Chapter 5
Ecosystems

5.1 Introduction

Much of this book has described knowledge graphs and their construction at a fairly technical level. In the introduction, we argued that domain-specific knowledge graphs have started to come into their own, using examples such as publications and academia, products and e-commerce, and social causes such as disaster relief. In this chapter, we take a much broader view of knowledge graphs and their impact. Specifically, we attempt an answer to questions such as, *how high has adoption of knowledge graphs been, and in what contexts? What bodes for the future of knowledge graphs?* Although there is a lot more still to come in knowledge graph research, some crystallization has occurred over the last few years (in some cases, decades), which will be the focus of this chapter.

5.2 Web of Linked Data

The Web has provided inarguable benefits but until recently, the same principles that enabled the Web of *documents* to emerge and succeed have not been applied to a hypothetical Web of *data*. Traditionally, data published on the Web was made available as raw CSV or XML dumps, marked up as HTML tables, thereby sacrificing much of its structure and semantics, or in other structured or semi-structured formats that were not intuitive for humans to read or understand (in their raw form). On the conventional Web, driven by hypertext, the *nature* of the relationship between two linked documents is implicit, as the data format, i.e. HTML, is not sufficiently expressive to enable richer semantics and modalities e.g., determining that individual entities described in a particular document are connected by typed links to related entities in the same (or other) documents. In the real world, in contrast, such relationships (between entities) form the basis for

M. Kejriwal, *Domain-Specific Knowledge Graph Construction*, SpringerBriefs
in Computer Science, https://doi.org/10.1007/978-3-030-12375-8_5

knowledge and interaction. A guiding question has been, *can knowledge graphs provide the backbone for enabling such rich, real-world like semantics on the Web?*

There are multiple strands of evidence to indicate that the answer may be yes, although the power of knowledge graphs in enabling semantics on the Web is not limitless [18, 163]. Nevertheless, knowledge graphs published in recent years have contributed greatly to an evolution of the Web from a global information space of linked documents to one where both documents and data are interlinked. Underpinning this evolution is a set of *best practices*, called *Linked Data* [25, 26, 75], for publishing and connecting structured data on the Web. The adoption of the Linked Data best practices has lead to the extension of the Web with a global data space connecting data from diverse domains such as people, books, scientific publications, music, proteins, drugs, statistical and scientific data, and reviews to only name a few. Such a Web enables new application types. There are generic Linked Data browsers which allow users to start browsing in *one* data source and then navigate along links into *related* data sources, analogous to how one could start on an HTML webpage on the conventional Web and then use it to browse to completely different webpages, hosted on servers across the world. There are Linked Data search engines that crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data, similar to queries posed over databases. The Web of Data also opens up new possibilities for *domain-specific* applications. Unlike Web 2.0 mashups, which work against a fixed set of data sources, Linked Data applications operate on top of an unbound, global data space. This enables them to adapt and deliver more complete answers as new data sources appear on the Web.

In its simplest form, Linked Data is about using the Web to create typed links between multi-source data elements such as concepts, entities and properties. These multiple sources may be as diverse as databases maintained by two organizations in different geographical locations, or simply heterogeneous systems within a single umbrella organization that have not traditionally been interoperable at the data level because of problems such as varying schemas, data types etc. Technically, Linked Data refers to data published on the Web in such a way that it is not only machine-readable, but its meaning is explicitly defined ('semantics') [163], it is linked to other external data sets, and can be linked *to* from external data sets.

While the primary units of the hypertext Web are HTML (HyperText Markup Language) documents connected by *untyped* hyperlinks, Linked Data relies on RDF (Resource Description Framework) documents [135]. However, rather than simply connecting these documents, Linked Data uses RDF to make typed statements that link arbitrary things in the world. The result, referred to as the Web of Data throughout this chapter, may more accurately be cast as a Web of *things* in the world, *described* by data on the Web.

5.2.1 Linked Data Principles

In the introduction we stated that Linked Data is a set of four best practices for publishing structured data on the Web [26]. Below, we state these four principles. The technology stack used for implementing these principles and publishing the data is described next, followed by the impact of the Linking Open Data (LOD) project [11], a direct consequence of the widespread adoption of these four principles.

1. **Use Uniform Resource Identifiers (URIs) as names for things.** Even though we primarily think of them as 'web addresses', URIs[1] are much more than just Uniform Resource Locators (URLs). In the broadest sense, a URI imposes constraints on, and sets a *standard* for [17], naming entities and units of data that people want to publish on the Web. In the case of HTML webpages, a URL serves nicely as the 'name' of the page. By using similar, albeit broader, standards for naming things, the first principle essentially ensures that we do not invent a new naming system from scratch. There are other benefits associated as well, as the second and third principles illustrate.
2. **Use Hypertext Transfer Protocol (HTTP) URIs so that the names can be looked up.** By associating HTTP lookup with URIs, the second principle ensures that the name of a thing is *dereferencable*. One simple way to do so is to ensure that URIs are also URLs. However, recall that the goal of Linked Data is to describe actual things, not just the description of things. By using techniques such as re-direction in conjunction with the first and second principles, it is possible to maintain this distinction. Intuitively, one could use a URI (not a URL) to name the thing itself, but when dereferenced, a re-direction could be used to direct the user to a URL which describes the thing. This is an elegant, rigorous way of ensuring that the names of things, as well as the descriptions of those things, could co-exist as separate artifacts on the Web.
3. **When a URI is looked up, provide useful information by using established standards such as RDF and SPARQL to publish and access information.** Resource Description Framework (RDF) and the SPARQL query language are important standards that have been developed over more than a decade by long-time researchers in the Semantic Web and Description Logics communities. The third principle ensures that when a URI is looked up, the data is not delivered in some ad-hoc format (e.g., as an Excel file), but instead conforms to well-established, open standards that can be consumed in a predictable way by a machine. Because the first and second principles ensure the use of HTTP and URIs, it is easier than it would be otherwise to implement the third principle. This also illustrates that the rules are not necessarily independent but build upon each other for effectiveness.

[1] In actuality, the first principle if even broader, allowing the use of *internationalized* resource identifiers rather than just URIs for naming things.

4. **Include links to other URIs, so that more relevant things can be discovered through navigation.** In a previous chapter, we covered the problem of Entity Resolution, which was a step designed to ensure that two or more entity 'mentions' referring to the same underlying entity would get 'resolved'. The mechanics on how such a resolution would happen, once the co-referent entities have been identified, were not described, since there is no one best practice. In the Linked Data scenario, a practitioner could simply publish an additional triple linking an entity in their dataset to equivalent entities in other datasets already existing as Linked Data. For example, as we cover later, DBpedia has emerged as a nexus for the openly published Linked Data on the Web, and since most entities in Wikipedia are included in DBpedia [8], linking entities in a dataset to DBpedia can often productively fulfill the fourth Linked Data principle. However, we also note that, while ER can be an important and well-defined mechanism for establishing links between entities in two different datasets, other relations can also be used. The knowledge graph embedding (KGE) techniques that we covered in the previous chapter could be a potent tool in this direction.

5.2.2 Technology Stack

The principles above illustrate that Linked Data is dependent on two technologies fundamental to the Web itself [27]: Uniform Resource Identifiers (URIs) and the Hypertext Transfer Protocol (HTTP). As we described earlier in the context of the first Linked Data principle, while Uniform Resource *Locators* (URLs) have become familiar as addresses for documents that can be located on the Web, Uniform Resource *Identifiers* provide a more generic means to identify any entity that exists in the world.

In the context of the second principle, where entities are identified by URIs using schemes such as http:// and https://, they can be looked up by dereferencing the URI leveraging the HTTP protocol. Thus, the HTTP protocol provides a simple, yet universal, mechanism either for retrieving resources that *can* be serialized as bytes, or retrieving (e.g., by using re-direction) *descriptions* of entities that cannot *physically* be uploaded and sent across networks.[2]

URIs and HTTP are supplemented by the RDF model, which is critical to implementing the vision of the Semantic Web and Linked Data. The use of RDF, and other technologies like SPARQL that execute on top of RDF triplestores to enable access to the data, is in response to the third Linked Data principle which requires information retrieved to be useful (importantly, both to humans and machines).

[2] For example, one could use the protocol for retrieving the description of a book, since the protocol cannot be used for sending the book itself across a network.

Although HTML allows us to structure and link documents on the Web, RDF provides a generic, graph-based data model with which to structure and link data that describes things (i.e. entities) in the world and the varied properties (typed links) that exist between entities. These typed links can have pre-defined semantics, such as owl:sameAs, since they come from a standard (widely used) upper-level vocabulary like SKOS, Dublin Core or RDFS [5]. These higher-level vocabularies are especially useful in facilitating the re-use of ontological terms and properties, ensuring more homogeneity than might be found. For example, properties like owl:sameAs are overwhelmingly used to capture, and publish, the results of Entity Resolution [96] and fulfill requirements such as the fourth Linked Data principle.

5.2.3 Linking Open Data

Because the Linked Data principles are recommended best practices, their success can only be measured in terms of impact and adoption. Perhaps the most visible evidence of impact has been an on-going, decentralized and international effort called the Linking Open Data (LOD) project (Fig. 5.1),[3] which has been described as 'a grassroots community effort founded in January 2007 and supported by the W3C Semantic Web Education and Outreach Group[4]'. The main goal of the effort is to *bootstrap* the Web of Data, and the adoption of the Linked Data principles, by identifying existing, open-license datasets, converting these datasets to RDF in accordance with Linked Data principles, and publishing them on the Web. An auxiliary goal is to facilitate more publishing of such datasets, with the hope that they become discoverable and usable by virtue of following the principles (especially the fourth principle, which encourages inter-linking).

Historically, the earliest participants (still accounting for a major portion of activity on LOD) were university academics, and small companies looking to gain a competitive advantage with high-risk technology. However, LOD has since become considerably more diverse, with significant current involvement from major players in media, government and tech such as the BBC, Thomson Reuters, New York Times and the Library of Congress. We posit that this growth is enabled by the open nature of the project, where anyone can participate simply by publishing a dataset according to the Linked Data principles and by interlinking it with existing datasets (a special case of the fourth principle). Although the growth is not as superlinear anymore as it was in the early stages of LOD, the ecosystem has remained popular. In the next section, we describe one of the success cases (DBpedia), which has found adopters across the spectrum in natural language processing, Semantic Web, and knowledge discovery.

[3] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[4] http://www.w3.org/2001/sw/sweo/

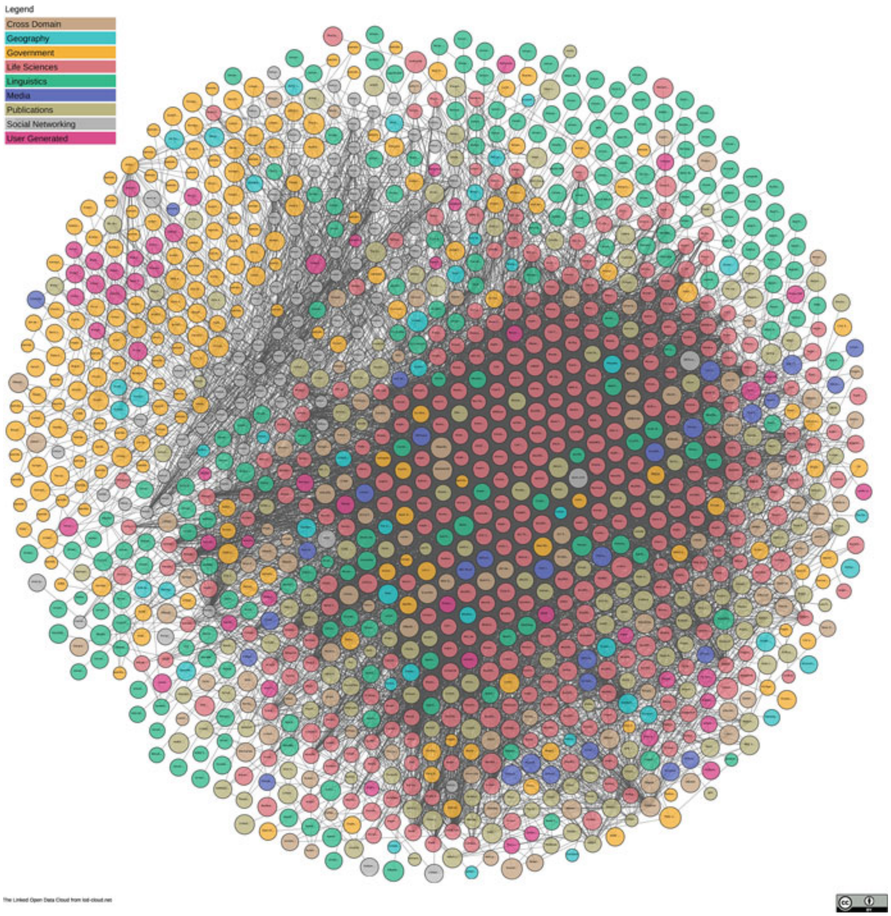
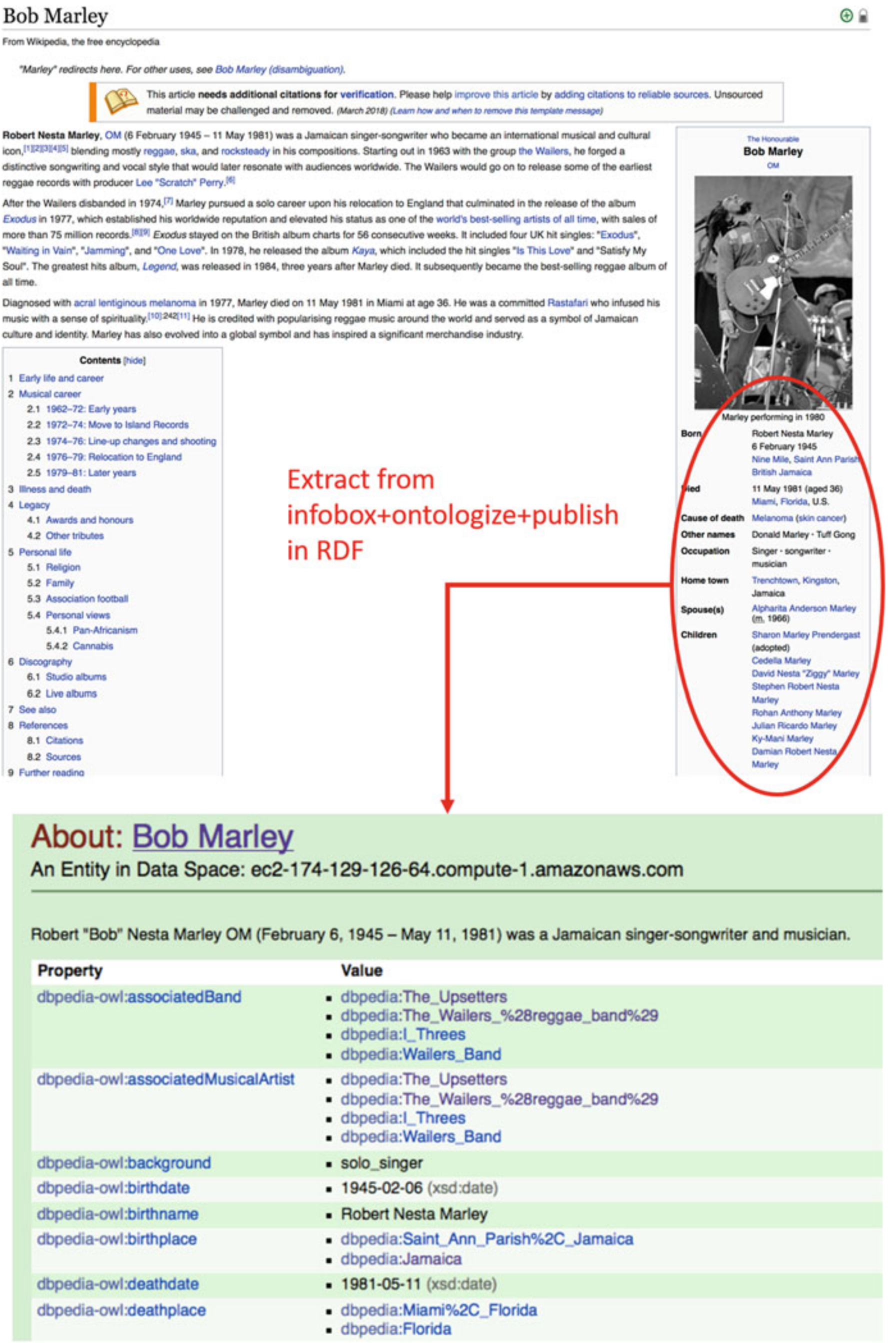


Fig. 5.1 A visualization of LOD datasets. Each node is a dataset and links represent connections between datasets in accordance with the fourth Linked Data principle. Colors represent domains e.g., social networking datasets are in grey. (Image courtesy of lod-cloud.net)

5.2.4 Example: DBpedia

The best example, and most well-known outcome of LOD, has been the DBpedia knowledge graph [8]. DBpedia was an early effort that sought to leverage the structured information on Wikipedia, which is itself a community-powered and crowdsourced encyclopedia. Figure 5.2 provides some intuition on how DBpedia would represent the entity ‘Bob Marley’ by extracting information from Bob Marley’s infobox on his Wikipedia page, ontologizing it with respect to the DBpedia ontology, and rendering it as RDF. All four Linked Data principles are obeyed in this transformation process. DBpedia is available both as RDF dumps, and as a queryable SPARQL endpoint.



Bob Marley

From Wikipedia, the free encyclopedia

"Marley" redirects here. For other uses, see Bob Marley (disambiguation).

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(March 2018) (Learn how and when to remove this template message)*

Robert Nesta Marley, OM (6 February 1945 – 11 May 1981) was a Jamaican singer-songwriter who became an international musical and cultural icon,[1][2][3][4][5] blending mostly reggae, ska, and rocksteady in his compositions. Starting out in 1963 with the group the Wailers, he forged a distinctive songwriting and vocal style that would later resonate with audiences worldwide. The Wailers would go on to release some of the earliest reggae records with producer Lee "Scratch" Perry.[6]

After the Wailers disbanded in 1974,[7] Marley pursued a solo career upon his relocation to England that culminated in the release of the album *Exodus* in 1977, which established his worldwide reputation and elevated his status as one of the world's best-selling artists of all time, with sales of more than 75 million records.[8][9] *Exodus* stayed on the British album charts for 56 consecutive weeks. It included four UK hit singles: "Exodus", "Waiting in Vain", "Jamming", and "One Love". In 1978, he released the album *Kaya*, which included the hit singles "Is This Love" and "Satisfy My Soul". The greatest hits album, *Legend*, was released in 1984, three years after Marley died. It subsequently became the best-selling reggae album of all time.

Diagnosed with acral lentiginous melanoma in 1977, Marley died on 11 May 1981 in Miami at age 36. He was a committed Rastafari who infused his music with a sense of spirituality.[10]:242[11] He is credited with popularising reggae music around the world and served as a symbol of Jamaican culture and identity. Marley has also evolved into a global symbol and has inspired a significant merchandise industry.

Contents [hide]

1 Early life and career
2 Musical career
 2.1 1962–72: Early years
 2.2 1972–74: Move to Island Records
 2.3 1974–76: Line-up changes and shooting
 2.4 1976–79: Relocation to England
 2.5 1979–81: Later years
3 Illness and death
4 Legacy
 4.1 Awards and honours
 4.2 Other tributes
5 Personal life
 5.1 Religion
 5.2 Family
 5.3 Association football
 5.4 Personal views
  5.4.1 Pan-Africanism
  5.4.2 Cannabis
6 Discography
 6.1 Studio albums
 6.2 Live albums
7 See also
8 References
 8.1 Citations
 8.2 Sources
9 Further reading

The Honourable
Bob Marley
OM

Marley performing in 1980

Born	Robert Nesta Marley 6 February 1945 Nine Mile, Saint Ann Parish British Jamaica
Died	11 May 1981 (aged 36) Miami, Florida, U.S.
Cause of death	Melanoma (skin cancer)
Other names	Donald Marley · Tuff Gong
Occupation	Singer · songwriter · musician
Home town	Trenchtown, Kingston, Jamaica
Spouse(s)	Alpharita Anderson Marley (m. 1966)
Children	Sharon Marley Prendergast (adopted) Cedella Marley David Nesta "Ziggy" Marley Stephen Robert Nesta Marley Rohan Anthony Marley Julian Ricardo Marley Ky-Mani Marley Damian Robert Nesta Marley

Extract from infobox+ontologize+publish in RDF

About: Bob Marley

An Entity in Data Space: ec2-174-129-126-64.compute-1.amazonaws.com

Robert "Bob" Nesta Marley OM (February 6, 1945 – May 11, 1981) was a Jamaican singer-songwriter and musician.

Property	Value
dbpedia-owl:associatedBand	▪ dbpedia:The_Upsetters ▪ dbpedia:The_Wailers_%28reggae_band%29 ▪ dbpedia:I_Threes ▪ dbpedia:Wailers_Band
dbpedia-owl:associatedMusicalArtist	▪ dbpedia:The_Upsetters ▪ dbpedia:The_Wailers_%28reggae_band%29 ▪ dbpedia:I_Threes ▪ dbpedia:Wailers_Band
dbpedia-owl:background	▪ solo_singer
dbpedia-owl:birthdate	▪ 1945-02-06 (xsd:date)
dbpedia-owl:birthname	▪ Robert Nesta Marley
dbpedia-owl:birthplace	▪ dbpedia:Saint_Ann_Parish%2C_Jamaica ▪ dbpedia:Jamaica
dbpedia-owl:deathdate	▪ 1981-05-11 (xsd:date)
dbpedia-owl:deathplace	▪ dbpedia:Miami%2C_Florida ▪ dbpedia:Florida

Fig. 5.2 An example DBpedia dashboard fragment describing Bob Marley

DBpedia generally occupies the central position in LOD diagrams because of two inter-related characteristics: (1) it is open-world and contains many entities, concepts and properties of interest, since it is ultimately described from Wikipedia infoboxes, and (2) for various reasons, many publishers on LOD have chosen to link to DBpedia entities to fulfill the fourth Linked Data principle. Although DBpedia is not dynamically fresh in the same vein as Wikipedia, which is constantly maintained by a community-enforced system of edits, revisions and additions, DBpedia is *periodically* derived from Wikipedia by executing extractors on Wikipedia dumps. Thus it is *relatively* fresh compared to more static datasets on LOD.

Overall, DBpedia continues to be well-maintained and widely used. In part, this is because of its dependence on Wikipedia, which has continued to be popular, but also because numerous applications across the Semantic Web, knowledge discovery and NLP communities now leverage it for weak supervision and distant supervision-style problems. Just like Wikipedia, DBpedia is also multi-lingual, which opens up even more applications.

5.3 Google Knowledge Vault

The Google Knowledge Vault, which indirectly populates some of the search features in Google, is a Web-scale probabilistic knowledge base that combines *extractions* from Web content (obtained via analytics over text, tabular data, page structure, and even human annotations) with *prior knowledge* derived from existing knowledge repositories [55]. Because these are distinct information sources, supervised machine learning methods have to be used for knowledge *fusion*. At the time of publication, this Knowledge Vault (KV) was assumed to be substantially bigger than any previously published structured knowledge repository, and featured a probabilistic inference system that could compute calibrated probabilities of assertion correctness. The authors of the Knowledge Vault paper report results from several studies and experiments illustrating the utility of the method [55].

Fundamentally, the KV was no different at an architectural level (see Fig. 5.3 for the architectural description of the KV) than the workflow proposed in this book. That is, there were three main components. The first component was a layer of extractors. Recall that, in an earlier chapter, we provided extensive details on information extraction, which is among the first steps in constructing a domain-specific knowledge graph from raw data. The KV is no different, although it is not single-domain. Extraction methods include text (including relation extraction, although the authors run standard methods at much larger scale), and Web IE methods like parsing the DOM trees of HTML pages, and also tables extracted from HTML. The KV also contains data from pages annotated manually with elements from ontologies like schema.org and openGraphProtocol.org [67, 69]. Schema.org is described in more detail in the following section.

However, rather than learn all its knowledge about the world from just IE, the KV also relied on prior knowledge by using graph-based priors. In essence,

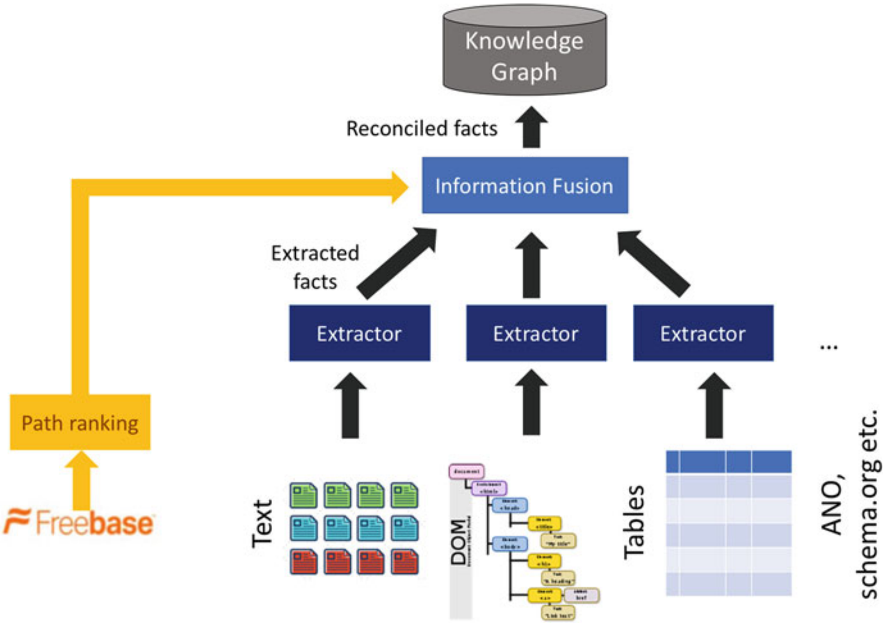


Fig. 5.3 A schematic showing the construction and workflow of the Google Knowledge Vault

these systems would learn the prior probability of each possible triple, based on triples stored in an existing KB. Technically, the procedure was different from the KGEs that we covered in the previous chapter; however, the concept was similar. For example, one of the applications of knowledge graph embeddings was *triples classification*, namely, the task of determining the probability of correctness of a hypothetical triple, given all the triples observed during training. Incorporating graph-based priors into the KV relied on a similar intuition.

Finally, the third key innovation in the KV architecture was an information or knowledge fusion box that would take the outputs of extractors, as well as those based on graph-based priors, and reconcile the facts based on supervised machine learning. Knowledge fusion was like the test phase in a triples classification system. In the actual paper, the authors consider several principled supervised machine learning methods.

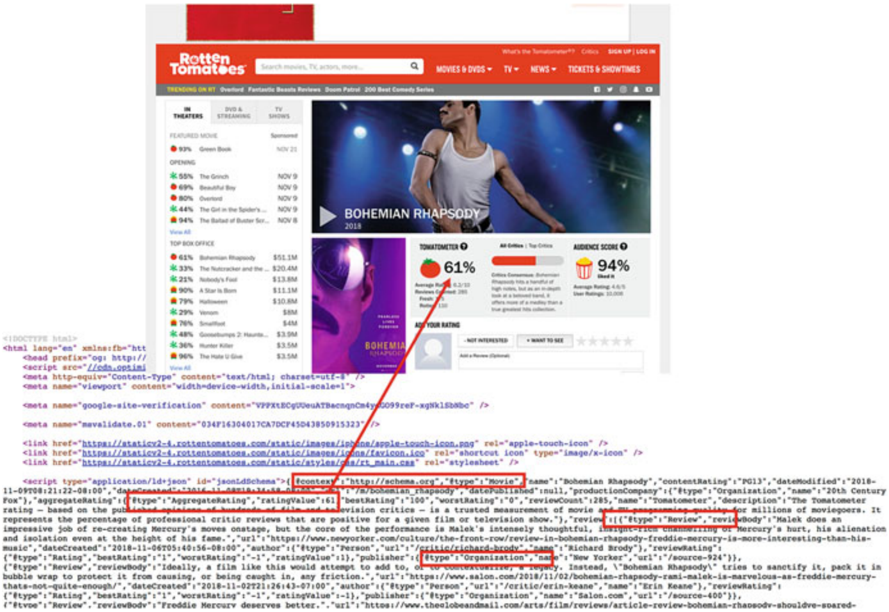
Although it is not known whether the KV constitutes the core technology powering the current iteration of the Google Knowledge Graph [164], its influence on the construction of Web-scale knowledge graphs from heterogeneous structured and unstructured data sources is undeniable. The effort has proven difficult to replicate in non-industrial settings, however. Some of the components, like information fusion, have also been superseded by recent innovations such as knowledge graph embeddings. However, the role of extractors and the leverage of prior knowledge in reconciling contradictions continue to be important in existing KG construction pipelines.

5.4 Schema.org

Schema.org is a shared vocabulary that webmasters can use to structure metadata on their websites and to help search engines understand the content being published [69]. Although the term *schema.org* would seem to suggest a website (which it is, leading to the project's homepage) it is contextually used to refer to the vocabulary itself, the markup on the webpages as well as the overall project, which is described as 'a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond'.

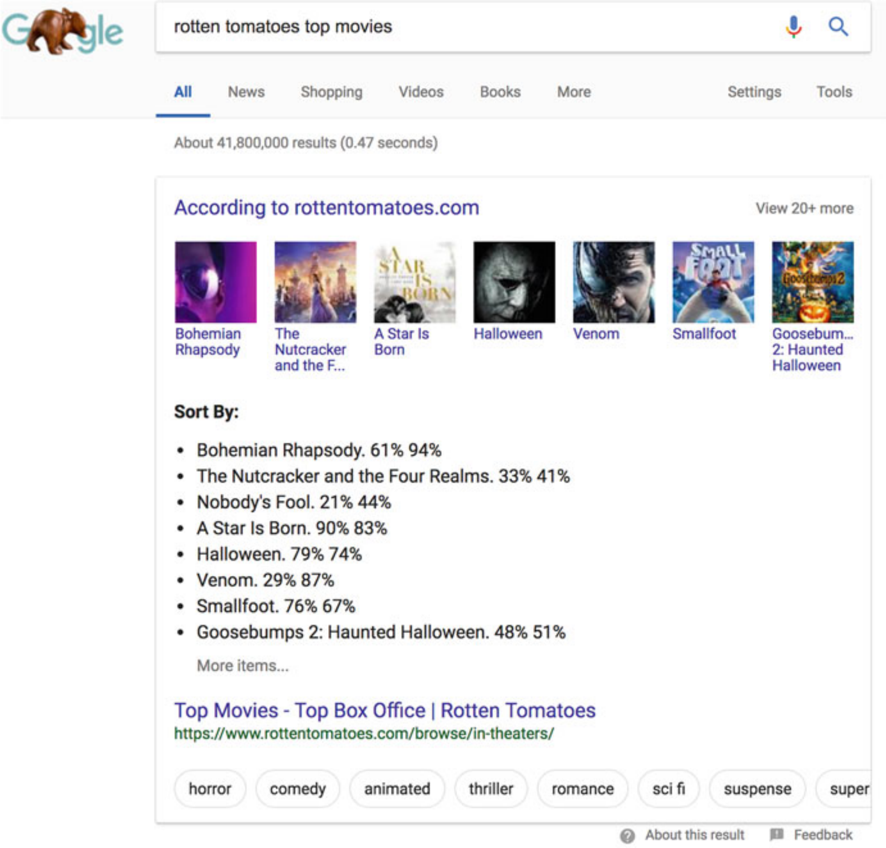
As an example, consider the movie 'Bohemian Rhapsody' as described by Rotten Tomatoes, a popular movie and review aggregation website. Many of the dynamic elements on the page (such as the reviews and the Tomatometer rating) have semantics associated with them according to the concepts and properties in the schema.org vocabulary. We highlight some example snippets in Fig. 5.4, with the aggregate rating being one example of an element that is visually rendered on the screen. When a search engine like Google scrapes this data, it is able to make use of this information to provide users with a better search experience (e.g., providing better answers to queries like 'rotten tomatoes top movies', see Fig. 5.5).

In fact, the initiative itself was launched on 2 June 2011 by Bing, Google and Yahoo! to create, support and develop a common set of schemas for structured



```
<!DOCTYPE html>
<html lang="en" xmlns:fb="htt
    <head prefix="og: http://
    <script src="//cdn.optimi
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
    <meta name="viewport" content="width=device-width,initial-scale=1">

    <meta name="google-site-verification" content="VPPXtECgUUeuATBacnqnCm4yGO99reF-xgNklSbNbc" />

    <meta name="msvalidate.01" content="034F16304017CA7DCF45D43850915323" />

    <link href="https://staticv2-4.rottentomatoes.com/static/images/iphone/apple-touch-icon.png" rel="apple-touch-icon" />
    <link href="https://staticv2-4.rottentomatoes.com/static/images/icons/favicon.ico" rel="shortcut icon" type="image/x-icon" />
    <link href="https://staticv2-4.rottentomatoes.com/static/styles/css/rt_main.css" rel="stylesheet" />

    <script type="application/ld+json" id="jsonLdSchema">{"@context":"http://schema.org","@type":"Movie","name":"Bohemian Rhapsody","contentRating":"PG13","dateModified":"2018-
11-09T08:21:22-08:00","dateCreated":"2016-11-08T19:34:58-08:00","id":"/m/bohemian_rhapsody","datePublished":null,"productionCompany":{"@type":"Organization","name":"20th Century
Fox"},"aggregateRating":{"@type":"AggregateRating","ratingValue":61,"bestRating":"100","worstRating":"0","reviewCount":285,"name":"Tomatometer","description":"The Tomatometer
rating – based on the published opinions of hundreds of film and television critics – is a trusted measurement of movie and TV programming quality for millions of moviegoers. It
represents the percentage of professional critic reviews that are positive for a given film or television show."},"review":[{"@type":"Review","reviewBody":"Malek does an
impressive job of re-creating Mercury's moves onstage, but the core of the performance is Malek's intensely thoughtful, insight-rich channelling of Mercury's hurt, his alienation
and isolation even at the height of his fame.","url":"https://www.newyorker.com/culture/the-front-row/review-in-bohemian-rhapsody-freddie-mercury-is-more-interesting-than-his-
music","dateCreated":"2018-11-06T05:40:56-08:00","author":{"@type":"Person","url":"/critic/richard-brody","name":"Richard Brody"},"reviewRating":
{"@type":"Rating","bestRating":"1","worstRating":"-1","ratingValue":1},"publisher":{"@type":"Organization","name":"New Yorker","url":"/source-924"}},
{"@type":"Review","reviewBody":"Ideally, a film like this would attempt to add to, or to contextualize, a legacy. Instead, \"Bohemian Rhapsody\" tries to sanctify it, pack it in
bubble wrap to protect it from causing, or being caught in, any friction.","url":"https://www.salon.com/2018/11/02/bohemian-rhapsody-rami-malek-is-marvelous-as-freddie-mercury-
thats-not-quite-enough/","dateCreated":"2018-11-02T21:26:43-07:00","author":{"@type":"Person","url":"/critic/erin-keane","name":"Erin Keane"},"reviewRating":
{"@type":"Rating","bestRating":"1","worstRating":"-1","ratingValue":-1},"publisher":{"@type":"Organization","name":"Salon.com","url":"/source-400"}},
{"@type":"Review","reviewBody":"Freddie Mercury deserves better.","url":"https://www.theglobeandmail.com/arts/film/reviews/article-review-bohemian-rhapsody-shouldve-soared-
```

Fig. 5.4 Example of schema.org snippets embedded in Rotten Tomato's HTML webpages



About 41,800,000 results (0.47 seconds)

According to rottentomatoes.com

Sort By:

- Bohemian Rhapsody. 61% 94%
- The Nutcracker and the Four Realms. 33% 41%
- Nobody's Fool. 21% 44%
- A Star Is Born. 90% 83%
- Halloween. 79% 74%
- Venom. 29% 87%
- Smallfoot. 76% 67%
- Goosebumps 2: Haunted Halloween. 48% 51%

More items...

Top Movies - Top Box Office | Rotten Tomatoes
https://www.rottentomatoes.com/browse/in-theaters/

Fig. 5.5 Example of dynamic, enhanced search by commercial search providers using extracted schema.org information

markup on webpages. This would indicate that enhanced search has been one of the primary goals for the initiative. Later that year, Yandex joined the initiative. The main motivation behind using schema.org is that the markup can be recognized by search engine spiders and other parsers, which enables a layer of semantics to be incorporated into search engine optimization. Recall that the Semantic Web (which includes Linked Data [26]) vision was similar [18], but the principles of Linked Data do not necessarily apply to schema.org. As one example, much of the published schema.org markup does not even attempt to link to other schema.org markup. Consequently, the schema.org ecosystem is less like KGs in the LOD universe [11], and more like isolated knowledge fragments that provide local context to the webpages in which they are embedded. However, there have been efforts to try and integrate LOD and schema.org with some recent papers proposing to find and add schema.org fragments to the LOD. The work is still in its early stages, however. At the very least, viable solutions to Entity Resolution will be required to facilitate such a vision successfully, since without ER, the fourth principle of Linked Data cannot be fulfilled.

Much of the vocabulary on Schema.org was inspired by earlier formats, such as microformats, FOAF, and OpenCyc. Microformats, with its most dominant representative hCard, continue (as of 2015) to be published widely on the web, where the deployment of Schema.org has strongly increased between 2012 and 2014. In 2015, Google began supporting the JSON-LD format, and as of September, 2017 recommended using JSON-LD for structured data whenever possible. Tools are also widely available to validate schema.org markup on published webpages. For example, tools such as the Google Structured Data Testing Tool, Yandex Microformat validator, and Bing Markup Validator can be used to test the validity of published or scraped schema.org data. In documentation released by Google,[5] it could be reasonably inferred that certain schema.org classes and properties, particularly people and organizations, influence the results of Google's Knowledge Graph.

5.5 Where is the Future Going?

Even the brief examples illustrated in this chapter show that KG ecosystems are continuing to flourish and come into their own, powering a full range of applications across communities as diverse as NLP, semantic search and conversational AI [65]. These ecosystems are different enough that, at first sight, one might be tempted to think that they are evolving independently of each other. Yet there are connections, some of which are only starting to materialize. For example, some authors have started publishing work on how to reconcile decentralized, highly disconnected schema.org knowledge fragments with the larger Linked Data ecosystem. Freebase, which was used to power the KV and has since been taken over by Google, was an essential part of the Linking Open Data project in its initial phase, and has since been replaced with Wikidata [174]. Nevertheless, despite all these connections, the question remains: is there some way to reconcile all of these different KGs under a single umbrella, one that is open and accessible to all?

An example of one such initiative, still in a seedling stage, is the Open Knowledge Network[6] (OKN), which is attempting to jumpstart and realize the long-held vision of a *common semantic information infrastructure* for the future [116]. Recognizing the motivation that natural interfaces to large knowledge structures have the potential to impact science, education and business to an extent comparable to the WWW, the OKN initiative argues that KG-centric services like Alexa and

[5] https://developers.google.com/search/docs/guides/enhance-site#add-your-sites-name-logo-and-social-links

[6] https://okfn.org/network/

Cortana, or the Google search engine, are limited in their scope of knowledge, not open to direct access or contributors beyond their corporate firewalls, and can only answer relatively limited questions in their business areas. OKN wants to pioneer an architecture that will allow stakeholders to encode knowledge for their topics of interest and be able to hook them into the larger network, without having to go through gatekeepers (such as Google or Apple). Furthermore, once the knowledge is encoded, access to this should not be restricted to a small priesthood of SQL or other programmatic interface users. There will be a wide range of interfaces, including natural language interfaces, graphical interfaces and visualizations which no one has even invented yet. Developers will be able to independently create more sophisticated programs for answering queries, providing summaries that help regular people make decisions in their lives.

In order to realize the vision of an open Web-scale knowledge network, an attempt like the Google Knowledge Vault is required but at a scale that (arguably) is even more extensive. As ambitious as this may sound, the steering members of OKN argue that the technologies for realizing such a network already exist. However, it is also undeniable that there are many hurdles in realizing such an ambition, including obvious issues of cost, incentives and maintenance. From a purely research standpoint however, the OKN would be far more comprehensive than any existing KG ecosystem, and would likely trigger revolutionary advances in KG-centric applications.

While the OKN is likely a longer-term initiative that will require the coalescing of multiple research communities, there are several medium-term research challenges that researchers have already started focusing on. Entity Resolution continues to be a vital area of research, especially considering our arguments in the earlier chapter on how existing systems continue to fall short on several requirements that are essential for conducting ER on large KGs. Information extraction (IE) also continues to advance each year, though some kinds of IE are witnessing more attention than others. NER research, for example, seems to have plateaued, but relation and event extraction systems continue to be presented, even at the time of writing. Low-supervision IE has also seen a surge of interest. Finally, IE for languages other than English, and particularly for ‘low-resource’ languages for which good translation services are not available, has seen an increased surge in research interest due to programs funded by agencies like DARPA.