Chapter 2
Meet the Data

This chapter introduces the basic types of data sources, as well as specific datasets and resources, that we will be working with in later chapters of the book. These may be placed on a spectrum of varying degrees of structure, from unstructured to structured data, as shown in Fig. 2.1.

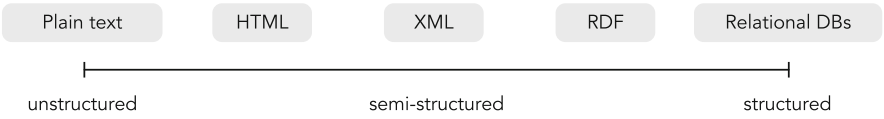

Fig. 2.1 The data spectrum

On the *unstructured* end of the spectrum we have plain text. Typically, these are documents written in natural language.[1] As a matter of fact, almost any type of data can be converted into plain text, including web pages, emails, spreadsheets, and database records. Of course, such a conversion would result in an undesired loss of internal document structure and semantics. It is nevertheless always an option to treat data as unstructured, by not making any assumptions about the particular data format. Search in unstructured text is often referred to as *full-text search*.

On the opposite end of the spectrum there is *structured* data, which is typically stored in relational databases; it is highly organized, tabular, and governed by a strict schema. Search in this type of data is performed using formal query languages, like SQL. These languages allow for a very precise formulation of information needs, but require expert knowledge of the query language and of the underlying database schema. This generally renders them unsuitable for ordinary users.

The data we will mostly be dealing with is neither of two extremes and falls somewhere "in the middle." Therefore, it is termed *semi-structured*. It is

[1] Written in natural language does not imply that the text has to be grammatical (or even sensible).

K. Balog, *Entity-Oriented Search*, The Information Retrieval Series 39,
https://doi.org/10.1007/978-3-319-93935-3_2

Table 2.1 Comparison of unstructured, semi-structured, and structured data search

	Unstructured	Semi-structured	Structured
Unit of retrieval	Documents	Objects	Tuples
Schema	No	Self-describing	Fixed
Queries	Keyword	Keyword++	Formal languages

characterized by the lack of a fixed, rigid schema. Also, there is no clear separation between the data and the schema; instead, it uses a self-describing structure (tags or other markers). Semi-structured data can be most easily viewed as a combination of unstructured and structured elements. Let us point out that text is rarely completely without structure. Even simple documents typically have a title (or a filename, that is often meaningful). In HTML documents, markup tags specify elements such as headings, paragraphs, and tables. Emails have sender, recipient, subject, and body fields. What is important to notice here is that these document elements or fields may or may not be present. This differs from structured data, where every field specified by the schema ahead of time must be given some permitted value. Therefore, documents with optional, self-describing elements naturally belong to the category of semi-structured data. Furthermore, relational database records may also be viewed as semi-structured data, by converting them to a set of hierarchically nested elements. Performing such conversions can in fact simplify data processing for entity-oriented applications. Using a semi-structured entity representation, all data related to a given entity is available in a single entry for that entity. Therefore, no aggregation via foreign-key relationships is needed. Table 2.1 summarizes data search over the unstructured-structured spectrum.

The remainder of this chapter is organized according to the main types of data sources we will be working with: the Web (Sect. 2.1), Wikipedia (Sect. 2.2), and knowledge bases (Sect. 2.3).

2.1 The Web

The World Wide Web (WWW), commonly known simply as "the Web," is probably the most widely used information resource and service today. The idea of the Web (what today would be considered Web 1.0) was introduced by Tim Berners-Lee in 1989. Beginning in 2002, a new version, dubbed "Web 2.0" started to gain traction, facilitating a more active participation by users, such that they changed from mere consumers to become also creators and publishers of content. The early years of the Web 2.0 era were landmarked by the launch of some of today's biggest social media sites, including Facebook (2004), YouTube (2005), and Twitter (2006). Finally, the Semantic Web (or Web 3.0) was proposed as an extension of the current Web [3, 20]. It represents the next major evolution of the Web that enables data to be understood by computers, which could then perform tasks intelligently on behalf of

Table 2.2 Publicly available web crawls

Name	Time period	Size	#Documents
ClueWeb09 full[a]	Jan 2009–Feb 2009	5 TB	1B
ClueWeb09 (Category B)		230 GB	50M
ClueWeb12[b]	Feb 2012–May 2012	5.6 TB	733M
ClueWeb12 (Category B)		400 GB	52M
Common Crawl[c]	May 2017	58 TB	2.96B
KBA stream corpus 2014[d]	Oct 2011–Apr 2013	10.9 TB	1.2B

Size refers to compressed data
[a] https://lemurproject.org/clueweb09/
[b] https://lemurproject.org/clueweb12/
[c] http://commoncrawl.org/2017/06/may-2017-crawl-archive-now-available/
[d] http://trec-kba.org/kba-stream-corpus-2014.shtml

users. The term *Semantic Web* refers both to this as-of-yet-unrealized future vision and to a collection of standards and technologies for knowledge representation (cf. Sect. 1.2.4).

Web pages are more than just plain text; one of their distinctive characteristics is their hypertext structure, defined by the HTML markup. HTML tags describe the internal document structure, such as headings, paragraphs, lists, tables, and so on. Additionally, HTML documents contain hyperlinks (or simply "links") to other pages (or resources) on the Web. Links are utilized in at least two major ways. First, the networked nature of the Web may be leveraged to identify important or authoritative pages or sites. Second, many of the links also have a textual label, referred to as *anchor text*. Anchor text is "incredibly useful for search engines because it provides some extra description of the page being pointed to" [23].

2.1.1 Datasets and Resources

We introduce a number of publicly available web crawls that have been used in the context of entity-oriented search. Table 2.2 presents a summary.

ClueWeb09/12 The ClueWeb09 dataset consists of about one billion web pages in 10 languages,[2] collected in January and February 2009. The crawl aims to be a representative sample of what is out there on the Web (which includes SPAM and pornography). ClueWeb09 was used by several tracks of the TREC conference. The data is distributed in gzipped files that are in WARC format. About half of the collection is in English; this is referred to as the "Category A" subset. Further, the first segment of Category A, comprising about 50 million pages, is referred to

[2] English, Chinese, Spanish, Japanese, French, German, Portuguese, Arabic, Italian, and Korean.

as the "Category B" subset.[3] The Category B subset also includes the full English Wikipedia. These two subsets may be obtained separately if one does not need the full collection.

ClueWeb12 is successor to the ClueWeb09 web dataset, collected between February and May 2012. The crawl was initially seeded with URLs from ClueWeb09 (with the highest PageRank values, and then removing likely SPAM pages) and with some of the most popular sites in English-speaking countries (as reported by Alexa[4]). Additionally, domains of tweeted URLs were also injected into the crawl on a regular basis. A blacklist was used to avoid sites that promote pornography, malware, and the like. The full dataset contains about 733 million pages. Similarly to ClueWeb09, a "Category B" subset of about 50 million English pages is also made available.

Common Crawl Common Crawl[5] is a nonprofit organization that regularly crawls the Web and makes the data publicly available. The datasets are hosted on Amazon S3 as part of the Amazon Public Datasets program.[6] As of May 2017, the crawl contains 2.96 billion web pages and over 250 TB of uncompressed content (in WARC format). The Web Data Commons project[7] extracts structured data from the Common Crawl and makes those publicly available (e.g., the Hyperlink Graph Dataset and the Web Table Corpus).

KBA Stream Corpus The KBA Stream Corpus 2014 is a focused crawl, which concentrates on news and social media (blogs and tweets). The 2014 version contains 1.2 billion documents over a period of 19 months (and subsumes the 2012 and 2013 KBA Stream Corpora). See Sect. 6.2.5.1 for a more detailed description.

2.2 Wikipedia

Wikipedia is one of the most popular web sites in the world and a trusted source of information for many people. Wikipedia defines itself as "a multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on a model of openly editable content."[8] Content is created through the collaborative effort of a community of users, facilitated by a *wiki* platform. There are various mechanisms in place to maintain high-quality content, including the verifiability policy (i.e., readers should be able to check that the information comes

[3] The Category B subset was mainly intended for research groups that were not yet ready at that time to scale up to one billion documents, but it is still widely used.

[4] http://www.alexa.com/.

[5] http://commoncrawl.org/.

[6] https://aws.amazon.com/public-datasets/.

[7] http://webdatacommons.org/.

[8] https://en.wikipedia.org/wiki/Wikipedia:About.

from a reliable source) and a clear set of editorial guidelines. The collaborative editing model makes it possible to distribute the effort required to create and maintain up-to-date content across a multitude of users. At the time of writing, Wikipedia is available in nearly 300 languages, although English is by far the most popular, with over five million articles. As stated by Mesgari et al. [15], "Wikipedia may be the best-developed attempt thus far to gather all human knowledge in one place."

What makes Wikipedia highly relevant for entity-oriented search is that most of its entries can be considered as (semi-structured) representations of entities. At its core, Wikipedia is a collection of pages (or articles, i.e., encyclopedic entries) that are well interconnected by hyperlinks. On top of that, Wikipedia offers several (complementary) ways to group articles, including categories, lists, and navigation templates. In the remainder of this section, we first look at the anatomy of a regular Wikipedia article and then review (some of the) other, special-purpose page types. We note that it is not our aim to provide a comprehensive treatment of all the types of pages in Wikipedia. For instance, in addition to the encyclopedic content, there are also pages devoted to the administration of Wikipedia (discussion and user pages, policy pages and guidelines, etc.); although hugely important, these are outside our present scope of interest.

2.2.1 The Anatomy of a Wikipedia Article

A typical Wikipedia article focuses on a particular entity (e.g., a well-known person, as shown in Fig. 2.2) or concept (e.g., "democracy").[9] Such articles typically contain, among others, the following elements (the letters in parentheses refer to Fig. 2.2):

- Title (I.)
- Lead section (II.)
  - Disambiguation links (II.a)
  - Infobox (II.b)
  - Introductory text (II.c)
- Table of contents (III.)
- Body content (IV.)
- Appendices and bottom matter (V.)
  - References and notes (V.a)
  - External links (V.b)
  - Categories (V.c)

[9] We refer back to Sect. 1.1.1 for a discussion on the difference between entities and concepts.

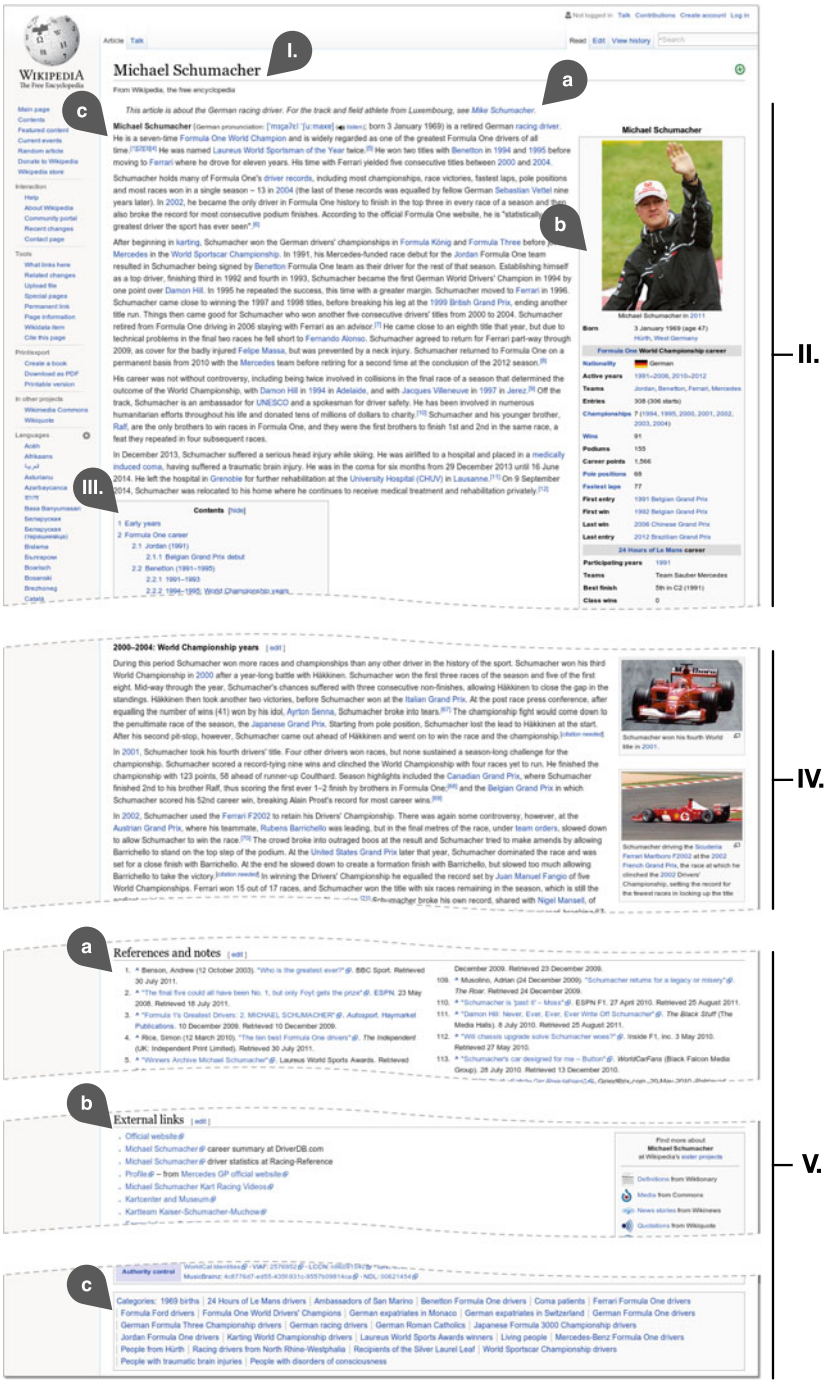

Fig. 2.2 The Wikipedia page of MICHAEL SCHUMACHER

The lead section of a Wikipedia article is the part between the title heading and the table of contents. It serves as an introduction to the article and provides a summary of its contents. The lead section may contain several (optional) elements, including disambiguation links, maintenance tags, infobox, image, navigational boxes, and introductory text. We will further elaborate on the title, infobox, and introductory text elements below.

The main body of the article may be divided into sections, each with a section heading. The sections may be nested in a hierarchy. When there are at least four sections, a navigable table of contents gets automatically generated and displayed between the lead section and the first heading.

The body of the article may be followed by optional appendix and footer sections, including internal links to related Wikipedia articles ("see also"), references and notes (that cite sources), further reading (links to relevant publications that have not been used as sources), internal links organized into navigational boxes, and categories.

2.2.1.1 Title

Each Wikipedia article is uniquely identified by its *page title*. The title of the page is typically the most common name for the entity (or concept) described in the article. When the name is ambiguous, the pages of the other namesakes are disambiguated by adding further qualifiers to their title within parentheses. For instance, MICHAEL JORDAN refers to the American (former) professional basketball player, and the page about the English footballer with the same name has the title MICHAEL JORDAN (FOOTBALLER). Note that the page title is case-sensitive (except the first character). For special pages, the page title may be prefixed with a namespace, separated with a colon, e.g., "Category:German racing drivers." We will look at some of the Wikipedia namespaces later in this section.

2.2.1.2 Infobox

The infobox is a panel that summarizes information related to the subject of the article. In desktop view, it appears at the top right of the page, next to the lead section; in mobile view it is displayed at the very top of the page. In the case of entity pages, the infobox summarizes key facts about the entity in the form of property-value pairs. Therefore, infoboxes represent an important source for extracting structured information about entities (cf. Sect. 2.3.2). A large number of infobox templates exist, which are created and maintained collaboratively, with the aim to standardize information across articles that belong to the same category. Infoboxes, however, are "free form," meaning that what ultimately gets included in the infobox of a given article is determined through discussion and consensus among the editors.

```
Schumacher holds many of Formula One's [[List of Formula One driver records|
driver records]], including most championships, race victories, fastest laps,
pole positions and most races won in a single season - 13 in [[2004 Formula
One season|2004]] (the last of these records was equalled by fellow German
[[Sebastian Vettel]] 9 years later). In [[2002 Formula One season|2002]],
he became the only driver in Formula One history to finish in the top three
in every race of a season and then also broke the record for most consecutive
podium finishes. According to the official Formula One website, he is
"statistically the greatest driver the sport has ever seen".
```

Listing 2.1 Wikitext markup showing internal links to other Wikipedia pages

2.2.1.3 Introductory Text

Most Wikipedia articles include an introductory text, the "lead," which is a brief summary of the article—normally, no more than four paragraphs long. This should be written in a way that it creates interest in the article. The first sentence and the opening paragraph bear special importance. The first sentence "can be thought of as the definition of the entity described in the article" [11]. The first paragraph offers a more elaborate definition, but still without being too detailed. DBpedia, e.g., treats the first paragraph as the "short abstract" and the full introductory text as the "long abstract" of the entity (cf. Sect. 2.3.2).

2.2.2 Links

Internal links are an important feature of Wikipedia as they allow "readers to deepen their understanding of a topic by conveniently accessing other articles."[10] Listing 2.1 shows the original wiki markup for the second paragraph of the introductory text from Schumacher's Wikipedia page from Fig. 2.2. Links are created by enclosing the title of a target page in double square brackets (`[[...]]`). Optionally, an alternative label, i.e., *anchor text*, may be provided after the vertical bar (`|`). Linking is governed by a detailed set of guidelines. A key rule given to editors is to link only the first occurrence of an entity or concept in the text of the article.

The value of links extends beyond navigational purposes; they capture semantic relationships between articles. In addition, anchor texts are a rich source of entity name variants. Wikipedia links may be used, among others, to help identify and disambiguate entity mentions in text (cf. Chap. 5).

[10] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking.

2.2.3 Special-Purpose Pages

Not all Wikipedia articles are entity pages. In this subsection and the next, we discuss two specific kinds of special-purpose pages.

2.2.3.1 Redirect Pages

Each entity in Wikipedia has a dedicated article and is uniquely identified by the page title of that article. The page title is the most common (canonical) name of the entity. Entities, however, may be referred to by multiple names (aliases). The purpose of *redirect pages* is to allow entities to be referred to by their name variants. Additionally, redirect pages are also often created for common misspellings of entity names. Redirect pages have no content themselves, they merely act as pointers from alternative surface forms to the canonical entity page. Whenever the user visits a redirect page, she will automatically be taken to the "main" article representing the entity. For example, pages redirecting to UNITED STATES include acronyms (U.S.A., U.S., USA, US), foreign translations (ESTADOS UNIDOS), misspellings (UNTIED STATES), and synonyms (YANKEE LAND). Mind that Wikipedia page titles are unique, thus each redirect may refer to a single entity, i.e., the most popular entity with that name (e.g., OBAMA redirects to BARACK OBAMA).

2.2.3.2 Disambiguation Pages

Disambiguation pages serve a reverse role: They are created for ambiguous names and list all entities that share that name. That is, they enumerate all possible meanings of a name. Disambiguation pages are always suffixed by "(disambiguation)" in the title. For example, the BENJAMIN FRANKLIN (DISAMBIGUATION) page lists eight different people, four ships, seven art and literature productions, along with a number of other possible entities, including the $100 bill and various buildings.

2.2.4 Categories, Lists, and Navigation Templates

Wikipedia offers three complementary ways to group related articles together: categories, lists, and navigation templates. These are independent of each other, and each method follows a particular set of guidelines and standards. It is not uncommon that a topic is simultaneously covered by a category, a list, and a navigation template. For example, the "Formula One constructors" article is grouped in all three ways: as a category, a list, and a template.

2.2.4.1 Categories

Categories mainly serve navigational purposes: they provide navigational links for the reader to browse sets of related pages; see V.c on Fig. 2.2. Each Wikipedia article should be assigned to at least one category; most articles are members of several categories. Each article designates to what categories it belongs, and the category page is automatically populated based on what articles declare membership to that category.

Category pages can be distinguished from regular articles by the "Category:" prefix in the page title. That is, a category is a page itself in the *Category* namespace (all other content we have discussed so far is in the *main* namespace). Each category page contains an introductory text (that can be edited like an article), and two automatically generated lists: one with subcategories and another with articles that belong to the category. Different kinds of categories may be distinguished, with the first two being of primary importance:

- *Topic categories* are named after a topic (usually corresponding to a Wikipedia article with the same name on that topic), e.g., "Formula One."
- *Set categories* are named after a particular class (usually in plural), e.g., "German racing drivers."
- *Set-and-topic categories* are a combination of the above two types, e.g., "Formula One drivers of Brawn."
- *Container categories* only contain other categories.
- *Universal categories* provide a comprehensive list of articles that are otherwise divided into subcategories, e.g., "1969 births."
- *Administration categories* are mainly used by editors for management purposes, e.g., "Clean up categories."

Categories are also organized in a hierarchy; each category should be a subcategory of some other category (except a single root category, called "Contents"). This categorization, however, is not a well-defined "is-a" hierarchy, but a (directed) graph; a category may have multiple parent categories and there might be cycles along the path to ancestors. There are various alternative ways to turn this graph into a tree, depending on where we start, i.e., what are selected as top-level categories. Below are two possible starting points:

- Fundamental categories[11] define four fundamental ontological categories: physical entities ("physical universe"), biological entities ("life"), social entities ("society"), and intellectual entities ("concepts").
- Wikipedia's portal for Categories[12] provides another starting point with a set of 27 main categories covering most of the knowledge domains.

[11] https://en.wikipedia.org/wiki/Category:Fundamental_categories.

[12] https://en.wikipedia.org/wiki/Portal:Contents/Categories.

According to Wikipedia's guidelines, the general rule to categorization, apart from certain exceptions, is that an article (1) should be categorized as low down in the category hierarchy as possible and (2) should usually not be in both a category and its subcategory.

2.2.4.2 Lists

Lists, as contrasted with categories, provide a means for manual categorization of articles. Lists have a number of advantages over categories. They can be maintained from a centralized location (at the list page itself), and there is more control over the presentation of the content (order of items, formatting, etc.). Importantly, they can also include "missing" articles, that is, items that do not have a Wikipedia page yet. Unfortunately, lists are more difficult to process automatically. Also, for some topics (e.g., people from a particular country), lists would be infeasible to maintain, due to the large number of entries.

2.2.4.3 Navigation Templates

Navigation templates are manual compilations of links that may be included in multiple articles and edited in a central place, i.e., the template page. They provide a navigation system with consistent look and organization for related articles. Navigation templates are meant to be compact and should offer a useful grouping of the linked articles (e.g., by topic, by era, etc.); for that, they may use custom formatting, beyond standard lists or tables. Template inclusion is bidirectional: every article that includes a given navigation template should also be contained as a link in that template. Like categories and lists, templates can also be utilized, e.g., for the task of completing a set of entities with other semantically related entities.

2.2.5 Resources

Wikipedia is based on the MediaWiki software,[13] which is a free open source wiki package. MediaWiki uses an extensible lightweight wiki markup language. Wikipedia may be downloaded in various formats, including XML and SQL dumps or static HTML files.[14] Page view statistics are also made publicly available for download.[15] Wikipedia's content is also accessible via the MediaWiki API in

[13] https://www.mediawiki.org.

[14] https://dumps.wikimedia.org/.

[15] https://dumps.wikimedia.org/other/analytics/.

various formats (including JSON and XML).[16] There exists a broad selection of tools for browsing, editing, analyzing, and visualizing Wikipedia.[17] In addition to the official MediaWiki parser, a number of alternative and special-purpose parsers have been created.[18]

2.3 Knowledge Bases

It is important to realize that Wikipedia has been created for human consumption. For machines, the content in this form is hardly accessible. Specifically, it is not a machine-readable structured knowledge model. In our terminology, Wikipedia is a knowledge repository.

The first usage of the term *knowledge base* is connected to expert systems, dating back to the 1970s. Expert systems, one of the earliest forms of (successful) AI software, are designed to solve (or aid humans in solving) complex problems, such as medical diagnosis, by reasoning about knowledge [6]. These systems have rather different data needs than what is supported by relational databases ("tables with strings and numbers"[19]). For such systems, knowledge needs to be represented explicitly. A knowledge base is comprised of a large set of assertions about the world. To reflect how humans organize information, these assertions describe (specific) entities and their relationships. An AI system can then solve complex tasks, such as participating in a natural language conversation, by exploiting the KB.[20]

One of the earliest attempts at such an AI system was the Cyc project that started in 1984 with the objective of building a comprehensive KB to represent everyday commonsense knowledge. The development has been ongoing for over 30 years now and is still far from finished. The main limitation of Cyc is that it relies on human knowledge engineers. In a system with ever-growing complexity, it becomes increasingly difficult to add new objects. While the Cyc project is still alive, it appears that manually codifying knowledge using formal logic and an extensive ontology is not the way forward with respect to the ultimate goal of natural language understanding.

Knowledge bases are bound to be incomplete; there is always additional information to be added or updated. Modern information access approaches embrace this inherent incompleteness. A KB is often regarded as a "semantic backbone" and used in combination with unstructured resources. Instead of relying on large

[16] https://www.mediawiki.org/wiki/API:Main_page.

[17] https://en.wikipedia.org/wiki/Wikipedia:Tools.

[18] https://www.mediawiki.org/wiki/Alternative_parsers.

[19] We admit that this is a gross oversimplification.

[20] To make this possible, the AI system also requires an *ontology* (a finite set of rules governing object relationships) and a (logical) inference engine.

ontologies, generally a rather lightweight approach is taken by using some form of subsumption ("is-a") ontology. When the emphasis is on the relationships between entities, a knowledge base is often referred to as a *knowledge graph*.

There exist general purpose as well as domain-specific knowledge bases. DBpedia and YAGO are academic projects that each derive a KB automatically from Wikipedia. Freebase is a community-curated KB that was the open core of Google's Knowledge Graph. It was, however, closed down in 2015 and the data is currently being transferred to Wikidata. All major search providers have their own proprietary knowledge base. Examples include Google's Knowledge Graph,[21] Microsoft's Satori,[22] and Facebook's Entity Graph.[23] Unfortunately, there is very little information available about these beyond popular science introductions.

Before discussing a number of specific knowledge bases, we first explain some fundamentals.

2.3.1 A Knowledge Base Primer

Knowledge bases will be instrumental to most tasks and approaches that will be discussed in this book. Thus, in this section, we explain the core underlying concepts, as well as the RDF data model for representing knowledge in a structured format.

A knowledge base may be divided into two layers:

- On the *schema level* lies a knowledge model, which defines semantic classes (i.e., entity types), the relationships between classes and instances, properties that (instances of) classes can have, and possibly additional constraints and restrictions on them (e.g., range of allowed values for a certain property). Classes are typically organized in a subsumption hierarchy (i.e., a *type taxonomy*).
- The *instance level* comprises a set of assertions about specific entities, describing their names, types, attributes, and relationships with each other.

The predominant language on the Web for describing instances is RDF, which we shall introduce in greater detail in Sect. 2.3.1.2. The KB schema may be encoded using a declarative language, such as RDFS (for expressing taxonomical relationships)[24] or OWL (for full-fledged ontological modeling).[25]

[21] https://googleblog.blogspot.no/2012/05/introducing-knowledge-graph-things-not.html.

[22] https://blogs.bing.com/search/2014/03/31/150-million-more-reasons-to-love-bing-everyday/.

[23] http://www.technologyreview.com/news/511591/facebook-nudges-users-to-catalog-the-real-world/.

[24] https://www.w3.org/TR/rdf-schema/.

[25] https://www.w3.org/OWL/.

2.3.1.1 Knowledge Bases vs. Ontologies

In information retrieval and natural language processing, knowledge bases have become central to machine understanding of natural language over the past decade. More recently, and especially in the search industry, often the term *knowledge graph* is used. In the fields of artificial intelligence and the Semantic Web, people have been using *ontologies* for similar, or even more ambitious, goals since the 1990s. The question naturally arises: What is the difference between a knowledge base and an ontology? Is it only a matter of choice of terminology or is there more to it? The simple answer is that a knowledge base can be represented as an ontology. For example, the YAGO knowledge base refers to itself as an ontology [21]. But the two are not exactly the same.

To understand the difference, we should first clarify what an ontology is. The word carries several connotations, depending on the particular discipline and research field. Perhaps the most widely cited definition is by Gruber [9], which states that "an ontology is an explicit specification of a conceptualization." According to Navigli [16], "an ontology is a set of definitions in a formal language for concepts that describe the world of interest, including the relationships that connect these concepts." Put simply, an ontology is a means to formalizing knowledge. Building blocks of an ontology include (1) *classes* (or *concepts*), (2) *objects* (or *instances*), (3) *relations*, connecting classes and objects to one another, (4) *attributes* (or *properties*), representing relations intrinsic to specific objects, (5) *restrictions* on relations, and (6) *rules* and *axioms*, which are assertions in a logical form [16]. The conceptual design of an ontology revolves around the possible concepts and relations that are to be encoded. The instance level (i.e., individual objects) may not even be involved in the process or included in the ontology. Knowledge bases, on the other hand, place the main emphasis on individual objects and their properties. The formal constraints imposed by the ontology on those objects are only of interest when the knowledge base is being populated with new facts.

In summary, both knowledge bases and ontologies attempt to capture a useful representation of a (physical or virtual) world, with the overall objective to solve (complex) problems. Ontologies are schema-oriented ("top-down" design) and focus on describing the concepts and relationships within a given domain with the highest possible expressiveness. Conversely, knowledge bases are fact-oriented ("bottom-up" design), with an emphasis on describing specific entities.

2.3.1.2 RDF

The Resource Description Framework (RDF) is a language designed to describe "things," which are referred to as *resources*. A resource denotes either an entity (object), an entity type (class), or a relation. Each resource is assigned a Uniform Resource Identifier (URI), making it uniquely and globally identifiable. Each RDF statement is a triple, consisting of *subject*, *predicate*, and *object* components.

- The *subject* is always a URI, denoting a resource.
- The *predicate* is also a URI, corresponding to a relationship or property of the subject resource.
- The *object* is either a URI (referring to another resource) or an (optionally typed) literal.

Consider the following piece of information:

> *Michael Schumacher (born 3 January 1969) is a retired German racing driver, who raced in Formula One for Ferrari.*

It may be represented as the following set of RDF statements (often referred to as *SPO triples*, for short). Each triple represents an atomic factual statement in the knowledge base. URIs are enclosed in angle brackets and are shortened for improved readability (see Table 2.4 for the full URIs of the namespaces `dbr`, `foaf`, etc.); literal values are in quotes.

<dbr:Michael_Schumacher>	<foaf:name>	"Schumacher, Michael"
<dbr:Michael_Schumacher>	<dbo:birthPlace>	<dbr:West_Germany>
<dbr:Michael_Schumacher>	<dbo:birthDate>	"1969-01-03"
<dbr:Michael_Schumacher>	<rdf:type>	<dbo:RacingDriver>
<dbr:Michael_Schumacher>	<dct:subject>	<dbc:Ferrari_Formula_One_drivers>

Note that the expressivity of the RDF representation depends very much on the vocabulary of predicates. This particular example is taken from DBpedia (which we shall introduce in detail in the next section), where the object `<dbc:Ferrari_Formula_One_drivers>` identifies a certain category of entities and the predicate `<dct:subject>` assigns the subject entity to that category. Using the DBpedia Ontology, there is (currently) no way to express this relationship more precisely, emphasizing that the person *was driving* for that team but not anymore.

Mind that a relationship between two entities is a directed link (labeled with a predicate). For instance, the fact that SCHUMACHER won the 1996 SPANISH GRAND PRIX may be described as either of the following two statements:

<dbr:1996_Spanish_Grand_Prix>	<dbp:firstDriver>	<dbr:Michael_Schumacher>
<dbr:Michael_Schumacher>	<dbp:firstDriverOf>	<dbr:1996_Spanish_Grand_Prix>

In reality, only the first one is an actual triple; the second is a made up example, as there is no `<dbp:firstDriverOf>` predicate in DBpedia. Even if there was one, this second triple would only introduce redundant information. What is important here is that information pertinent to a given entity is contained in triples with that entity standing as either subject or object. When the entity stands as object, the triple may be reversed by taking the inverse of the predicate, like:

<dbr:Michael_Schumacher>	**is <dbp:firstDriver> of**	<dbr:1996_Spanish_Grand_Prix>

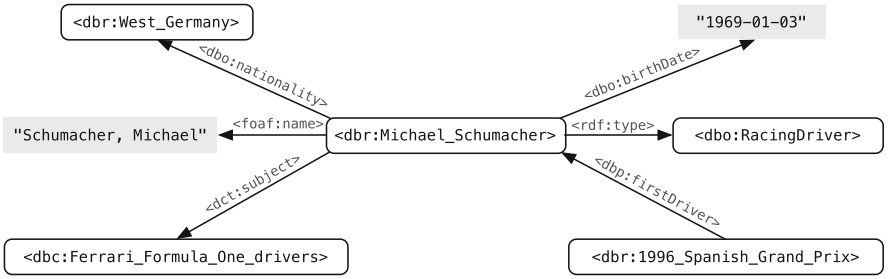

Fig. 2.3 Excerpt of an RDF graph (taken from DBpedia). URIs (i.e., entities) are represented by rounded rectangles, literals (i.e., attribute values) are denoted by shaded rectangles

Conceptually, the set of RDF triples forms a large, directed, labelled graph (referred to as the *RDF graph*). Each RDF triple corresponds to a pair of nodes in the graph (subject and object), connected by an edge (predicate). Figure 2.3 displays the graph corresponding to the triples from our running example.

Note that RDF describes the instance level in the knowledge base. To cope with the knowledge base schema (concepts and relations), an ontology representation language is needed, such as RDFS or OWL. Essentially, what RDFS and OWL provide are vocabularies for ontological modeling. RDFS (RDF Schema) provides a vocabulary for encoding taxonomies and is often preferred when lightweight modeling is sufficient. OWL (Web Ontology Language) builds upon RDFS and comes with the full expressive power of description logics. For the serialization, storage, and retrieval of RDF data we refer to Sect. 2.3.7.

2.3.2 DBpedia

In layman's terms, DBpedia[26] is a "database version of Wikipedia." More precisely, DBpedia is a knowledge base that is derived by extracting structured data from Wikipedia [12]. One powerful aspect of DBpedia is that it is not a result of a one-off process but rather of a continuous community effort, with numerous releases since its inception in 2007. Over the years, DBpedia has developed into an interlinking hub in the Web of Data (which will be discussed in Sect. 2.3.6). Another distinguishing feature of DBpedia is that it is available in multiple languages. We base our discussion below on the latest release that is available at the time of writing, DBpedia 2016-10, and especially on the English version. For an overview of DBpedia's evolution over time and for details on the localized versions, we refer to [12]. Due to DBpedia's importance as a resource, we shall provide an in-depth treatment of its main components.

[26] http://dbpedia.org/.

2.3.2.1 Ontology

The DBpedia Ontology is a cross-domain ontology that has been manually created based on the most frequently used Wikipedia infoboxes (cf. II.b on Fig. 2.2). The current version contains 685 classes, which are organized in a six-level deep subsumption hierarchy. The ontology is intentionally kept this shallow so that it can be easily visualized and navigated. Each class within the ontology is described by a number of properties; for each property, the range of possible values is also defined. Figure 2.4 shows a small excerpt from the DBpedia Ontology.

The maintenance of the ontology is a community effort that is facilitated by the DBpedia Mappings Wiki. Using this wiki, users can collaboratively create and edit mappings from different infobox templates to classes and properties in the DBpedia Ontology. These mappings are instrumental in extracting high-quality data as they alleviate problems arising from the heterogeneity of Wikipedia's infoboxes. We elaborate more on this in the following subsection. Crowdsourcing turned out to be a powerful tool for extending and refining the ontology. The number of properties has grown from 720 in 2009, to 1650 in 2014, and to 2795 in 2016. The number of classes has increased at a similar pace, from 170 in 2009, to 320 in 2014, and to 685 in 2016. The current version of DBpedia describes 4.58 million entities, out of which 4.22 million are classified in the ontology. The largest ontology classes

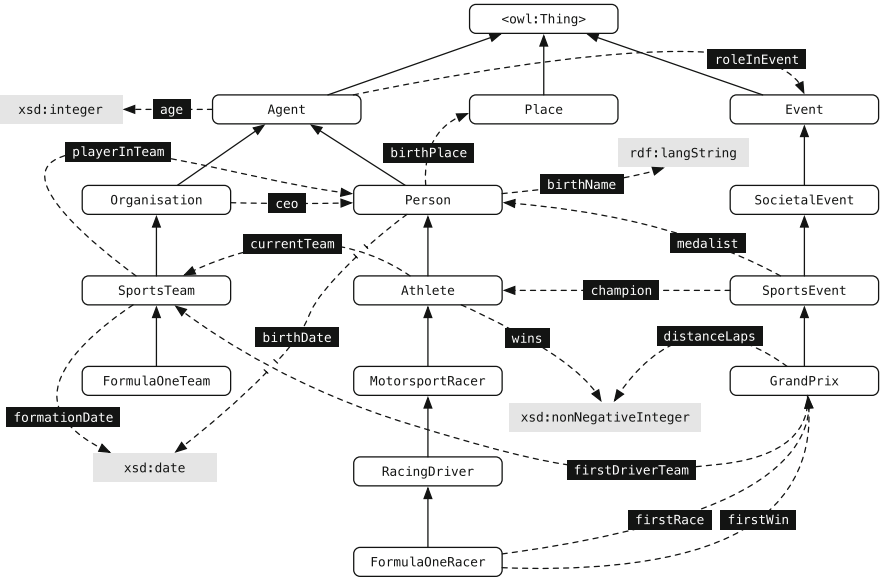

Fig. 2.4 Excerpt from the DBpedia Ontology. Classes are represented by rounded rectangles where arrows with solid lines indicate subclass relationships (from subclass to superclass). Properties are denoted by the dashed arrows with black labels. Value types are shown in gray boxes. Unless specifically indicated, classes/values are in the `dbo` namespace (cf. Table 2.4)

include persons (1.4M), places (735k), creative works like music albums and films (411k), and organizations (241k).

2.3.2.2 Extraction

The DBpedia extraction framework follows a pipeline architecture, where the input is a Wikipedia article and the output is a set of RDF statements extracted from that article. The framework encompasses a number of different purpose-built extractors; some of these are designed to grab a single property (e.g., the abstract or a link to an image depicting the entity), while others deal with specific parts of Wikipedia pages (e.g., infoboxes). The DBpedia extractors can be categorized into four main types:

Raw infobox extraction The most important source of structured information are the infoboxes. These list the main facts about a given entity as property-value pairs. The raw infobox extractor directly translates all Wikipedia infobox properties to DBpedia predicates. There is no normalization performed either on properties (i.e., RDF predicates) or values (i.e., RDF objects). This generic extraction method provides complete coverage of all infobox data. The predicates with the `dbp` prefix in the example triples in Sect. 2.3.1.2 are the results of this raw infobox extraction.

Mapping-based infobox extraction One major issue with infoboxes is inconsistency. A wide range of infobox templates are used in Wikipedia, which evolve over time. As a result, the same type of entity may be described by different templates; these templates may use different names for the same property (e.g., `birthplace` vs. `placeofbirth`). Further, attribute values may be expressed using a range of alternative formats and units of measurement. In order to enforce consistency, which is an important dimension of data quality, a homogenization of properties and values is necessary. This normalization (or homogenization) is done against the DBpedia Ontology and is made possible by the community-provided mappings specified in the DBpedia Mappings Wiki. Not only are predicates normalized but literal object values are also canonicalized to basic units according to the assigned datatypes. The mapping-based approach significantly increases the quality compared to the raw infobox data. The RDF statements generated by this extractor can be distinguished by the `dbo` prefix of the predicates.

Feature extraction A number of specialized extractors are developed for the extraction of a single feature from an article. These include, among others, abstract, categories, disambiguations, external links, geo-coordinates, homepage, image, label, page links, and redirects; we refer to Table 2.3 for the descriptions of these.

Statistical extraction The extractors in this last group are not part of the DBpedia "core." They were created with the intent to provide resources that can support computational linguistics tasks [14]. Unlike the core extractors, which are essentially rule-based, these employ statistical estimation techniques. Some of

Table 2.3 A selection of specific feature extractors in DBpedia

Name	Predicate	Description
Abstract	`dbo:abstract`	The first lines of the Wikipedia article
Categories	`dc:subject`	Wikipedia categories assigned to the article
Disambiguation	`dbo:wikiPageDisambiguates`	Disambiguation links
External links	`dbo:wikiPageExternalLink`	Links to external web pages
Geo-coordinates	`georss:point`	Geographical coordinates
Homepage	`foaf:homepage`	Link to the official homepage of an instance
Image	`foaf:depiction`	Link to the first image on the Wikipedia page
Label	`rdfs:label`	The page title of the Wikipedia article
Page links	`dbo:wikiPageWikiLink`	Links to other Wikipedia articles
Redirect	`dbo:wikiPageRedirects`	Wikipedia page to redirect to

See Table 2.4 for the URI prefixes

them deviate further from the regular extractors in that they aggregate data from all Wikipedia pages as opposed to operating on a single article. The resulting datasets include grammatical gender (for entities of type person), lexicalizations (alternative names for entities and concepts), topic signatures (strongest related terms), and thematic concepts (the main subject entities/concepts for Wikipedia categories).

2.3.2.3 Datasets and Resources

The output of each DBpedia extractor, for each language, is made available as a separate dataset. All datasets are provided in two serializations: as Turtle (N-triples) and as Turtle quads (N-Quads, which include context). The datasets can be divided into the following categories:

- *DBpedia Ontology*: The latest version of the ontology that was used while extracting all datasets.
- *Core datasets*: All infobox-based and specific feature extractors (including the ones listed in Table 2.3) belong here.
- *Links to other datasets*: DBpedia is interlinked with a large number of knowledge bases. The datasets in this group provide links to external resources both on the instance level (`owl:sameAs`), e.g., to Freebase and YAGO, and on the schema level (`owl:equivalentClass` and `owl:equivalentProperty`), most notably to schema.org.
- *NLP datasets*: This last group corresponds to the output of the statistical extractors.

Namespaces and Internationalization The generic DBpedia URI namespaces are listed in the upper block of Table 2.4. As part of the internationalization efforts, some datasets are available both in *localized* and in *canonicalized* version.

Table 2.4 Main URI namespaces used in DBpedia

Prefix	URL	Description
DBpedia namespaces		
`dbr`	`http://dbpedia.org/resource/`	One-to-one mapping between Wikipedia articles and DBpedia resources
`dbp`	`http://dbpedia.org/property/`	Properties from raw infobox extraction
`dbo`	`http://dbpedia.org/ontology/`	DBpedia Ontology
External namespaces[a]		
`dc`	`http://purl.org/dc/elements/1.1/`	Dublin core
`foaf`	`http://xmlns.com/foaf/0.1/`	Friend of a friend (FOAF)
`georss`	`http://www.georss.org/georss/`	Geographically encoded objects for RSS
`owl`	`http://www.w3.org/2002/07/owl#`	W3C web ontology language
`rdf`	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`	Standard W3C RDF vocabulary
`rdfs`	`http://www.w3.org/2000/01/rdf-schema#`	Extension of the basic RDF vocabulary

[a]The list is not intended to be exhaustive

The localized datasets include everything from the given language's Wikipedia and use language-specific URIs (`http://<lang>.dbpedia.org/resource/` and `http://<lang>.dbpedia.org/property/`). The canonicalized datasets, on the other hand, only contain resources that exist in the English edition of Wikipedia as well; here, the generic (language-agnostic) URI namespace is used.

SPARQL Endpoint The various datasets are not only available for download but can also be accessed and queried via a public SPARQL endpoint.[27] The endpoint is hosted using the Virtuoso Universal Server.

DBpedia Live Content on Wikipedia is changing. For example, in April 2016, there were 3.4M edits on the English Wikipedia.[28] New DBpedia versions are released fairly regularly, at least once per year. However, for certain applications or usage scenarios, this rate of updating might be too slow. Instead of the infrequent batch updates, a live synchronization mechanism would be preferable. DBpedia Live[29] is a module that does exactly this: It processes Wikipedia updates real-time and keeps DBpedia up-to-date. DBpedia Live offers a separate SPARQL endpoint.[30] Additionally, the "changesets" (added and removed triples) are made available and can be applied on top of a local copy of DBpedia with the help of a sync tool.[31]

[27] http://dbpedia.org/sparql.

[28] https://stats.wikimedia.org/EN/SummaryEN.htm.

[29] http://live.dbpedia.org/.

[30] http://live.dbpedia.org/sparql.

[31] https://github.com/dbpedia/dbpedia-live-mirror.

2.3.3 YAGO

YAGO[32] (which stands for Yet Another Great Ontology) is a similar effort to DBpedia in that it extracts structured information from Wikipedia, such that each Wikipedia article becomes an entity. Although they share similar aims, the underlying systems and philosophy are quite different. While DBpedia stays close to Wikipedia and aims to simply provide an RDF version of it, YAGO focuses on achieving high precision and consistent knowledge. Instead of relying on mappings collaboratively created by a community, YAGO's extraction is facilitated by expert-designed declarative rules. Each fact in YAGO is annotated with a confidence value. According to an empirical evaluation, the accuracy of the contained facts is about 95% [21].

2.3.3.1 Taxonomy

Another key difference between DBpedia and YAGO lies in the typing of entities. While DBpedia employs a small, manually curated ontology, YAGO constructs a deep subsumption hierarchy of entity types by connecting Wikipedia categories with WordNet concepts. WordNet[33] is a large lexical resource that groups words into sets of cognitive synonyms (synsets). Each synset expresses a distinct concept; ambiguous words (i.e., words with multiple meanings) belong to multiple synsets.

Wikipedia categories are organized in a directed graph, but this is not a strict hierarchy (cf. Sect. 2.2.4.1). Moreover, the relations between categories merely reflect the thematic structure. "Thus, the hierarchy is of little use from an ontological point of view" [21]. Hence, YAGO establishes a hierarchy of classes, where the upper levels are based on WordNet synsets and the leaves come from (a subset of the) Wikipedia leaf categories. This results in over 568K entity types, hierarchically organized in 19 levels.

2.3.3.2 Extensions

There have been two major extensions to the original YAGO knowledge base. YAGO2 [10] anchors knowledge in time and space, i.e., places entities and facts on their spatial and temporal dimension. Specifically, timestamps are defined for four main entity types: people, groups, artifacts, events. It is argued that "these four types cover almost all of the cases where entities have a meaningful existence time" [10]. Location is extracted for entities that have a "permanent spatial extent on Earth" [10], such as countries, cities, mountains, and rivers. A new super-

[32] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

[33] http://wordnet.princeton.edu/.

class (`yagoGeoEntity`) is introduced to the YAGO taxonomy, which groups together all geo-entities. Geo-entities are harvested from two sources: Wikipedia (based on associated geographical coordinates) and GeoNames[34] (a freely available geographical database). In summary, YAGO2 associates over 30 million facts with their occurrence time, and over 17 million facts with the location of occurrence. The time of existence is known for 47% and the location is known for 30% of all entities. According to a sample-based manual assessment, YAGO2 has a precision of 95%.

YAGO3 [13] extends YAGO to multiple languages, by running the YAGO extraction system on different language editions of Wikipedia. This brings in one million new entities and seven million facts over the original (English-only) YAGO.

2.3.3.3 Resources

YAGO is publicly available for download in TSV or Turtle formats, either in its entirety or just specific portions.[35] The YAGO type hierarchy is also offered in DBpedia as an alternative to the DBpedia Ontology. Mappings (i.e., "same-as" links) between YAGO and DBpedia instances are provided in both directions.

2.3.4 Freebase

Freebase[36] is an open and large collaborative knowledge base [5]. It was launched in 2007 by the software company Metaweb, which was acquired by Google in 2010. Freebase "was used as the open core of the Google Knowledge Graph" [19]. In December 2014, Google announced that it would shut down Freebase and help with the transfer of content from Freebase to Wikidata.[37] The migration process is briefly elaborated on in the next subsection.

The content of Freebase has been partially imported from open data sources, such as Wikipedia, MusicBrainz,[38] and the Notable Names Database (NNDB).[39] Another part of the data comes from user-submitted wiki contributions. Freebase encouraged users to create entries for less popular entities (which would have not made it to Wikipedia). Instead of using controlled ontologies, Freebase adopted a folksonomy approach, in which users could use types much like tags. Each type has a number of properties (i.e., predicates) associated with it.

[34] http://www.geonames.org.

[35] http://yago-knowledge.org.

[36] https://developers.google.com/freebase/.

[37] https://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc.

[38] https://musicbrainz.org/.

[39] http://www.nndb.com/.

Google has made some important data releases using Freebase, specifically, entity annotations for the ClueWeb09 and ClueWeb12 corpora (cf. Sect. 5.9.2) and for the KBA Stream Corpus 2014 (cf. Sect. 6.2.5.1). The latest Freebase dump, from 31 March 2015, is still available for download. It contains 1.9 billion triples and about 39 million entities (referred to as *topics* in Freebase).

2.3.5 Wikidata

Wikidata[40] is a free collaborative knowledge base operated by the Wikimedia Foundation [22]. Its goal is to provide the same information as Wikipedia, but in a structured format. Launched in October 2012, Wikidata "has quickly become one of the most active Wikimedia projects" [22]. As of 2017, it has over 7K active monthly contributors (those making at least 5 edits per month). Unlike the previous KBs we have discussed, Wikidata does not consider statements as facts, but rather as *claims*, each having a list of references to sources supporting that claim. Claims can contradict each other and coexist, thereby allowing opposing views to be expressed (e.g., different political positions). Essentially, claims are property-value pairs for a given *item*, which is Wikidata lingo for an entity. There is support for two special values: "unknown" (e.g., a person's exact day of death) and "no value" (e.g., Australia has no bordering countries); these cases are different from data being incomplete. Claims can also have additional subordinate property-value pairs, called *qualifiers*. Qualifiers can store contextual information (e.g., the validity time for an assertion, such as the population of a city in a certain year, according to a particular source). Importantly, Wikidata is multilingual by design and uses language-independent entity IDs.

Wikidata relies on crowdsourced manual curation to ensure data quality. With the retirement of Freebase, Google decided to offer the content of Freebase to Wikidata. This migration, which is still underway at the time of writing, is not without challenges. One of the main difficulties is rooted in the "cultural" differences between the two involved communities; they "have a very different background, subtly different goals and understandings of their tasks, and different requirements regarding their data" [19]. One specific challenge is that Wikidata is eager to have references for their statements, which are not present in Freebase. Such references are obtained from the Google Knowledge Vault [8], then checked and curated manually by Wikidata contributors using a purpose-built tool, called the *Primary Sources Tool*; we refer to [19] for details. As of June 2017, Wikidata contains over 158 million statements about 26.9 million entities. Wikidata offers copies of its

[40] https://wikidata.org/.

Fig. 2.5 Result snippet from a Google search result page

```
<section class="ar_recipe_index full-page" itemscope
         itemtype="http://schema.org/Recipe">
    <link href="http://allrecipes.com/recipe/132929/easy-chicken-satay/"
          itemprop="url" />
    <meta itemprop="mainEntityOfPage" content="True" />
```

Listing 2.2 Excerpt from a recipe's HTML page annotated with Microdata meta tags. Source: http://allrecipes.com/recipe/132929/easy-chicken-satay/

content for download in JSON, RDF, and XML formats,[41] and also provides access via a search API.[42]

2.3.6 The Web of Data

The amount of structured data available on the Web is growing steadily. A large part of it is contained in various knowledge bases, like DBpedia and Freebase. In addition, increasingly more data is being exposed in the form of semantic annotations added to traditional web pages using metadata standards, such as Microdata, RDFa, and JSON-LD. There is a strong incentive for websites for marking up their content with semantic metadata: It allows search engines to enhance the presentation of the site's content in search results. An important standardization development was the introduction of `schema.org`, a common vocabulary used by major search providers (including Google, Microsoft, and Yandex) for describing commonly used entity types (including people, organizations, events, products, books, movies, recipes, etc.). Figure 2.5 displays a snippet from a Google search result page; Listing 2.2 shows an excerpt from the HTML source of the corresponding HTML page with Microdata annotations.

Historically, data made available in RDF format was referred to as *Semantic Web* data. One of the founding principles behind the Semantic Web is that data should be interlinked. These principles were summarized by Berners-Lee [2] in the following four simple rules:

[41] http://www.wikidata.org/wiki/Wikidata:Database_download.

[42] https://query.wikidata.org/.

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information.
4. Include links to other URIs, so that people can discover more things.

The term *Linked Data* (LD) refers to a set of best practices for publishing structured data on the Web. The key point about Linked Data is that it enables to connect entities (or, generally speaking, resources) across multiple knowledge bases. This is facilitated by a special "same-as" predicate, `<owl:sameAs>`, basically saying that the subject and object resources connected by that predicate are the same. For example, the following two statements connect the representations of the entity MICHAEL SCHUMACHER across DBpedia, Freebase, and Wikidata:

`<dbr:Michael_Schumacher>`	**`<owl:sameAs>`**	`<fb:m.053w4>`
`<dbr:Michael_Schumacher>`	**`<owl:sameAs>`**	`<wikidata:Q9671>`

These "same-as" links connect all Linked Data into a single global data graph. Mind that the term Linked Data should be used to describe the publishing practice, not the data itself. A knowledge base published using LD principles should be called *Linked Dataset*. To avoid the terminological confusion, we shall refer to the collection of structured data exposed on the Web in machine understandable format as the *Web of Data* (emphasizing the difference in nature to the traditional *Web of Documents*). *Linked Open Data* (LOD) may also be used as a casual synonym, emphasizing the fact that Linked Data is released under an open license. Figure 2.6 shows the Linked Open Data cloud, where edges indicate the presence of "same-as" links between two datasets (knowledge bases). Notice that DBpedia is a central hub here.

2.3.6.1 Datasets and Resources

In this book, we focus on two particular data collections, BTC-2009 and Sindice-2011, that have been used in the information retrieval community for entity-oriented research. For a more extensive list of datasets, tools, and Linked Data resources, see http://linkeddata.org/.

BTC-2009 The Billion Triples Challenge 2009 dataset (BTC-2009)[43] was created for the Semantic Web Challenge in 2009 [4]. The dataset was constructed by combining the crawls of multiple semantic search engines during February–March 2009. It comprises 1.1 billion RDF triples, describing 866 million distinct resources, and amounts to 17 GB in size (compressed).

[43] https://km.aifb.kit.edu/projects/btc-2009/.

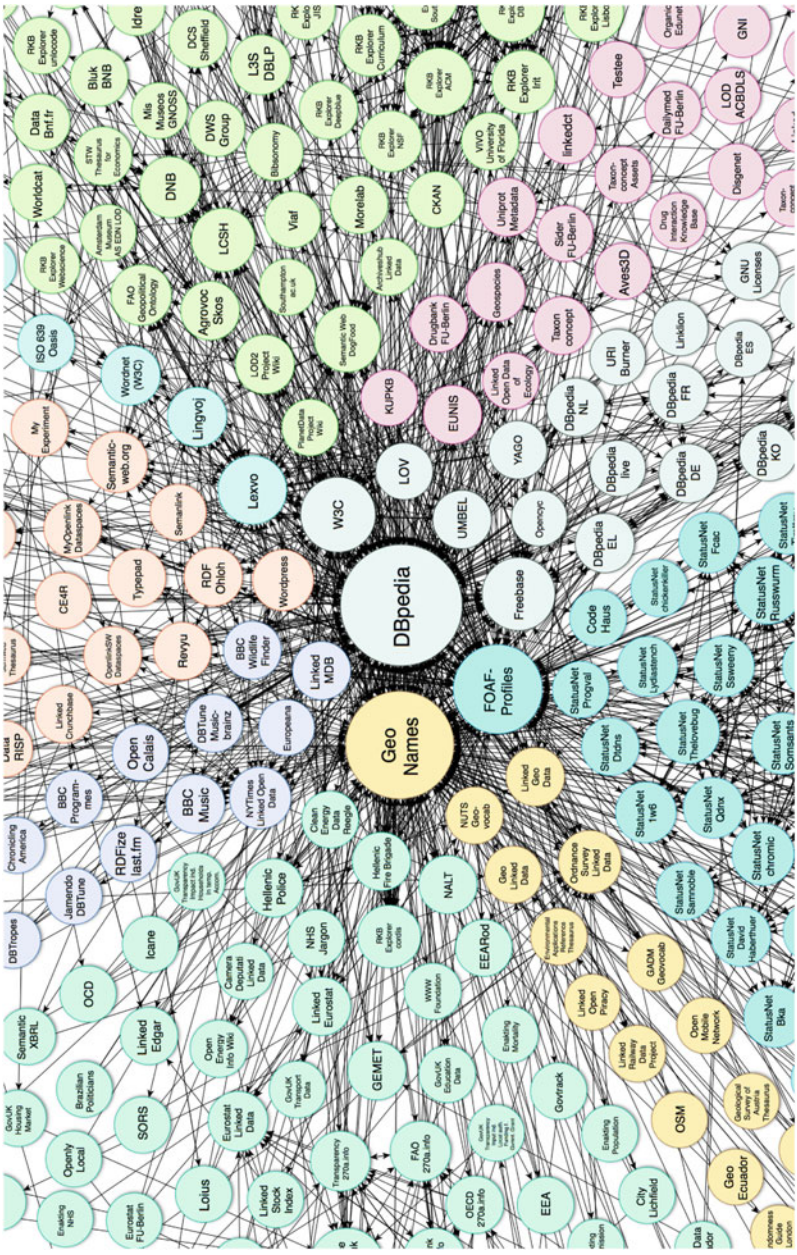

Fig. 2.6 Excerpt from the Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch, and Richard Cyganiak. Source: http://lod-cloud.net/

Sindice-2011 The Sindice-2011 dataset [7] was created in 2011 for the TREC Entity track [1] with the aim to provide a more accurate reflection of the at-the-time current Web of Data. The data has been collected by the Sindice semantic search engine [18] between 2009 and 2011. Sindice-2011 contains 11 billion RDF statements, describing 1.7 billion entities. The dataset is 1.3 TB in size (uncompressed).

2.3.7 Standards and Resources

RDF, RDFS, and OWL are all standards of the World Wide Web Consortium (W3C),[44] which is the main international standards organization for the World Wide Web. There exist numerous serializations for RDF data, e.g., Notation-3, Turtle, N-Triples, RDFa, and RDF/JSON. The choice of serialization depends on the context and usage scenario. For example, Turtle is probably the easiest serialization to use for human consumption and manipulation. If large volumes of data need to be interchanged between systems, then producing data dumps in N-Triples format is a common choice. If only HTML documents are produced, then RDFa is preferred. SPARQL[45] is a structured query language for retrieving and manipulating RDF data, and is also a W3C standard. Triplestores are special-purpose databases designed for storing and querying RDF data. Examples of triplestores include Apache Jena,[46] Virtuoso,[47] and RDF-3X [17].

2.4 Summary

This chapter has introduced the different kinds of data, from unstructured to structured, that we will be using in the coming chapters. The order in which we have discussed them—first the Web, then Wikipedia, and finally knowledge bases—reflects how the research focus in entity-oriented search is shifting toward relying increasingly more on structured data sources, and specifically on knowledge bases. Knowledge bases are rich in structure, but light on text; they are of high quality, but are also inherently incomplete. This stands in stark contrast with the Web, which is a virtually infinite source of heterogeneous, noisy, and text-heavy content that comes with limited structure. Their complementary nature makes the combination of knowledge bases and the Web a particularly attractive and fertile ground for entity-oriented (re)search.

[44] https://www.w3.org/.
[45] https://www.w3.org/TR/sparql11-query/.
[46] https://jena.apache.org/.
[47] https://virtuoso.openlinksw.com/.

References

1. Balog, K., Serdyukov, P., de Vries, A.P.: Overview of the TREC 2011 Entity track. In: The Twentieth Text REtrieval Conference Proceedings, TREC '11. NIST (2012)
2. Berners-Lee, T.: Linked data (2009)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5), 34–43 (2001)
4. Bizer, C., Mika, P.: Editorial: The semantic web challenge, 2009. Web Semantics: Science, Services and Agents on the World Wide Web **8**(4) (2010)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, pp. 1247–1250. ACM (2008). doi: 10.1145/1376616.1376746
6. Buchanan, B.G., Shortliffe, E.H.: Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley Series in Artificial Intelligence). Addison-Wesley Publishing Co. (1984)
7. Campinas, S., Ceccarelli, D., Perry, T.E., Delbru, R., Balog, K., Tummarello, G.: The Sindice-2011 dataset for entity-oriented search in the web of data. In: 1st International Workshop on Entity-Oriented Search, EOS '11 (2011)
8. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 601–610. ACM (2014). doi: 10.1145/2623330.2623623
9. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993). doi: 10.1006/knac.1993.1008
10. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194**, 28–61 (2013). doi: 10.1016/j.artint.2012.06.001
11. Kazama, J., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '07, pp. 698–707. Association for Computational Linguistics (2007)
12. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal (2012)
13. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A knowledge base from multilingual Wikipedias. In: Seventh Biennial Conference on Innovative Data Systems Research, CIDR '15 (2015)
14. Mendes, P.N., Jakob, M., Bizer, C.: DBpedia for NLP: A multilingual cross-domain knowledge base. In: Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC '12. ELRA (2012)
15. Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.Å., Lanamäki, A.: "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. Journal of the Association for Information Science and Technology **66**(2), 219–245 (2015). doi: 10.1002/asi.23172
16. Navigli, R.: Ontologies. In: Mitkov, R. (ed.) Ontologies. Oxford University Press (2017)
17. Neumann, T., Weikum, G.: RDF-3X: a risc-style engine for RDF. Proc. VLDB Endow. **1**(1), 647–659 (2008). doi: 10.14778/1453856.1453927
18. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: a document-oriented lookup index for open linked data. Int. J. Metadata Semant. Ontologies **3**(1), 37–52 (2008). doi: 10.1504/IJMSO.2008.021204

19. Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L.: From Freebase to Wikidata: The great migration. In: Proceedings of the 25th International Conference on World Wide Web, WWW '16, pp. 1419–1428. International World Wide Web Conferences Steering Committee (2016). doi: 10.1145/2872427.2874809
20. Shadbolt, N., Berners-Lee, T., Hall, W.: The semantic web revisited. IEEE Intelligent Systems **21**(3), 96–101 (2006). doi: 10.1109/MIS.2006.62
21. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 697–706. ACM (2007). doi: 10.1145/1242572.1242667
22. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledge base. Commun. ACM **57**(10), 78–85 (2014). doi: 10.1145/2629489
23. Zhai, C., Massung, S.: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ACM and Morgan & Claypool (2016)