/sparql11-overview/

[14] Artificial Intelligence and Enterprise Knowledge Graphs: Better Together (Dataversity, 2019), https://www.dataversity.net/artificial-intelligence-and-enterprise-knowledge-graphs-better-together/

[15] Property Graphs: Training Wheels on the way to Knowledge Graphs (Dave McComb, 2019), https://www.semanticarts.com/property-graphs-training-wheels-on-the-way-to-knowledge-graphs/

[16] Graph Query Language GQL, https://www.gqlstandards.org/

[17] The FAIR Guiding Principles for scientific data management and stewardship (Mark D. Wilkinson et al in: Scientific Data, 2016), https://doi.org/10.1038/sdata.2016.18

[18] Gartner, Inc: 'Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders' (Ehtisham Zaidi and Guido De Simoni, 2019), https://www.gartner.com/en/documents/3957301

[19] Hierarchical relationships are typically '*is a*' and '*is part of*' relationships. These cannot be distinguished in SKOS, but this is possible by using additional ontologies.

[20] SKOS Simple Knowledge Organization System - Reference (W3C, 2009), https://www.w3.org/TR/skos-reference/

**[21]** A Comprehensive Analysis of Knowledge Management Cycles (Haradhan Kumar Mohajan, 2016), https://mpra.ub.uni-muenchen.de/83088/

[22] PoolParty GraphViews, https://www.poolparty.biz/poolparty-graphviews

[23] Financial Industry Business Ontology (EDM Council), https://spec.edmcouncil.org/fibo/

[24] EU guidelines on ethics in artificial intelligence: Context and implementation (European Parliamentary Research Service, 2019), https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

[25] Explaining Explanations: An Overview of Interpretability of Machine Learning (Leilani H. Gilpin et al., 2019), https://arxiv.org/pdf/1806.00069.pdf

[26] What is a knowledge graph and how does one work? (Julian Aijal, 2019), https://thenextweb.com/podium/2019/06/11/what-is-a-knowledge-graph-and-how-does-one-work/

[27] PoolParty Semantic Suite, https://www.poolparty.biz/

[28] data.world, https://data.world/

[29] The Digitization of the World - From Edge to Core (David Reinsel et al, 2018), https://www.seagate.com/gb/en/our-story/data-age-2025/

[30] 1 Zettabyte = $10^{21}$ bytes compared to estimated $10^{24}$ stars in the whole universe compared to estimated $10^{15}$ synaptic connections in a human brain

[31] Knowledge Graphs (Aidan Hogan et al, 2020), https://arxiv.org/abs/2003.02320

[32] Automatic cluster labeling through Artificial Neural Networks (Lucas A. Lopes et al, 2014), https://doi.org/10.1109/IJCNN.2014.6889949

[33] Faceted navigation in ecommerce: How it helps customers and SEO (Maria Marinina, 2019), https://www.searchenginewatch.com/2019/03/13/faceted-navigation-in-ecommerce-how-helps-customers-and-seo/

[34] SPARQLing cocktails (PoolParty, 2017), http://integrator.poolparty.biz/sparqlingCocktails/cocktails

[35] Cost of drug development (Wikipedia, 2020), https://en.wikipedia.org/wiki/Cost_of_drug_development

[36] Linked Life Data, SPARQL endpoint (Ontotext), http://linkedlifedata.com/sparql

[37] GoT live explorer, http://graphofthings.org/

[38] The Graph of Things: A step towards the Live Knowledge Graph of connected things (Danh Le Phuoc et al, 2016), https://www.researchgate.net/publication/303789187

[39] Generating Digital Twin models using Knowledge Graphs for Industrial Production Lines (Agniva Banerjee et al, 2017), https://www.researchgate.net/publication/319356723

[40] Semantic Sensor Network Ontology (W3C, October 2017), https://www.w3.org/TR/vocab-ssn/

[41] Three Necessities For Maximizing Your Digital Twins Approach (Jans Aasman, 2019), https://www.forbes.com/sites/forbestechcouncil/2019/11/04/three-necessities-for-maximizing-your-digital-twins-approach/

[42] Gartner, Inc: 'Predicts 2020: Artificial Intelligence — the Road to Production' (Anthony Mullen et al, December 2019), https://www.gartner.com/en/documents/3975770

[43] From Word to Sense Embeddings: A Survey on Vector Representations of Meaning (Jose Camacho-Collados, Mohammad Taher Pilehvar, 2018), https://arxiv.org/abs/1805.04032

[44] What is DITA? (Kimber, 2017), https://www.xml.com/articles/2017/01/19/what-dita/

[45] LinkedIn Economic Graph, https://economicgraph.linkedin.com/

[46] The Fusion of Search and Recommendation Functionalities (PoolParty.biz, 2017), https://www.poolparty.biz/wp-content/uploads/2017/09/Recommender-Engine-Wine-Cheese-Pairing.pdf

[47] HR Recommender (PoolParty.biz, 2020), https://hr-recommender.poolparty.biz/

[48] European Skills, Competences, Qualifications and Occupations, https://ec.europa.eu/esco/

[49] Gartner, Inc: 'Using Conversational AI Middleware to Build Chatbots and Virtual Assistants' (Magnus Revang, October 2019), https://www.gartner.com/en/documents/3970980

[50] Google Knowledge Panel (Google, 2020), https://support.google.com/knowledgepanel

[51] Authentic Wiener Schnitzel Recipe (Jennifer McGavin, 2019), https://www.thespruceeats.com/wiener-schnitzel-recipe-1447089

[52] Get your recipes on Google (Google, 2020), https://developers.google.com/search/docs/data-types/recipe

[53] Gartner, Inc: 'Survey Analysis: Third Gartner CDO Survey—How Chief Data Officers Are Driving Business Impact' (Valerie Logan et al, May 2019), https://www.gartner.com/en/documents/3834265

[54] The Data-Centric Manifesto (2020), http://datacentricmanifesto.org/

[55] K is for Knowledge (George Anadiotis, 2019), https://www.zdnet.com/article/k-is-for-knowledge-application-and-data-integration-for-better-business-using-metadata-and-knowledge-graphs/

[56] GraphQL - A query language for your API, https://graphql.org/

[57] The Accidental Taxonomist, 2nd edition (Heather Hedden, 2016), http://www.hedden-information.com/accidental-taxonomist/

[58] The social economy: Unlocking value and productivity through social technologies. (McKinsey, 2012) https://www.mckinsey.com/industries/high-tech/our-insights/the-social-economy

[59] ALIGNED project: Quality-centric, software and data engineering (ALIGNED consortium, 2018), http://aligned-project.eu/

[60] Introducing the Linked Data Business Cube (Tassilo Pellegrini, 2014), https://semantic-web.com/2014/11/28/introducing-the-linked-data-business-cube/

[61] LOD2 - Creating Knowledge out of Interlinked Data (LOD2 consortium, 2014), https://cordis.europa.eu/project/id/257943

[62] RDF 1.1 Turtle - Terse RDF Triple Language (W3C, 2014), https://www.w3.org/TR/turtle/

[63] RDF 1.1 TriG (W3C, 2014), https://www.w3.org/TR/trig/

[64] SKOS: A Guide for Information Professionals (Priscilla Jane Frazier, 2015), http://www.ala.org/alcts/resources/z687/skos

[65] SKOS Reference document expressed as OWL ontology (W3C, 2009), http://www.w3.org/TR/skos-reference/skos-owl1-dl.rdf

[66] OWL 2 Web Ontology Language (W3C Recommendation, 2012), https://www.w3.org/TR/owl2-overview/

[67] What are Ontologies? (Ontotext), https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

[68] Introducing the Knowledge Graph: things, not strings (Singhl, 2012), https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

[69] Wikidata, https://www.wikidata.org/

[70] KBpedia - Open-Source Integrated Knowledge Structure, https://kbpedia.org/

[71] Basic Formal Ontology (BFO), https://basic-formal-ontology.org/

[72] Linked Open Data: The Essentials (Florian Bauer, Martin Kaltenböck, 2012), https://www.reeep.org/LOD-the-Essentials.pdf

[73] Linked Open Vocabularies (LOV), https://lov.linkeddata.es/dataset/lov/

[74] The Basel Register of Thesauri, Ontologies & Classifications (BARTOC), https://bartoc.org/

[75] How Semantic AI Is Shaking Up Business Models In The Banking Sector (Andreas Blumauer, 2020), https://www.forbes.com/sites/forbestechcouncil/2020/03/12/how-semantic-ai-is-shaking-up-business-models-in-the-banking-sector/

[76] What is FIBO? (EDM Council, 2020), https://spec.edmcouncil.org/fibo/

[77] STW Thesaurus for Economics, https://zbw.eu/stw/version/latest/about.en.html

[78] US SEC XBRL Taxonomies, https://xbrl.us/home/filers/sec-reporting/taxonomies/

[79] Currencies Name Authority List, https://data.europa.eu/euodp/data/dataset/currency

[80] The World Bank Vocabularies, https://vocabulary.worldbank.org/

[81] EuroVoc, https://op.europa.eu/en/web/eu-vocabularies

[82] UNBIS Thesaurus, http://metadata.un.org/thesaurus/

[83] Use of ESCO in Monster ES search engine (European Commission, 2019), https://ec.europa.eu/esco/portal/news/61ef2465-7b91-4c1f-a5a0-afc390f42b90

[84] PermID, https://permid.org/

[85] BioPortal, https://bioportal.bioontology.org/

[86] Open Biological and Biomedical Ontology (OBO) Foundry, http://www.obofoundry.org/

[87] The European Bioinformatics Institute (EMBL-EBI), https://www.ebi.ac.uk/services

[88] Australian Health Thesaurus, https://about.healthdirect.gov.au/australian-health-thesaurus

[89] Healthdirect Australia, https://www.healthdirect.gov.au/

[90] Europeana, https://www.europeana.eu/

[91] VIAF: The Virtual International Authority File, https://viaf.org/

[92] Getty Vocabularies, https://www.getty.edu/research/tools/vocabularies/

[93] BIBFRAME, https://www.loc.gov/bibframe/

[94] WorldCat Linked Data Vocabulary, https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html

[95] Library of Congress Subject Headings, http://id.loc.gov/authorities/subjects.html

[96] The Nomenclature for Museum Cataloging, https://www.nomenclature.info/

[97] Getty Vocabularies SPARQL endpoint, http://vocab.getty.edu/

[98] The CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/

[99] The bibliographic data portal of the National Library of Spain, http://datos.bne.es/

[100] British National Bibliography Linked Data Platform, https://bnb.data.bl.uk/

[101] Linked Data Service of the German National Library, https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/LDS/lds.html

[102] ArCo: the Italian Cultural Heritage Knowledge Graph (Valentina Anita Carriero et al, 2019), https://arxiv.org/abs/1905.02840

[103] Sustainable Development Goals Taxonomy, http://metadata.un.org/sdg/

[104] UNBIS Thesaurus, http://metadata.un.org/thesaurus/

[105] REEEP Climate Smart Thesaurus, http://data.reeep.org/thesaurus/guide

[106] GEneral Multilingual Environmental Thesaurus, https://www.eionet.europa.eu/gemet/

[107] AGROVOC, http://aims.fao.org/vest-registry/vocabularies/agrovoc

[108] Climate Tagger, https://www.climatetagger.net/

[109] Semantic Data Service, https://semantic.eea.europa.eu/

[110] GeoName, https://www.geonames.org/

[111] GeoNames Ontology, http://www.geonames.org/ontology/documentation.html

[112] Library of Congress Linked Data Services, https://id.loc.gov/

[113] EU Open Data Portal, https://data.europa.eu/euodp/en/data/

[114] INSPIRE Geoportal, https://inspire-geoportal.ec.europa.eu/

[115] Online vocabulary of the Geological Survey of Austria, https://thesaurus.geolba.ac.at/

[116] Card Sorting to Discover the Users' Model of the Information Space (Jakob Nielsen, 1995), https://www.nngroup.com/articles/usability-testing-1995-sun-microsystems-website/

[117] Organize your materials with the world's most widely used library classification system, https://www.oclc.org/en/dewey.html

[118] Taxonomy Governance Best Practices (Zach Wahl, 2017), https://enterprise-knowledge.com/taxonomy-governance-best-practices/

[119] Harvest Linked Data to Generate a Seed Thesaurus (PoolParty Manual, 2020), https://help.poolparty.biz/pages?pageId=35921550

[120] Efficient Knowledge Modelling Based on Your Text Corpora (PoolParty.biz, 2020), https://www.poolparty.biz/text-corpus-analysis

[121] The Fundamental Importance of Keeping an ABox and TBox Split (Michael K. Bergman, 2009), http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/

[122] Knowledge-based Artificial Intelligence (Michael K. Bergman, 2014), http://www.mkbergman.com/1816/knowledge-based-artificial-intelligence/

[123] Ontology Design Best Practices (Joe Hilger, 2017), https://enterprise-knowledge.com/ontology-design-best-practices-part/

[124] Linked Open Vocabularies, https://lov.linkeddata.es/dataset/lov

[125] R2RML: RDB to RDF Mapping Language (W3C Recommendation, September 2012), https://www.w3.org/TR/r2rml/

[126] Label unstructured data using Enterprise Knowledge Graphs (Artem Revenko, 2019), https://medium.com/semantic-tech-hotspot/label-unstructured-data-using-enterprise-knowledge-graphs-9d63f6f85ae1

[127] Resolving Language Problems (Andreas Blumauer, 2017), https://www.linkedin.com/pulse/resolving-language-problems-part-1-andreas-blumauer

[128] For example: OpenBioLink as a resource and evaluation framework for evaluating link prediction models on heterogeneous biomedical graph data, https://github.com/OpenBioLink/OpenBioLink

[129] Data fusion (Jens Bleiholder, 2009), https://dl.acm.org/doi/10.1145/1456650.1456651

[130] Find a SPARQL quick tutorials at https://docs.data.world/tutorials/sparql/

[131] A New Hope: The Rise of the Knowledge Graph (Jem Rayfield, 2019) https://www.ontotext.com/blog/the-rise-of-the-knowledge-graph/

[132] Validating RDF Data (Gayo et al, 2018), https://book.validatingrdf.com/

[133] SHACL - Shapes Constraint Language (W3C, 2017), https://www.w3.org/TR/shacl/

[134] http://graphdb.ontotext.com/documentation/standard/reasoning.html

[135] How to Build a Knowledge Graph in Four Steps: The Roadmap From Metadata to AI (Lulit Tesfaye, 2019), https://idm.net.au/article/0012676-how-build-knowledge-graph-four-steps-roadmap-metadata-ai

[136] How to Build a Knowledge Graph in Four Steps: The Roadmap From Metadata to AI (Lulit Tesfaye, 2019), https://idm.net.au/article/0012676-how-build-knowledge-graph-four-steps-roadmap-metadata-ai

[137] Upper ontology, (Wikipedia, 2020), https://en.wikipedia.org/wiki/Upper_ontology

[138] Cool URIs for the Semantic Web (Leo Sauermann, 2008), https://www.w3.org/TR/cooluris/

[139] Linked Data: Evolving the Web into a Global Data Space (Tom Heath and Christian Bizer, 2011)

http://linkeddatabook.com/editions/1.0/#htoc11

[140] What is a data catalog? (TechTarget, https://searchdatamanagement.techtarget.com/definition/data-catalog)

[141] Gartner, Inc: 'An Introduction to Graph Data Stores and Applicable Use Cases' (Sumit Pal, January 2019), https://www.gartner.com/document/3899263

[142] Oracle as a RDF Graph, https://www.oracle.com/database/technologies/spatialandgraph/rdf-graph-features.html

[143] Foundations of Modern Query Languages for Graph Databases (Renzo Angles et al, 2017), https://doi.org/10.1145/3104031

[144] What is RDF Triplestore? (Ontotext, 2020), https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/

[145] Reifying RDF: What Works Well With Wikidata? (Daniel Hernández, Aidan Hogan, and Markus Krötzsch, 2015), http://ceur-ws.org/Vol-1457/SSWS2015_paper3.pdf

[146] Position Statement: The RDF* and SPARQL* Approach to Annotate Statements in RDF and to Reconcile RDF and Property Graphs (Olaf Hartig, 2019), http://blog.liu.se/olafhartig/2019/01/10/position-statement-rdf-star-and-sparql-star/

[147] Gartner, Inc: 'An Introduction to Graph Data Stores and Applicable Use Cases' (Sumit Pal, January 2019), https://www.gartner.com/document/3899263

[148] OGC GeoSPARQL standard: https://www.ogc.org/standards/geosparql/

[149] Eclipse RDF4J, https://rdf4j.org/

[150] Apache Jena, https://jena.apache.org/

[151] Contextualizing Airbnb by Building Knowledge Graph (Xiaoya Wei, 2019), https://medium.com/airbnb-engineering/b7077e268d5a

[152] Introducing the Content Graph Explorer (Ian Piper, 2018), http://www.tellurasemantics.com/content-store/introducing-the-cge

[153] Visualizing the History of Pandemics (Nicholas LePan, 2020), https://www.visualcapitalist.com/history-of-pandemics-deadliest/

[154] After corona: The Resilient Society? (Zukunftsinstitut, 2020), https://www.youtube.com/watch?v=g0lncocYIiY

[155] Gartner, Inc: 'How COVID-19 Will Impact Government Digital Transformation and Innovation' (Andrea Di Maio, Ben Kaner, Michael Brown, 2020), https://www.gartner.com/en/documents/3982374

[156] Facebook sent home thousands of human moderators due to the coronavirus. Now the algorithms are in charge (The Washington Post, 2020), https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/

[157] Covid-19 Changed How the World Does Science, Together (The New York Times, 2020), https://www.nytimes.com/2020/04/01/world/europe/coronavirus-science-research-cooperation.html

[158] Building a Learning Organization (Olivier Serrat, 2017), https://link.springer.com/chapter/10.1007/978-981-10-0983-9_11

[159] A Survey of Semantic Technology and Ontology for e-Learning (Yi Wang, Ying Wang, 2019), http://www.semantic-web-journal.net/content/survey-semantic-technology-and-ontology-e-learning

[160] Creating Value through Open Data (European Commission, 2015), https://www.europeandataportal.eu/en/highlights/creating-value-through-open-data

[161] The State of Open Government Data in 2017 (Danny Lämmerhirt et al, 2017), https://index.okfn.org/insights/

[162] COVID-19 Open Research Dataset (Allen Institute for AI), https://pages.semanticscholar.org/coronavirus-research

[163] COVID-19 image data collection, https://github.com/ieee8023/covid-chestxray-dataset

[164] For example: SN SciGraph, https://scigraph.springernature.com/

[165] What Happens to AI When the World Stops(COVID-19)? (Ian Rowan, 2020), https://towardsdatascience.com/cf905a331b2f

[166] AI can help with the COVID-19 crisis - but the right human input is key (Matissa Hollister, 2020), https://www.weforum.org/agenda/2020/03/covid-19-crisis-artificial-intelligence-creativity/

[167] COVID-19 and Digital Rights (The Electronic Frontier Foundation), https://www.eff.org/issues/covid-19

[168] Who should be responsible for your data? The knowledge scientist (Juan Sequeda, 2019), https://www.infoworld.com/article/3448577/who-should-be-responsible-for-your-data-the-knowledge-scientist.html

[169] W3C Workshop on Web Standardization for Graph Data, https://www.w3.org/Data/events/data-ws-2019/

[170] https://blog.liu.se/olafhartig/2019/01/10/position-statement-rdf-star-and-sparql-star/

[171] Gartner, Inc: 'Predicts 2020: Digital Workplace Applications Led by the New Work Nucleus' (Lane Severson et al, December 2019), https://www.gartner.com/en/documents/3975994

[172] Linked Data Glossary (W3C, 2013), https://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/

[173] Gartner, Inc: Design Principles of Human-in-the-Loop Systems for Control, Performance and Transparency of AI (Anthony Mullen, Magnus Revang, Pieter den Hamer, 2019), https://www.gartner.com/en/documents/3970687

[174] What is inference? (Ontotext, 2020), http://graphdb.ontotext.com/free/devhub/inference.html

[175] Introduction to: Open World Assumption vs Closed World Assumption (Juan Sequeda, 2012), https://www.dataversity.net/introduction-to-open-world-assumption-vs-closed-world-assumption/

[176] Six Core Aspects of Semantic AI (Andreas Blumauer, 2018), https://www.datasciencecentral.com/profiles/blogs/six-core-aspects-of-semantic-ai

[177] Gartner, Inc: Design Principles of Human-in-the-Loop Systems for Control, Performance and Transparency of AI (Anthony Mullen, Magnus Revang, Pieter den Hamer, 2019),

https://www.gartner.com/en/documents/3970687