Data Integration for Open Data on the Web

Sebastian Neumaier[1], Axel Polleres[1,2], Simon Steyskal[1], and Jürgen Umbrich[1]

[1] Vienna University of Economics and Business, Austria
[2] Complexity Science Hub Vienna, Austria

Abstract. In this lecture we will discuss and introduce challenges of integrating openly available Web data and how to solve them. Firstly, while we will address this topic from the viewpoint of Semantic Web research, not all data is readily available as *RDF* or *Linked Data*, so we will give an introduction to different *data formats* prevalent on the Web, namely, standard formats for publishing and exchanging tabular, tree-shaped, and graph data. Secondly, not all *Open Data* is really completely open, so we will discuss and address issues around *licences*, terms of usage associated with Open Data, as well as documentation of data *provenance*. Thirdly, we will discuss issues connected with (meta-)data quality issues associated with Open Data on the Web and how Semantic Web techniques and vocabularies can be used to describe and remedy them. Fourth, we will address issues about *searchability* and *integration* of Open Data and discuss in how far *semantic search* can help to overcome these. We close with briefly summarizing further issues not covered explicitly herein, such as multi-linguality, temporal aspects (archiving, evolution, temporal querying), as well as how/whether OWL and RDFS reasoning on top of integrated open data could be help.

1 Introduction

Over the last decade we have seen the World Wide Web being populated more and more by "machines". The world wide Web has evolved from its original form as a network of linked Documents, readable by humans to more and more a Web of data and APIs. That is, nowadays, even if we interact as humans with Web pages, in most cases (i) the contents of Web pages are generated from Databases in the backend, (ii) the Web content we see as humans contains annotations readable by machines, and even (iii) the way we interact with Web pages generates data (frighteningly, even often without the users being aware of), collected and stored again in databases around the globe. It is therefore valid to say that the Web of Data has become a reality and – to some extent – even the vision of the Semantic Web. In fact, this vision of the Semantic Web has itself evolved over the decades, starting with Berners-Lee et al.'s seminal article in 2001 [13] that already envisioned the future Web as "federating particular knowledge bases and databases to perform anticipated tasks for humans and their agents". Based on these ideas a lot of effort and research has been devoted

to the World Wide Web Consortium (W3C) Semantic Web activity,[3] which in 2013 has been subsumed by – i.e., renamed to – "Data Activity".[4]

In many aspects, the Semantic Web has not necessarily evolved as expected, and the biggest success stories so far do less depend on formal logics [37] than we may have expected, but more on the availability of data. Another recent article by Bernstein et al. [14] takes a backwards look on the community and summarizes successes of the Semantic Web community such as the establishment of lightweight annotation vocabularies like `Schema.org` on Web pages, or praising the uptake of large companies such as Google, Yahoo!, Microsoft, and Facebook who are developing large knowledge graphs, which however, so far these companies mostly keep closed.

Thus, if Web researchers outside of these companies want to tap into the rich sources of Data available now on the Web they need to develop their own data workflows to find relevant and usable data. To their help, more and more Open Data is being published on the Web, that is, data that is made freely available by mostly public institutions (Open Government Data) both for transparency reasons and with the goal to "fuel" a Data Economy, pushed both by the EU [29] and the G8 [72].

The present lecture notes may be viewed as partially an experience report as well as – hopefully – a guide through challenges arising when using (Open) data from the Web. The authors have been involved over the past view years in several projects and publications around the topic of Open Data integration, monitoring, and processing. The main challenges we have come across in all these projects are largely overlapping and therefore we decided to present them in the present chapter:

1. **Where to find Open Data?** (Section 2) Most Open Data nowadays can be found on so called Open Data Portals, that is, data catalogs, typically allowing API access and hosting dataset descriptions and links to actual data resources.
2. **"Low-level" data heterogeneity** (Section 3) As we will see, most of the structured data provided as Open Data is not readily available as RDF or Linked Data – the preferred formats for semantic data access described in other chapters of this volume. Different formats are much more prevalent, plus encoding issues make it difficult to access those datasets.
3. **Licenses and Provenance** (Section 4) Not all *Open Data* is really completely open, since most data on the Web is attached to different licences, terms and conditions, so we will discuss how and whether these licenses can be interpreted by machines, or, respectively how the provenance of different integrated data sources can be tracked.
4. **Quality issues** (Section 5) A major challenge for data – also often related to its provenance – is quality; on the one hand the re-use of poor quality data is obviously not advisable, but on the other hand different applications might have different demands/definitions of quality.

[3] `https://www.w3.org/2001/sw/`, last accessed 30/03/2017

[4] `https://www.w3.org/2013/data/`, last accessed 30/03/2017

5. **How to find data – Searchability?** (Section 6) Last, but not least, we will look into current solutions for search in Open Data, which we pose as a major open research challenge: whereas crawling and (keyword-based search) of human readable websites work well, this is not yet the case for structured data on the Web; we will discuss why and sketch some routes ahead.

Besides these main questions, we will conclude with summarizing issues and open questions around integrating Open Data from the Web not covered explicitly herein in Section 7, such as multi-linguality, temporal aspects (archiving, evolution, temporal querying), as well as how/whether OWL and RDFS reasoning on top of integrated open data could be help.

2 Where to find Web Data?

If we look for sources of openly available data that is widely discussed in the literature, we mainly can identify three starting points, which are partially overlapping:

- User-created open data bases
- The Linked Open Data "Cloud"
- Webcrawls
- Open Data Portals

User-created open data bases, through efforts such as Wikipedia, are large amounts of data and data-bases that have been co-created by user communities distributed around the globe; the most important ones being listed as follows:

- **DBpedia** [44] is a community effort that has created one of the biggest and most important cross-domain dataset in RDF [19] in the focal point of the so called Linked Open Data (LOD) cloud [6]. At its core is a set of declarative mappings extracting data from Wikipedia *infoboxes* and tables into RDF and it is accessible as well as through dumps also through an open query interface supporting the SPARQL [33] query language. DBpedia can therefore be well called one of the cornerstones of Semantic Web and Linked Data research being the subject and center of a large number of research papers over the past few years. Reported numbers vary as DBpedia is modular and steadily growing with Wikipeda, e.g. in 2015 DBpedia contained overall more than 3B RDF Statements [5], whereof the English DBpedia contributed 837M statements (RDF triples). Those 837M RDF triples alone amount to 4.7GB when stored in the compressed RDF format HDT [30] [6]. However, as we will see there are many, indeed far bigger other openly accessible data sources, that yet remain to be integrated, which are rather in the focus of the present chapter.

[5] `http://wiki.dbpedia.org/about/facts-figures`, last accessed 30/03/2017
[6] `http://www.rdfhdt.org/datasets/`, last accessed 30/03/2017

– **Wikidata** [74] a similar, but conceptually different effort has been started in 2012 to bring order into data items in Wikipedia, with the idea to – instead of extracting data from semi-structured Wikipages – build a database for data observations with fixed properties and datatypes, mainly with the idea to avoid extraction errors and provide means to record provenance directly with the data, with likewise 100s of millions of facts in the meantime: exact numbers are hard to give, but [71] report some statistics of the status of 2015, when Freebase was included into Wikidata; we note that counting RDF triples[7] is only partially useful, since the data representation of Wikidata is not directly comparable with the one from DBpedia [35, 36].
– **OpenStreetmap** as another example of an openly available data base that has largely been created by users contains a vast amount of geographic features to obtain an openly available and re-usable map; with currently 739.7GB (uncompressed) data in OSM's native XML format (and still 33GB compressed).[8]

The Linked Open Data "Cloud" – already mentioned above – is a manually curated collection of datasets that are published on the Web openly, adhering to the so-called Linked Data principles, defined as follows [12] (cf. chapters of previous editions of the Reasoning Web book series for good overview articles):

LDP1: use URIs as names for things;
LDP2: use HTTP URIs so those names can be dereferenced;
LDP3: return useful – herein we assume RDF – information upon dereferencing of those URIs; and
LDP4: include links using externally dereferenceable URIs.[9]

The latest iteration of the LOD Cloud [1] contains – with DBpedia in its center – hundreds of datasets with equal or even larger sizes than DBpedia, documenting a significant growth of Linked Data over the past years. Still, while often in the Semantic Web literature the LOD cloud and the "Web of Data" are implicitly equated, there is a lot of structured data available on the Web (a) either, while using RDF, not being linked to other datasets, or (b) provided in other, popular formats than RDF.

Running Web crawls is the only way to actually find and discover structured Web Data, which is both resource intensive and challenging in terms of respecting politeness rules when crawling. However, some Web crawls have been made openly available, such as the Common Crawl corpus which contains "petabytes of data collected over the last 7 years"[10]. Indeed the project has already been used to collect and analyse the availability (and quality) of structured data on the Web, e.g. in the Web Data Commons Project [50, 51].

[7] Executing the SPARQL query `SELECT (count(*) as ?C ) WHERE {?S ?P ?O }` on `https://query.wikidata.org/` gives 1.7B triples, last accessed 30/03/2017.

[8] `http://wiki.openstreetmap.org/wiki/Planet.osm`, last accessed 30/03/2017

[9] That is, within your published RDF graph, use HTTP URIs pointing to other dereferenceable documents, that possibly contain further RDF graphs.

[10] `http://commoncrawl.org/`, last accessed 30/03/2017

Table 1: Top-10 portals, ordered by datasets.

domain of portal URL	Origin	Software	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{R}\rvert$
data.gov	US	CKAN	192,738	170,524
www.data.gc.ca	Canada	CKAN	147,364	428,141
transparenz.hamburg.de	Germany	CKAN	69,147	101,874
data.noaa.gov	US	CKAN	57,934	148,343
geothermaldata.org	US	CKAN	56,388	59,804
data.gov.au	Australia	CKAN	42,116	77,900
data.gov.uk	UK	CKAN	41,615	80,980
hubofdata.ru	Russia	CKAN	28,393	62,700
openresearchdata.ch	Switzerland	CKAN	20,667	161,259
govdata.de	Germany	CKAN	19,334	55,860

Open Data portals are collections or catalogs that index metadata and link to actual data resources which have become popular over the past few years through various Open Government Data Initiatives, but also in the private sector. Apart from all the other sources mentioned so far, most of the data published openly is indexed in some kind of Open Data Portal. We therefore will discuss these portals in the rest of this paper in more detail.

Open Data portals

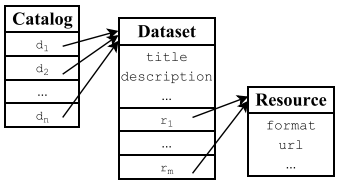


Fig. 1: High-level structure of a Data Catalog.

Most of the current "open" data form part of a dataset that is published in Open Data portals which are basically catalogues similar to digital libraries (cf. Figure 1): in such catalogues, a *dataset* aggregates a group of data files (referred to as *resources* or distributions) which are available for access or download in one or more formats (e.g., CSV, PDF, Microsoft Excel, etc.). Additionally, a dataset contains *metadata* (i.e., basic descriptive information in structured format) about these resources, e.g. authorship, provenance or licensing information. Most of these portals rely on existing software frameworks, such as CKAN[11] or Socrata,[12] that offer UI, search, and API functionalities. CKAN is the most prominent portal software framework used for publishing Open Data and is used by several governmental portals including `data.gov.uk` and `data.gov`.

For example, the Humanitarian Data Exchange[13] (see Figure 2) is a portal by the United Nations. It aggregates and publishes data about the context in which

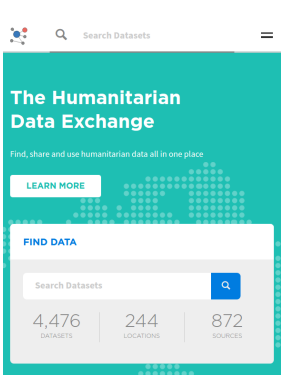
[11] `https://ckan.org/`, last accessed 30/3/2017

[12] `https://socrata.com/`, last accessed 30/3/2017

[13] `https://data.humdata.org/`, last accessed 27/3/2017

a humanitarian crisis is occurring (e.g., damage assessments and geospatial data) and data about the people affected by the crisis. The datasets on this portal are described using several metadata fields, and the metadata description can be retrieved in JSON format using the Web API of the data portal (cf. Figure 2).

The metadata description of these datasets provide download links for the actual content. For instance, the particular dataset description in Figure 2 – a dataset reporting the amounts paid by refugees to facilitate their movement to Europe – holds a URL which refers to a table (a CSV file) containing the corresponding data, displayed in Table 2.



=

```
{
  "name":"amounts-paid-by-refugees-...",
  "title": "Amounts paid by refugees...",
  "license": "Creative Commons Attribution",
  "tags": [
    "europe",
    "mediterranean",
    "refugee"
  ],
  "resources": [
    {
      "format": "CSV",
      "name": "The Money Trail - South - Prices",
      "created": "2015-10-28T21:20:40.006453",
      "url": "https://docs.google.com/...",
    }
  ],
  ...
}
```

Fig. 2: Example dataset description from the Humanitarian Data Exchange portal.

Table 2: The tabular content of the dataset in Figure 2

Route	Period	Ref crossing	Total in EUR 2014
Central Med	2010-2015	285,700	3,643,000,000
East Borders	2010-2015	5,217	72,000,000
East Med Land	2010-2015	108,089	1,751,000,000
East Med Sea	2010-2015	61,922	1,053,000,000
West African	2010-2015	1,040	4,000,000
West Balkans	2010-2015	74,347	1,589,000,000
West Med	2010-2015	29,487	251,000,000

3 Data Formats on the Web

When we discuss different available data on the Web, we already emphasized that – despite being subject of a lot of research – RDF and Linked Data are not

necessary the prevalent formats for published data on the Web. An analysis of the datasets systematically catalogued in Open Data portals will confirm this. Likewise, we will have to discuss *metadata* formats on these portals.

Data formats on Open Data portals. Table 3 shows the top used formats and the number of unique resources together with their number of portals they appear, adapted from [58], where we analysed and crawled metadata from 260 Open Data Portals for cues to the data formats in which different datasets are provided. Note, that these numbers are based on available metadata information of the datasets and can be higher due to varying spellings, misspellings, and missing metadata. Therefore, these numbers should be considered as a lower bound for the respective formats. Bold highlighted values indicate that the format is considered as open as per the Open Definition [12]:[14] the open definition sets out several guidelines of which data formats are to be considered "open", according to which we have analysed assessed openness by a list of compliant formats, cf. [58].

Table 3: Most frequent formats.

	`format`	$\|resources\|$	%	$\|portals\|$
1	**HTML**	491,891	25	74
2	PDF	182,026	9.2	83
3	**CSV**	179,892	9.1	108
4	XLS(X)	120,703	6.1	89
5	**XML**	90,074	4.6	79
6	**ZIP**	50,116	2.5	74
	...			
11	**JSON**	28,923	1.5	77
16	**RDF**	10,445	0.5	28

A surprising observation is that ∼10% of all the resources are published as PDF files. This is remarkable, because strictly speaking PDF cannot be considered as an Open Data format: while PDFs may contain structured data (e.g. in tables) there are no standard ways to extract such structured data from PDFs - or general-purpose document formats in general. Therefore, PDFs cannot be considered as machine-readable, nor as a suitable way for publishing Open Data. As we also see, RDF does not appear among the top-15 formats for Open Data publishing.[15] This underlines the previously stated hypothesis that – especially in the area of Open Government Data – openly available datasets on data portals are mostly not published as RDF or Linked Data.

[14] `http://opendefinition.org/ofd/`, last accessed 30/03/2017

[15] The numbers for the RDF serializations JSON-LD (8 resources) and TTL (55) are vanishingly small.

Also, JSON does not appear among the top ten formats in terms of numbers of published data resources on Open Data portals. Still, we include those main formats in our discussion below, as

- particularly JSON and RDF play a significant role in metadata descriptions,
- JSON is the prevalent format for many Web APIs,
- RDF, as we saw, is apart from the Linked Data cloud prevalent in Web pages and crawls through its support as an annotation format by popular search engines.

In the following we introduce some of these popular, well known, data formats on the Web and categorize them by their structure, namely, graph-based, tree-shaped, and tabular formats.

3.1 Graph-based formats

RDF, W3C recommendation since 2004 [41] and "refurbished" in 2014 [23, 19], was originally conceived as a metadata model language for describing resources on the web. It evolved (also through deployment) to a universal model and format to describe arbitrary relations between resources identified, typically, by URIs, such that they can be read and understood by machines.

RDF itself consists of statements in the form of *subject*, *predicate*, *object* triples. RDF triples can be displayed as graphs where the subjects and objects are nodes and the predicates are directed edges. RDF uses vocabularies to define the set of elements that can be used in an application. Vocabularies are similar to schemas for RDF datasets and can also define the domain and range of predicates. The graph in Figure 3 represents the metadata description of the dataset in Figure 2 in the DCAT (Data Catalog) vocabulary [48].[16]

There exist several formats to serialize RDF data. Most prominent is RDF/XML, the XML serialization first introduced in the course of 1999 W3C specification of the RDF data model, but there are also a more readable/concise textual serialization formats such as the line-based N-Triples [21] and the "Terse RDF Language" TURTLE [10] syntax. More recent, in 2014, W3C released the first recommendation for JSON-LD [68]. JSON-LD is an extension for the JSON format (see below) mostly allowing to specify namespaces for identifiers and support of URIs (supporting Linked Data principles natively in JSON) which allows the serialization of RDF as JSON, or vice versa, the transformation of JSON as RDF: conventional JSON parser and databases can be used; users of JSON-LD which are mainly interested in conventional JSON, are not required to understand RDF and do not have to use the Linked Data additions.

[16] DCAT is a vocabulary commonly used for describing general metadata about datasets. See Section 5.2 for mapping and homogenization of metadata descriptions using standard vocabularies.

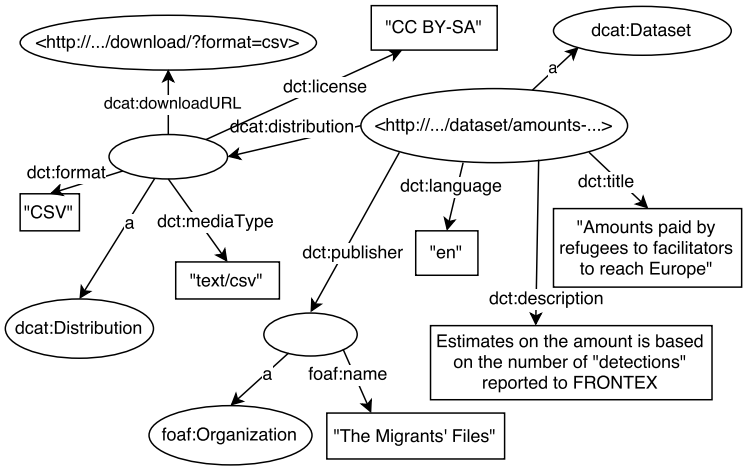


Fig. 3: RDF graph of DCAT metadata mapping of Figure 2

3.2 Tree-shaped formats

The JSON file format [18] is a so-called semi-structured file format, i.e., where documents are loosely structured without a fixed schema (as for example data in relational databases) as attribute–value pairs where values can be primitive (Strings, numbers, Booleans), arrays (sequences of values enclosed in square brackets '[',']'), or nested JSON objects (enclosed in curly braces '{','}'), thus – essentially – providing a serialization format for tree-shaped, nested structures. For an example for JSON we refer to Figure 2.

Initially, the JSON format was mainly intended to transmit data between servers and web applications, supported by web services and APIs. In the context of Open Data we often find JSON as a format to describe metadata but also to publish the actual data: also raw tabular data can easily be transformed into semi-structured and tree-based formats like JSON[17] and, therefore, is often used as alternative representation to access the data. On the other hand, JSON is the de facto standard for retrieving metadata from Open Data portals.

XML. For the sake of completeness, due to its long history, and also due to its still striking prevalence as a data exchange format of choice, we shall also mention some observations on XML. This prevalence is not really surprising since many industry standards and tools export and deliver XML, which is then used as the output for for many legacy applications or still popular for many Web APIs, e.g., in the area of geographical information systems (e.g. KML,[18]

[17] For instance, see Converter Tools on `https://project-open-data.cio.gov/`, last accessed 24/03/2017

[18] `https://developers.google.com/kml/documentation/`, last accessed 24/03/2017

GML,[19] WFS,[20] etc.). Likewise, XML has a large number of associated standards around it such as query, navigation, transformation and schema languages like XQuery,[21] XPath,[22] XSLT[23], and XML Schema[24] which are still actively developed, supported by semi-structured database systems, and other tools. XML by itself has been subject to extensive research, for example in the fields of data exchange [4, Part III] or query languages [8]. Particularly, in the context of the Semantic Web, there have also been proposals to combine XQuery with SPARQL, cf. for instance [15, 26] and references therein. The issue of interoperability between RDF and XML indeed is further discussed within the W3C in their recently started "RDF and XML Interoperability Community Group" [25] see also [16] for a summary. So, whereas JSON has probably better support in terms of developer-friendliness and recent uptake particularly through Web APIs, there is still a strong community with well-established standards behind XML technologies. For instance, schema languages or query languages for JSON exist as proposals, but their formal underpinning is still under discussion, cf. e.g. [63, 17]. Another approach would be to adopt, reuse and extend XML technologies to work on JSON itself, as for instance proposed in [26]. On an abstract level, there is not much to argue about JSON and XML just being two syntactic variants for serializing arbitrary, tree-shaped data.

3.3 Tabular data formats

Last but not least, potentially driven also by the fact that the vast majority of Open Data on the Web originates from relational databases or simply from spreadsheets, a large part of the Web of Open Data consists of tabular data. This is illustrated by the fact that two of the most prominent formats for publishing Open Data in Table 3 cover tabular data: CSV and XLS. Note particularly that both of these formats are present on more Open Data portals than for instance XML.

While XLS (the export format of Microsoft Excel) is obviously a proprietary open format, CSV (comma-separated values) is a simple, open format with a standard specification allowing to serialize arbitrary tables as text (RFC4180) [67]. However, as we have shown in a recent analysis [54], compliance with this standard across published CSVs is not consistent: in Open Data corpus containing 200K tabular resources with a total file size of 413GB we found out that out of the resources in Open Data portals labelled as a tabular only 50% can be considered CSV files. In this work we also investigated different use of delimiters, the availability of (multiple) header rows or cases where single CSV files actually contain multiple tables as common problems.

[19] `http://www.opengeospatial.org/standards/gml`, last accessed 24/03/2017
[20] `http://www.opengeospatial.org/standards/wfs`, last accessed 24/03/2017
[21] `https://www.w3.org/TR/xquery-30/`, last accessed 24/03/2017
[22] `https://www.w3.org/TR/xpath-30/`, last accessed 24/03/2017
[23] `https://www.w3.org/TR/xslt-30/`, last accessed 24/03/2017
[24] `https://www.w3.org/XML/Schema`, last accessed 24/03/2017
[25] `https://www.w3.org/community/rax/`, last accessed 24/03/2017

Last, but not least, as opposed to tabular data in relational databases, which typically adhere to a fixed schema and constraints, these constraints, datatype information and other schema information is typically lost when being exported and re-published as CSVs. This loss can be compensated partially by adding this information as additional metadata to the published tables; one particular format for such kind of metadata has been recently standardized by the W3C [65]. For more details on the importance of metadata we refer also to Section 5 below.

3.4 Data Formats – Summary

Overall, while data formats are often only considered syntactic sugar, one should not underestimate the issues about conversions, scripts parsing errors, stability of tools, etc. where often significant amounts of work incurs. While any data can be converted/represented in principle into a CSV, XML, or RDF serialization, one should keep in mind that a canonical, "dumb" serialization in RDF by itself, does not "add" any "semantics".

For instance, a naive RDF conversion (in Turtle syntax) of the CSV in Table 2 could look as follows in Fig. 4, but would obviously not make the data more "machine-readbable" or easier to process.

```
@prefix : <http://www.example.org/> .

:c1 rdfs:label "Route".
:c2 rdfs:label "Period".
:c3 rdfs:label "Ref_crossing".
:c4 rdfs:label "Total in EUR 2014".

[:c1 "Central Med"; :c2 "2010-2015", :c3 "285,700"; :c4 "3,643,000,000"].
[:c1 "East Borders"; :c2 "2010-2015"; :c3 "5,217"; :c4 "72,000,000" ].
[:c1 "East Med Land" ; :c2 "2010-2015"; :c3 "108,089" ; :c4 "1,751,000,000"].
[:c1 "East Med Sea"; :c2 "2010-2015" ; :c3 "61,922"; :c4"1,053,000,000"].
[:c1 "West African"; :c2 "2010-2015"; :c3 "1,040"; :c4 "4,000,000"].
[:c1 "West Balkans"; :c2 "2010-2015"; :c3 "74,347"; :c4 "1,589,000,000"].
[:c1 "West Med"; :c2 "2010-2015"; :c3 "29,487"; :c4 "251,000,000"].
```

Fig. 4: Naive conversion of tabular data into RDF

We would leave coming up with a likewise naive (and probably useless) conversion to XML or JSON to the reader: the real intelligence in mapping such data lies in finding suitable ontologies to describe the properties representing columns c1 to c4, recognizing the datatypes of the column values, linking names such as "East Med Sea" to actual entities occurring in other datasets, etc. Still, typically, in data processing workflows more than 80% of the the effort to data conversion, pre-processing and cleansing tasks.

Within the Semantic Web, or to be more precise, within the closed scope of Linked Data this problem and the steps involved have been discussed in depth in the literature [7, 60]. A partial instantiation of a platform which shall provide a cleansed and integrated version of the Web of Linked Data is presented by the

LOD-Laundromat [11] project: here, the authors present a cleansed unified store of Linked Data as an experimental platform for the whole Web of Linked Data, mostly containing the all datasets of the current LOD cloud, are made available. Querying this platform efficiently and investigating the properties of this subset of the Web of Data is a subject of active ongoing research, despite only Linked RDF data has been considered: however, building such a platform for the scale of arbitrary Open Data on the Web, or even only for the data accumulated in Open Data portals would demand a solution at a much larger scale, handling more tedious cleansing, data format conversion and schema integration problems.

4 Licensing and Provenance of Data

Publishing data on the Web is more than just making it publicly accessible. When it comes to consuming publicly accessible data, it is crucial for data consumers to be able to assess the trustworthiness of the data as well as being able to use it on a secure legal basis and to know where the data is coming from, or how it has been pre-processed. As such, if data is to be published on the Web, appropriate metadata (e.g., describing the data's provenance and licensing information) should be published alongside with it, thus making published data as self-descriptive as possible (cf. [34]).

4.1 Open Data Licensing in Practice

While metadata about terms and conditions under which a dataset can be reused are essential for its users, according to the Linked Open Data Cloud web page, only less than 8% of the linked data datesets provide license information[26]

Within Open data portals, the situation seems slightly better overall: more than 50% of the monitored datasets in the Open Data portals in the Portalwatch project (see Section 5 below) announce somehow in the metadata some kind of license information [58]. The most prevalent license keys used in Open Data portals [58] are listed in Table 4.

While most of the provided license definitions lack a machine-readable description that would allow automated compatibility checks of different licenses or alike, some are not even compliant with *Open Definition* conformant data licenses (cf. Table 5).

In order to circumvent these shortcomings, different RDF vocabularies have been introduced to formally describe licenses as well as provenance information of datasets, two of which (ODRL and PROV) we will briefly introduce in the next two subsections.

[26] `http://lod-cloud.net/state/state_2014/#toc10`, last accessed 01/05/2017.

[27] `http://open-data.europa.eu/kos/licence/EuropeanCommission`, last accessed 24/03/2017

Table 4: Top-10 licenses.

`license_id`	$\|datasets\|$	%	$\|portals\|$
ca-ogl-lgo	239662	32.3	1
notspecified	193043	26	71
dl-de-by-2.0	55117	7.4	7
CC-BY-4.0	49198	6.6	84
us-pd	35288	4.8	1
OGL-UK-3.0	33164	4.5	18
other-nc	27705	3.7	21
CC0-1.0	9931	1.3	36
dl-de-by-1.0	9608	1.3	6
Europ.Comm.[27]	8604	1.2	2
others	80164	10.8	

Table 5: Open Definition conformant data licenses [40]

License
Creative Commons Zero (CC0)
Creative Commons Attribution 4.0 (CC-BY-4.0)
Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)
Open Data Commons Attribution License (ODC-BY)
Open Data Commons Public Domain Dedication and Licence (ODC-PDDL)
Open Data Commons Open Database License (ODC-ODbL)

4.2 Making Licenses machine-readable

The Open Digital Rights Language (ODRL) [39] is a comprehensive policy expression language (representable with a resp. RDF vocabulary) that has been demonstrated to be suitable for expressing fine-grained access restrictions, access policies, as well as licensing information for Linked Data as shown in [20, 69].

An *ODRL Policy* is composed of a set of *ODRL Rules* and an *ODRL Conflict Resolution Strategy*, which is used by the enforcement mechanism to ensure that when conflicts among rules occur, a system either grants access, denies access or generates an error in a non-ambiguous manner.

An *ODRL Rule* either permits or prohibits the execution of a certain action on an asset (e.g. the data requested by the data consumer). The scope of such rules can be further refined by explicitly specifying the party/parties that the rule applies to (e.g. Alice is allowed to access some dataset), using constraints (e.g. access is allowed until a certain date) or in case of permission rules by defining duties (e.g. a payment of 10 euros is required).

Listing 1.1 demonstrates how ODRL can be used to represent the *CreativeCommons* license `CC-BY 4.0`.

Listing 1.1: `CC-BY 4.0` represented in ODRL

```
<http://purl.org/NET/rdflicense/cc-by4.0>
      a odrl:Policy ;
      rdfs:label "Creative Commons CC-BY" ;
      rdfs:seeAlso
         <http://creativecommons.org/licenses/by/4.0/legalcode> ;
      dct:source <http://creativecommons.org/licenses/by/4.0/> ;
      dct:hasVersion "4.0" ;
      dct:language <http://www.lexvo.org/page/iso639-3/eng> ;
      odrl:permission [
         odrl:action cc:Distribution,
                     cc:Reproduction, cc:DerivativeWorks ;
         odrl:duty [
            odrl:action cc:Notice, cc:Attribution
         ]
      ] .
```

Policy Conflict Resolution A rule that permits or prohibits the execution of an action on an asset could potentially affect related actions on that same asset. Explicit relationships among actions in ODRL are defined using a subsumption hierarchy, which states that an action α_1 is a broader term for action α_2 and thus might influence its permission/prohibition (cf. *Figure* 5). On the other hand implicit dependencies indicate that the permission associated with an action α_1 requires another action α_2 to be permitted also. Implicit dependencies can only be identified by interpreting the natural language description of the respective *ODRL Actions* (cf. *Figure* 6). As such, when it comes to the enforcement of access policies defined in ODRL, there is a need for a reasoning engine which is capable of catering for both explicit and implicit dependencies between actions.

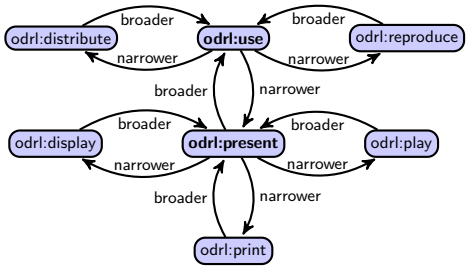


Fig. 5: Example of explicit dependencies in ODRL.

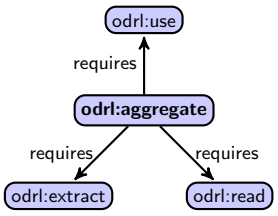


Fig. 6: Example of implicit dependencies in ODRL.

4.3 Tracking the Provenance of Data

In order to handle the unique challenges of diverse and unverified RDF data spread over RDF datasets published at different URIs by different data publishers across the Web, the inclusion of a notion of provenance is necessary. The W3C PROV Working Group [49] was chartered to address these issues and developed an RDF vocabulary to enable annotation of datasets with interchangeable provenance information. On a high level PROV distinguishes between entities, agents, and activities (see Figure 7). A `prov:Entity` can be all kinds of things, digital

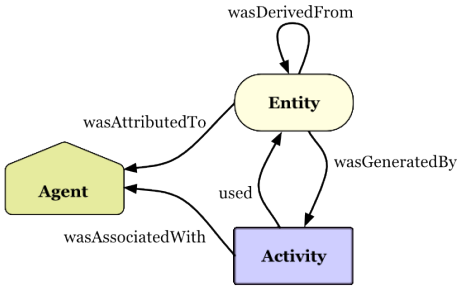


Fig. 7: The core concepts of PROV. Source: Taken from [49]

or not, which are created or modified. Activities are the processes which create or modify entities. An `prov:Agent` is something or someone who is responsible for a `prov:Activity` (and indirectly also for an entity).

Listing 1.2 illustrates a PROV example (all other triples removed) of two observations, where observation `ex:obs123` was derived from another observation `ex:obs789` via an activity `ex:activity456` on the 1st of January 2017 at 01:01. This derivation was executed according to the rule `ex:rule937` with an agent `ex:fred` being responsible. This use of the PROV vocabulary models tracking of source observations, a timestamp, the conversion rule and the responsible agent (which could be a person or software component). The PROV vocabulary could thus be used to annotated whole datasets, or single observations (data points) within such dataset, or, respectively any derivations and aggregations made from open data sources re-published elsewhere.

Listing 1.2: PROV example

```
ex:obs123 a prov:Entity ;
    prov:generatedAtTime "2017-01-01T01:01:01"^^xsd:dateTime;
    prov:wasGeneratedBy ex:activity456 ;
    prov:wasDerivedFrom ex:obs789 .

ex:activity456 a prov:Activity;
    prov:qualifiedAssociation [
```

```
    a Association ;
    prov:wasAssociatedWith ex:fred ;
    prov:hadPlan ex:rule397 .
] .
```

5 Metadata Quality Issues and Vocabularies

The Open Data Portalwatch project [58] has originally been set up as a framework for monitoring and *quality assessment* of (governmental) Open Data portals, see `http://data.wu.ac.at/portalwatch`. It monitors data from portals using the CKAN, Socrata, and OpenDataSoft software frameworks, as well as portals providing their metadata in the DCAT RDF vocabulary.

Currently, as of the second week of 2017, the framework monitors 261 portals, which describe in total about 854k datasets with more than 2 million distributions, i.e., download URLs (cf. Table 6). As we monitor and crawl the metadata of these portals in a weekly fashion, we can use the gathered insights in two ways to enrich the crawled metadata of these portals: namely, (i) we publish and serve the integrated and homogenized metadata descriptions in a weekly, versioned manner, (ii) we enrich these metadata descriptions by assessed quality measures along several dimensions. These dimensions and metrics are defined on top of the DCAT vocabulary, which allows us to treat and assess the content independent of the portal's software and own metadata schema.

Table 6: Monitored portals and datasets in Portalwatch

	total	**CKAN**	**Socrata**	**OpenDataSoft**	**DCAT**
portals	261	149	99	11	2
datasets	854,013	767,364	81,268	3,340	2,041
URLs	2,057,924	1,964,971	104,298	12,398	6,092

The quality assessment is performed along the following dimensions: (i) The *existence* dimension consists of metrics checking for important information, e.g., if there is contact information in the metadata. (ii) The metrics of the *conformance* dimension check if the available information adheres to a certain format, e.g., if the contact information is a valid email address. (iii) The *open data* dimension's metrics test if the specified format and license information is suitable to classify a dataset as open. The formalization of all quality metrics currently assessed on the Portalwatch platform and implementation details can be found in [58].

5.1 Heterogeneous metadata descriptions

Different Open Data portals use different metadata keys to describe the datasets they host, mostly dependent on the software framwwork under which the portal runs: while the schema for metadata descriptions on Socrata and OpenDataSoft portals are fixed and predefined (they use their own vocabulary and metadata keys), CKAN provides a higher flexibility in terms of own, per portal, metadata schema and vocabulary. Thus, overall, the metadata that can be gathered from Open Data Portals show a high degree of heterogeneity.

In order to provide the metadata in a standard vocabulary, there exists a CKAN-to-DCAT extension for the CKAN software that defines mappings for datasets and their resources to the corresponding DCAT classes `dcat:Dataset` and `dcat:Distribution` and offers it via the CKAN API. However, in general it cannot be assumed that this extension is deployed for all CKAN portals: we were able to retrieve the DCAT descriptions of datasets for 93 of the 149 active CKAN portals monitored by Portalwatch [59].

Also, the CKAN software allows portal providers to include additional metadata fields in the metadata schema. When retrieving the metadata description for a dataset via the CKAN API, these keys are included in the resulting JSON. However, it is neither guaranteed that the CKAN-to-DCAT conversion of the CKAN metadata contains these extra fields, nor that these extra fields, if exported, are available in a standardized way.

We analysed the metadata of 749k datasets over all 149 CKAN portals and extracted a total of 3746 distinct extra metadata fields [59]. Table 7 lists the most frequently used fields sorted by the number of portals they appear in; most frequent `spatial` in 29 portals. Most of these cross-portal extra keys are generated by widely used CKAN extensions. The keys in Table 7 are all generated by the harvesting[28] and spatial extension.[29]

We manually selected mappings for the most frequent extra keys if they are not already included in the mapping; the selected properties are listed in the "DCAT key" column in Table 7 and are included in the homogenized, re-exposed, metadata descriptions, cf. Section 5.2. In case of an ?-cell, we were not able to choose an appropriate DCAT core property.

5.2 Homogenizing metadata using DCAT and other metadata vocabularies

The W3C identified the issue of heterogeneous metadata schemas across the data portals, and proposed an RDF vocabulary to solve this issue: The metadata standard DCAT [48] (Data Catalog Vocabulary) describes data catalogs and corresponding datasets. It models the datasets and their distributions (published data in different formats) and re-uses various existing vocabularies such as Dublin Core terms [75], and the SKOS [52] vocabulary.

[28] `http://extensions.ckan.org/extension/harvest/`, last accessed 24/03/2017

[29] `http://docs.ckan.org/projects/ckanext-spatial/en/latest/`, last accessed 24/03/2017

Table 7: Most frequent extra keys

Extra key	Portals	Datasets	Mapping
spatial	29	315,652	dct:spatial
harvest_object_id	29	514,489	?
harvest_source_id	28	486,388	?
harvest_source_title	28	486,287	?
guid	21	276,144	dct:identifier
contact-email	17	272,208	dcat:contactPoint
spatial-reference-system	16	263,012	?
metadata-date	15	265,373	dct:issued

The recent DCAT application profile for data portals in Europe (DCAT-AP)[30] extends the DCAT core vocabulary and aims towards the integration of datasets from different European data portals. In its current version (v1.1) it extends the existing DCAT schema by a set of additional properties. DCAT-AP allows to specify the version and the period of time of a dataset. Further, it classifies certain predicates as "optional", "recommended" or "mandatory". For instance, in DCAT-AP it is mandatory for a `dcat:Distribution` to hold a `dcat:accessURL`.

An earlier approach, in 2011, is the the VoID vocabulary [3] published by W3C as an Interest Group Note. VoID – the Vocabulary for Interlinked Datasets – is an RDF schema for describing metadata about linked datasets: it has been developed specifically for data in RDF representation and is therefore complementary to the DCAT model and not fully suitable to model metadata on Open Data portals (which usually host resources in various formats) in general.

In 2011 Fürber and Hepp [32] proposed an ontology for data quality management that allows the formulation of data quality, cleansing rules, a classification of data quality problems and the computation of data quality scores. The classes and properties of this ontology include concrete data quality dimensions (e.g., completeness and accuracy) and concrete data cleansing rules (such as whitespace removal) and provides a total of about 50 classes and 50 properties. The ontology allows a detailed modelling of data quality management systems, and might be partially applicable and useful in our system and to our data. However, in the Open Data Portalwatch we decided to follow the W3C Data on the Web Best Practices and use the more lightweight Data Quality Vocabulary for describing the quality assessment dimensions and steps.

More recently, in 2015 Assaf et al. [5] propose HDL, an harmonized dataset model. HDL is mainly based on a set of frequent CKAN keys. On this basis, the authors define mappings from other metadata schemas, including Socrata, DCAT and Schema.org.

[30] `https://joinup.ec.europa.eu/asset/dcat_application_profile/description`, last accessed 24/03/2017

Metadata mapping by the Open Data Portalwatch framework. In order to offer the harvested datasets in the Portalwatch project in a homogenized and standardised way, we implemented a system that re-exposes data extracted from Open Data portal APIs such as CKAN [59]: the output formats include a subset of W3C's DCAT with extensions and `Schema.org`'s Dataset-oriented vocabulary.[31] We enrich the integrated metadata by the quality measurements of the Portalwatch framework available as RDF data using the Data Quality Vocabulary[32] (DQV). To further describe tabular data in our dataset corpus we use simple heuristics to generate additional metadata using the vocabulary defined by the W3C CSV on the Web working group [65], which we likewise add to our enriched metadata. We use the PROV ontology (cf. Section 4.3) to record and annotate the provenance of our generated/published data (which is partially generated by using heuristics). The example graph in Figure 8 displays the generated data for the DCAT dataset, the quality measurements, the CSV metadata, and the provenance information.

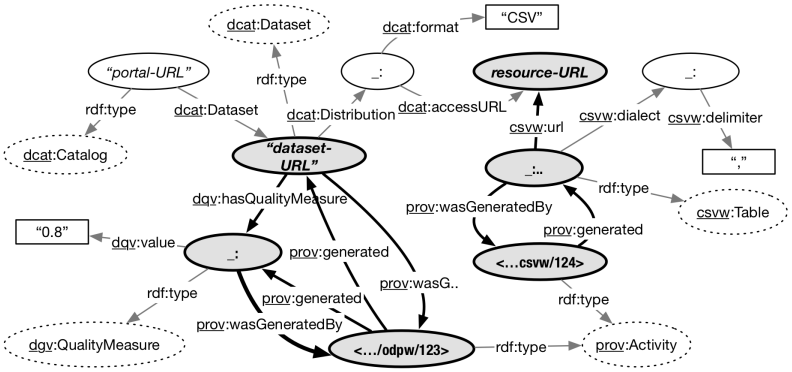


Fig. 8: The mapped DCAT dataset is further enriched by three additional datasets (indicated by the bold edges): (i) each DCAT dataset is associated to a set of quality measurements; (ii) there is additional provenance information available for the generated RDF graph; (iii) in case the corresponding distribution is a table we generated CSV specific metadata such as the delimiter and the column headers.

6 Searchability and Semantic Annotation

The popular Open Data portal software frameworks (e.g., CKAN, Socrata) offer search interfaces and APIs. However, the APIs typically allow only search over

[31] Google Research Blog entry, `https://research.googleblog.com/2017/01/facilitating-discovery-of-public.html`, last accessed 27/01/2017.

[32] `https://www.w3.org/TR/vocab-dqv/`, last accessed 24/03/2017

the metadata descriptions of the datasets, i.e., the title, descriptions and tags, and therefore rely on complete and detailed meta-information. Nevertheless, if an user wants to find data for a specific entity this search might be not successful. For instance, a search for data about "Vienna" at the Humanitarian Data Exchange portal gives no results, even though there are relevant datasets in the portal such as "World – Population of Capital Cities".

6.1 Open Data Search: state of the art

Overall, to the best of our knowledge, there is not much substantial research in the area of search and querying for Open Data. A straightforward approach to offer search over the data is to index the documents as text files into typical keyword search systems. Keyword search is already addressed and partially solved by full-text search indices, as they exist by search engines such as Google. However, these systems do not exploit the underlying structure of the dataset. For instance, a default full-text indexer considers a CSV table as a document and the cells get indexed as (unstructured) tokens. A search query for tables containing the terms "Vienna" and "Berlin" in the same column is not possible using these existing search systems. In order to enable such a structured search over the content of tables an alternative data model is required.

In a current table search prototype[33] we enable these query use-cases while utilizing existing state-of-the-art document-based search engines. We use the search engine Elasticsearch[34] and index the rows and columns of a table as separated documents, i.e., we add a new document for each column and for each row containing all values of the respective row/column. By doing so we store each single cell twice in the search system. This particular data model enables to define multi-keyword search over rows and columns. For instance, queries for which the terms "Vienna" and "Berlin" appear within the same column.

Recently, the Open Data Network project[35] addresses the searchability issue by providing a search and query answering framework on top of Socrata portals. The UI allows to start a search with a keyword and suggested matching datasets or already registered questions. However, the system relies on the existing Socrata portal ecosystem with its relevant data API[36]. This API allows to programmatically access the uploaded data and apply filters on columns and rows.

The core challenge for search & query over tabular data is to process and and build an index over a large corpus of heterogeneous tables. In 2016, we assessed the table heterogeneity for over 200k Open Data CSV files [54]. We found that a typical Open Data CSV file has less than 100kB (the biggest with over 25GB) and consists of 14 columns and 379 rows. An interesting observation was that ∼50% of the inspected header values were composed of camel case, suggesting that the

[33] `http://data.wu.ac.at/csvengine`, last accessed 24/03/2017
[34] `https://www.elastic.co/products/elasticsearch`, last accessed 24/03/2017
[35] `https://www.opendatanetwork.com`, last accessed 24/03/2017
[36] `https://dev.socrata.com`, last accessed 24/03/2017

table was exported from a relation table. Regarding the data types, roughly half of the columns consists of numerical data types. As such, Open Data CSV tables have different numbers of columns and rows and column values can belong to different data types. Some of the CSV files contain multiple tables and the tables itself can be non well-formed, meaning that there exists multiple-headers or the rows with aggregated values over the previous rows.

To the best of our knowledge, the research regarding querying over thousands of heterogeneous tables is fairly sparse. One of the initial work towards search and query over tables was the work by Das Sarma et. al. in 2012[25]. The authors propose a system to find for a given input table a set of related Web tables. The approach relies on the assumptions that tables have an "entity" column (e.g. the player column in a table about tennis players) and introduces relatedness metrics for tables (either for joining two tables or appending one table to the other). the authors propose a set of high-level features for *grouping* tables to handle the large amount of heterogeneous tables and to reduce the search space for a given input table. Eventually, the system itself returns tables which either can be joined with the input table (via the entity column) or can be append to the input table (adding new rows).

The idea of finding related tables is also closely relate to the research of finding inclusion dependencies (IND), that are relation such as column$A \subseteq$ column$B$. A core application for these dependencies is the discovery of foreign key relations across tables, but they are also used in data integration [53] scenarios, query optimization, and schema redesign [62]. The task of finding INDs gets harder with the number of tables and columns and the scalable and efficient discovery of inclusion dependencies across several tables is a well-known challenge in database research [9, 62, 43]. The state of the art research combines probabilistic and exact data structures to approximate the INDs in relational datasets. The algorithm guarantees to correctly find all INDs and only adds false positives INDs with a low probability [42].

Another promising direction is the work of Liu et. al. in 2014 which investigates the fundamental differences between relation data and JSON data management [46]. Consequently, the authors derive three architectural principles to facilitate a schema-less development within traditional relation database management systems. The first principle is to store JSON as JSON in the RDBMS. The second principle is to use the query language SQL as a Set-oriented Query Language rather than a Structured Query Language. The third principle is to use available partial schema-aware indexing methods but also schema agnostic indexing. While this work focuses on JSON and XML, it would be interesting to study and establish similar principles for tabular data and how this can be applied and benefit for search and querying.

Enabling search and querying over Open Data could benefit from many insights from the research around semantic search systems. The earlier semantic search systems such as Watson [24], Swoogle [27] or FalconS [22] provided search and simple querying over collections of RDF data. More advanced systems, such as SWSE [38] or Sindice.com [61] focused on indexing RDF document at web-

Portal	Tables	$\overline{cols}$	$\overline{num.cols}$	w/o Header	Num. H.	Mapped
AT	968	13	8	154	6,482	1,323
EU	357	20	4	223	1,233	349

Table 8: Header mapping of CSVs in Open Data portals

scale. SWSE is a scalable entity lookup system operating over an integrated data, while Sindice.com provided keyword search and entity lookups using an inverted document index. Surprisingly, published research around semantic search slowed down. However, the big search engine players on the market such as Google or Bing utilise semantic search approaches to provide search over their internal knowledge graph.

6.2 Annotation, labelling, and integration of tabular data

Text-based search engines such as Elasticsearch, however, do not integrate any semantic information of the data sources and therefore do not enable search based on concepts, synonyms or related content. For instance, to enable a search for the concept "population" over a set of resources (that do not contain the string "population"), it is required that the tables (and their columns, respectively) are labelled and annotated correctly.

There exists an extensive body of research in the Semantic Web community in semantic annotation and linking of tabular data sources. The majority of these approaches [2, 28, 45, 55, 66, 70, 73, 76] assume well-formed relational tables and try to derive semantic labels for attributes in these structured data sources (such as columns in tables) which are used to (i) map the schema of the data source to ontologies or existing semantic models or (ii) categorize the content of a data source.

Given an existing knowledge base, these approaches try to discover concepts and named entities in the table, as well as relations among them, and link them to elements and properties in the knowledge base. This typically involves finding potential candidates from the knowledge base that match particular table components (e.g., column header, or cell content) and applying inference algorithms to decide the best mappings.

However, in typical Open Data portals many data sources exist where such textual descriptions (such as column headers or cell labels) are missing or cannot be mapped straightforwardly to known concepts or properties using linguistic approaches, particularly when tables contain many numerical columns for which we cannot establish a semantic mapping in such manner. Indeed, a major part of the datasets published in Open Data portals comprise tabular data containing many numerical columns with missing or non human-readable headers (organisational identifiers, sensor codes, internal abbreviations for attributes like "population count", or geo-coding systems for areas instead of their names, e.g. for districts, etc.) [47].

In [57] we verified this observation by inspecting 1200 tables collected from the European Open Data portal and the Austrian Government Open Data

Portal and attempted to map the header values using the BabelNet service (`http://babelnet.org`): Table 8 lists our findings; an interesting observation is that the AT portal has an average number of 20 columns per table with an average of 8 numerical columns, while the EU portal has larger tables with an average of 4 out of 20 columns being numerical. Regarding the descriptiveness of possible column headers, we observed that 28% of the tables have missing header rows. Eventually, we extracted headers from 7714 out of around 10K numerical columns and used the BabelNet service to retrieve possible mappings. We received only 1472 columns mappings to BabelNet concepts or instances, confirming our assumption that many headers in Open Data CSV files cannot easily be semantically mapped.

Therefore, we propose in [57] an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values, i.e., the numerical data in a certain column. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. We assign different labels/contexts with different confidence values and this way our approach could potentially be combined with the previous introduced textual labelling techniques for further label refinement.

7 Conclusions, including Further Issues and Challenges

In this chapter we gave a rough overview over the still persisting challenge of integrating and finding data on the Web. We focused on Open Data and provided some starting points for finding large amounts of nowadays available structured data, the processing of which still remains a major challenge: on the one hand, because the introduction of Semantic Web Standards such as RDF and OWL did not yet find adoption and there is still a large variety in terms of formats to publish structured data on the Web. On the other hand, even the use of such standard formats alone would not alleviate the issue of findability of said data. Proper search and indexing techniques for structured data and its metadata need to be devised. Moreover, metadata needs to be self-descriptive, that is, it needs to not only describe what published datasets contain, but also how the data was generated (provenance) or under which terms it can be used (licenses). Overall, one could say that despite the increased availability of data on the Web, (i) there are still a number of challenges to be solved before we can call it a Semantic Web, and (ii) one often needs to be ready to manually pre-process and align data before automated reasoning techniques can be applied. Projects such as the Open Data Portalwatch, a monitoring framework for Open Data portals worldwide, from which most of our insights presented in this paper were derived, are just a starting point in the direction of making this Web of data machine-processable: there is a number of aspects that we did *not* cover herein, such as monitoring the evolution of datasets, archiving such evolving data, or querying Web data over time, cf. [31] for some initial research on this topic.

Nor did we discuss attempts to reason over Web data "in the wild" using OWL and RDFS, which we had investigated on the narrower scope of Linked Data some years ago [64], but which will impose far more challenges when taking into account the vast amounts of data not yet linked to the so called Linked Data cloud, but available through Open Data Portals. Lastly, another major issue we did not discuss in depth is multi-linguality: data (content) as well as metadata associated with Open Data is published in different languages with different language descriptions and thereby a lot of "Open" information is only accessible to speakers of the respective languages, leave aside impossible to integrate for machines: still recent progress in machine translation or multi-lingual Linked Data corpora like Babelnet [56] could contribute to solving this puzzle.

You will find further starting points in these directions in the present volume, or also previous editions of the Reasoning Web summer school. We hope these starting points serve as an inspiration for further research on making machines understand openly available data on the Web and thus bringing us closer to the original vision of the Semantic Web, an ongoing journey.

Acknowledgements The work presented in this paper has been supported by the Austrian Research Promotion Agency (FFG) under the projects ADEQUATe (grant no. 849982) and DALICC (grant no. 855396).

References

1. Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch, and Richard Cyganiak. Linking open data cloud diagram 2017, 2017.
2. Marco D. Adelfio and Hanan Samet. Schema extraction for tabular data on the web. *Proceedings of the VLDB Endowment*, 6(6):421–432, April 2013.
3. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. `https://www.w3.org/TR/void/`, March 2011.
4. Marcelo Arenas, Pablo Barceló, Leonid Libkin, and Filip Murlak. *Foundations of Data Exchange*. Cambridge University Press, 2014.
5. Ahmad Assaf, Raphaël Troncy, and Aline Senart. HDL - Towards a harmonized dataset model for open data portals. In *PROFILES 2015, 2nd International Workshop on Dataset Profiling & Federated Search for Linked Data, Main conference ESWC15, 31 May-4 June 2015, Portoroz, Slovenia*, Portoroz, Slovenia, 05 2015. CEUR-WS.org.
6. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proc. of ISWC*, 2007.
7. Sören Auer and Jens Lehmann. Creating knowledge out of interlinked data. *Semantic Web*, 1(1-2):97–104, 2010.
8. James Bailey, François Bry, Tim Furche, and Sebastian Schaffert. Web and semantic web query languages: A survey. In *Reasoning Web, First International Summer School 2005*, pages 35–133, Msida, Malta, July 2005.

9. Jana Bauckmann, Ziawasch Abedjan, Ulf Leser, Heiko Müller, and Felix Naumann. Discovering conditional inclusion dependencies. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 2094–2098, 2012.
10. David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. RDF 1.1 Turtle: The Terse RDF Triple Language. W3C Recommendation, February 2014. `http://www.w3.org/TR/turtle/`.
11. Wouter Beek, Laurens Rietveld, Stefan Schlobach, and Frank van Harmelen. LOD laundromat: Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing*, 20(2):78–81, 2016.
12. Tim Berners-Lee. Linked Data. W3C Design Issues, July 2006. From `http://www.w3.org/DesignIssues/LinkedData.html`; retr. 2017/03/31.
13. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, pages 29–37, May 2001.
14. Abraham Bernstein, James Hendler, and Natasha Noy. The semantic web. *Communications of the ACM*, 59(9):35–37, January 2016.
15. Stefan Bischof, Stefan Decker, Thomas Krennwallner, Nuno Lopes, and Axel Polleres. Mapping between RDF and XML with XSPARQL. 1(3):147–185, 2012.
16. Marta Borriello, Christian Dirschl, Axel Polleres, Phil Ritchie, Frank Salliau, Felix Sasaki, and Giannis Stoitsis. From xml to rdf step by step: approaches for leveraging xml workflows with linked data. In *XML Prague 2016 – Conference Proceedings*, pages 121–138, Prague, Czech Republic, February 2016.
17. Pierre Bourhis, Juan L. Reutter, Fernando Suárez, and Domagoj Vrgoc. JSON: data model, query languages and schema specification. *CoRR*, abs/1701.02221, 2017.
18. T. Bray. The JavaScript Object Notation (JSON) Data Interchange Format. Internet Engineering Task Force (IETF) RFC 7159, March 2014.
19. Dan Brickley and R.V. Guha. RDF Schema 1.1. W3C Recommendation, February 2014. `http://www.w3.org/TR/rdf-schema/`.
20. Elena Cabrio, Alessio Palmero Aprosio, and Serena Villata. These are your rights. In *Proceedings of the 11th Extended Semantic Web Conference (ESWC)*, 2014.
21. Gavin Carothers and Andy Seaborne. RDF 1.1 N-Triples: A line-based syntax for an RDF graph. W3C Recommendation, February 2014. `http://www.w3.org/TR/rdf-schema/`.
22. Gong Cheng, Weiyi Ge, and Yuzhong Qu. Falcons: Searching and browsing entities on the semantic web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1101–1102, New York, NY, USA, 2008. ACM.
23. Richard Cyganiak, David Wood, Markus Lanthaler, , Graham Klyne, Jeremy J. Carroll, and Brian Mcbride. RDF 1.1 Concepts and Abstract Syntax. Technical report, 2014.
24. Mathieu d'Aquin and Enrico Motta. Watson, more than a semantic web search engine. *Semant. web*, 2(1):55–63, January 2011.
25. Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 817–828. ACM, 2012.
26. Daniele Dell'Aglio, Axel Polleres, Nuno Lopes, and Stefan Bischof. Querying the web of data with XSPARQL 1.1. In *ISWC2014 Developers Workshop*, volume 1268 of *CEUR Workshop Proceedings*. CEUR-WS.org, October 2014.

27. Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 652–659, New York, NY, USA, 2004. ACM.
28. Ivan Ermilov, Sören Auer, and Claus Stadler. User-driven semantic mapping of tabular data. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 105–112, New York, NY, USA, 2013. ACM.
29. European Commission. Towards a thriving data-driven economy, July 2014.
30. Javier D. Fernández, Miguel A. Martınez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF Representation for Publication and Exchange (HDT). 19(2), 2013.
31. Javier David Fernández Garcia, Jürgen Umbrich, Magnus Knuth, and Axel Polleres. Evaluating query and storage strategies for RDF archives. In *12th International Conference on Semantic Systems (SEMANTICS)*, ACM International Conference Proceedings Series, pages 41–48. ACM, September 2016.
32. Christian Fürber and Martin Hepp. Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 1st International Workshop on Linked Web Data Management*, LWDM '11, pages 1–8, New York, NY, USA, 2011. ACM.
33. Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. W3C Recommendation, March 2013. available at `http://www.w3.org/TR/sparql11-query/`.
34. Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.
35. Daniel Hernández, Aidan Hogan, and Markus Krötzsch. Reifying RDF: what works well with wikidata? In *Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems co-located with 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 11, 2015.*, pages 32–47, 2015.
36. Daniel Hernández, Aidan Hogan, Cristian Riveros, Carlos Rojas, and Enzo Zerega. Querying wikidata: Comparing sparql, relational and graph databases. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, pages 88–103, 2016.
37. Pascal Hitzler, Jens Lehmann, and Axel Polleres. Logics for the semantic web. In Dov M. Gabbay, Jörg H. Siekmann, and John Woods, editors, *Computational Logic*, volume 9 of *Handbook of the History of Logic*, pages 679–710. Elesevier, 2014.
38. Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. *J. Web Sem.*, 9(4):365–401, 2011.
39. Renato Iannella and Serena Villata. Odrl information model. W3C Working Draft, 2017. `https://www.w3.org/TR/odrl-model/`.
40. Open Knowledge International. Open Definition Conformant Licenses, April 2017. From `http://opendefinition.org/licenses/`; retr. 2017/04/28.
41. Graham Klyne and Jeremy J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, 2004.
42. Sebastian Kruse, Thorsten Papenbrock, Christian Dullweber, Moritz Finke, Manuel Hegner, Martin Zabel, Christian Zöllner, and Felix Naumann. Fast approximate discovery of inclusion dependencies. In *Datenbanksysteme für Business,*

Technologie und Web (BTW 2017), 17. Fachtagung des GI-Fachbereichs ,,Datenbanken und Informationssysteme" (DBIS), 6.-10. März 2017, Stuttgart, Germany, Proceedings, pages 207–226, 2017.

43. Sebastian Kruse, Thorsten Papenbrock, and Felix Naumann. Scaling out the discovery of inclusion dependencies. In *Datenbanksysteme für Business, Technologie und Web (BTW), 16. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 4.-6.3.2015 in Hamburg, Germany. Proceedings*, pages 445–454, 2015.
44. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
45. Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. *PVLDB*, 3(1):1338–1347, 2010.
46. Zhen Hua Liu, Beda Hammerschmidt, and Doug McMahon. Json data management: Supporting schema-less development in rdbms. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1247–1258, New York, NY, USA, 2014. ACM.
47. Vanessa Lopez, Spyros Kotoulas, Marco Luca Sbodio, Martin Stephenson, Aris Gkoulalas-Divanis, and Pol Mac Aonghusa. Queriocity: A linked data platform for urban information management. In *The Semantic Web - ISWC 2012*, pages 148–163, 2012.
48. Fadi Maali and John Erickson. Data Catalog Vocabulary (DCAT). `http://www.w3.org/TR/vocab-dcat/`, January 2014.
49. Deborah McGuinness, Timothy Lebo, and Satya Sahoo. The PROV Ontology (PROV-O). `http://www.w3.org/TR/prov-o/`, April 2013.
50. Robert Meusel, Petar Petrovski, and Christian Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 277–292, 2014.
51. Robert Meusel, Dominique Ritze, and Heiko Paulheim. Towards more accurate statistical profiling of deployed schema.org microdata. *J. Data and Information Quality*, 8(1):3:1–3:31, 2016.
52. Alistair Miles and Sean Bechhofer. Simple knowledge organization system reference. Recommendation, W3C, August 18 2009.
53. Renée J. Miller, Mauricio A. Hernández, Laura M. Haas, Lingling Yan, C. T. Howard Ho, Ronald Fagin, and Lucian Popa. The clio project: Managing heterogeneity. *SIGMOD Rec.*, 30(1):78–83, March 2001.
54. Johann Mitlöhner, Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Characteristics of open data csv files. In *2nd International Conference on Open and Big Data*, August 2016. invited paper.
55. Varish Mulwad, Tim Finin, and Anupam Joshi. Semantic message passing for generating linked data from tables. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 363–378, 2013.
56. Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.

57. Sebastian Neumaier, Jürgen Umbrich, Josiane Xavier Parreira, and Axel Polleres. Multi-level semantic labelling of numerical values. In *The 15th International Semantic Web Conference*, Kobe, Japan, 2016.
58. Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *J. Data and Information Quality*, 8(1):2:1–2:29, 2016.
59. Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Lifting data portals to the web of data. In *WWW2017 Workshop on Linked Data on the Web (LDOW2017), Perth, Australia, April 3-7, 2017*, 2017.
60. Axel-Cyrille Ngonga Ngomo, Sören Auer, Jens Lehmann, and Amrapali Zaveri. Introduction to linked data and its lifecycle on the web. In *Reasoning Web 2014: Reasoning on the Web in the Big Data Era - 10th International Summer School*.
61. Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *IJMSO*, 3(1):37–52, 2008.
62. Thorsten Papenbrock, Sebastian Kruse, Jorge-Arnulfo Quiané-Ruiz, and Felix Naumann. Divide & conquer-based inclusion dependency discovery. *PVLDB*, 8(7):774–785, 2015.
63. Felipe Pezoa, Juan L. Reutter, Fernando Suárez, Martín Ugarte, and Domagoj Vrgoc. Foundations of JSON schema. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 263–273, 2016.
64. Axel Polleres, Aidan Hogan, Renaud Delbru, and Jürgen Umbrich. RDFS & OWL reasoning for linked data. In Sebastian Rudolph, Georg Gottlob, Ian Horrocks, and Frank van Harmelen, editors, *Reasoning Web. Semantic Technologies for Intelligent Data Access (Reasoning Web 2013)*, volume 8067, pages 91–149. Springer, Mannheim, Germany, July 2013.
65. Rufus Pollock, Jeni Tennison, Gregg Kellogg, and Ivan Herman. Metadata Vocabulary for Tabular Data. `https://www.w3.org/TR/2015/REC-tabular-metadata-20151217/`, December 2015. W3C Recommendation.
66. S. K. Ramnandan, Amol Mittal, Craig A. Knoblock, and Pedro A. Szekely. Assigning semantic labels to data sources. In *ESWC 2015*, pages 403–417, 2015.
67. Y. Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180 (Informational), October 2005.
68. Manu Sporny, Gregg Kellogg, and Markus Lanthaler. JSON-LD 1.0A JSON-based Serialization for Linked Data. `http://www.w3.org/TR/json-ld/`, January 2014.
69. Simon Steyskal and Axel Polleres. Defining expressive access policies for linked data using the ODRL ontology 2.0. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014*, 2014.
70. Mohsen Taheriyan, Craig A. Knoblock, Pedro Szekely, and Jose Luis Ambite. A Scalable Approach to Learn Semantic Models of Structured Sources. In *Proceedings of the 8th IEEE International Conference on Semantic Computing (ICSC 2014)*, 2014.
71. Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From freebase to wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 1419–1428, 2016.
72. The Open Data Charter. G8 open data charter and technical annex, 2013.
73. Petros Venetis, Alon Y. Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. Recovering semantics of tables on the web. *PVLDB*, 4(9):528–538, 2011.

74. Denny Vrandecic and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.
75. Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. Technical report, USA, 1998.
76. Ziqi Zhang. Towards efficient and effective semantic table interpretation. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 487–502, 2014.