Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 66–77

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 14th International Conference on Semantic Systems

The Human Face of the Web of Data: A Cross-sectional Study of Labels

Lucie-Aimée Kaffee, Elena Simperl

ECS, University of Southampton, UK

Abstract

Labels in the web of data are the key element for humans to access the data. We introduce a framework to measure the coverage of information with labels. The framework is based on a set of metrics including completeness, unambiguity, multilinguality, labeled object usage, and monolingual islands. We apply this framework on seven diverse datasets, from the web of data, a collaborative knowledge base, open governmental and GLAM data. We gain an insight into the current state of labels and multilinguality on the web of data. Comparing a set of differently sourced datasets can help data publishers to understand what they can improve and what other ways of collecting and data can be adopted.

Keywords: Linked Data, Web of Data, Labels, Human Accessibility, Multilingual
2010 MSC: 00-01, 99-00

1. Introduction

The web of data is an invaluable resource for humans and computers alike. While its main benefits are often explained in the context of the linked data principles, in many applications making linked data genuinely useful also means attaching natural language representations to URIs, if needed in several languages. There are many examples to illustrate this, from search [1], text generation [2], browsing [3] and visualisation [4] to question answering [5, 6] and ontology modeling [7].

In linked data, resources can be accompanied by human-readable labels, descriptions or comments using a range of pre-defined properties. Additionally, text can be marked with a language tag, such as *@en* for English to support multilingual applications.

Ell et al. have introduced a framework to study the human readability of the web of data [8]. The framework consists of a method to collect different natural language representations of URIs in a linked dataset and a set of metrics to assess different dimensions of human readability: completeness, efficiency of access, unambiguity and multilinguality. They apply the framework on the 2010 edition of the Billion Triple Challenge (BTC) corpus, a representative sample of the web of data at the time and conclude that more labels are needed to encourage uptake of the use of the data in a greater range.

Seven years and many success stories later, we were interested to see how things have progressed and whether there are any noticeable differences among datasets in different domains, which have been created in different contexts. We selected a sample of seven datasets as follows:

Email addresses: kaffee@soton.ac.uk (Lucie-Aimée Kaffee), E.Simperl@soton.ac.uk (Elena Simperl)

10.1016/j.procs.2018.09.007

BTC 2010 The dataset used in [8] was our baseline. We wanted to reproduce the initial findings to calibrate our implementation of the analysis framework. While Ell et al. use a subset of the dataset, we apply our framework to the complete dataset.

BTC 2014 An update of the 2010 dataset used by Ell at al.; a representative cross-section of the Linked Open Data Cloud, crawled on the Web.

Wikidata One of the largest knowledge bases of our times, created and maintained by a distributed community of editors, supported by bots.

Government data (two datasets) The public sector has been one of the greatest promoters of linked data in recent times and many open government datasets are available as linked data. We selected two datasets from two domains from the UK and Taiwan, respectively. Both governments are known for their advanced open data policies, as demonstrated for several years in a row in global studies such as the Open Data Index.[1] Both datasets are published by a central unit within the government and updated regularly.

GLAM data (two datasets) GLAM (Galleries, Libraries, Archives and Museums) have been among the first verticals that not only published substantial amounts of data as linked data, but are also actively using it to provide access to its digital collections [9]. Again, we picked datasets from countries speaking English and other languages, in this case from Switzerland, known for its multilingual fabric. Both datasets are published by the National Libraries in those countries.

In drawing the sample we aimed to select datasets that are diverse with respect to their country of origin, domain, provenance, governance and language (or languages) supported. In this way, we have a more complete picture of the human face of the web of data and of those areas that require further improvement.

We extend the metrics of Ell et al. and create a comprehensive framework, that can be used for any linked data graph to assess its human accessibility in terms of labels.

We found that datasets are widely labeled, however lack of multilinguality. Often, entities are labeled only in one language, even if the dataset contains labels in multiple languages. Another important finding is the lack of labeling of highly reused objects.

The remainder of this paper is organised as follows. We will start in Section 2 by introducing the framework by Ell et al. [8] that we used as starting point for our analysis. We then give details on the seven datasets in our study (Section 3) and the framework developed to analyse these resources (Section 4). We outline the results in Section 5 and discuss their implications and the limitations of the study in Section 6. We frame our contribution in the context of related studies around data quality and multilinguality in Section 7, before concluding with a summary of findings and planned future work in Section 8.

2. Background

Ell et al. introduced a framework to analyse label coverage of linked data resources [8], which sets a baseline for investigating the human readability of the web of data.

The metrics focus on non-information resources (NIR) or *entities* (in the following used interchangeably), which are identified by hash URIs or can be resolved in a 302 or 303 response in the HTML header. NIRs describe things, such as cities or people; classes of things, such as the set of all cities; as well as their properties, such as the number of people in a city or the quality of a city to be the capital of a country. To allow people to engage with linked data effectively, whether as part of a end-user application such as question answering system, or in a technical context, such as when editing a knowledge base, NIRs must have human readable representations. The framework consists of two steps: extraction of labeling properties from the datasource at hand, and a set of metrics. As part of the metrics, NIRs are analysed based on

[1] `https://index.okfn.org/`, accessed on January 8th 2018.

Dataset	# Triples	Year	# LP	Most Used LP
BTC10	3,171,793,030	2010	36	`http://www.w3.org/2000/01/rdf-schema#label`
BTC14	4,090,758,596	2014	36	`http://www.w3.org/2000/01/rdf-schema#label`
WD	2,199,382,887	2017	3	`http://www.w3.org/2000/01/rdf-schema#label`
SchSu	15,347	2015	1	`http://www.w3.org/2000/01/rdf-schema#label`
TaiPS	42,938	2017	3	`http://linked-data.moi.gov.tw/ontology/moi/name`
BNL	4,620,557	2017	8	`http://www.w3.org/2000/01/rdf-schema#label`
SNL	9,900,417	2016	3	`http://purl.org/dc/elements/1.1/title`

Table 1. Dataset statistics, such as size in total number of triples, year last updated, number of labeling properties (LP) used, and most used labeling property

completeness, efficient accessibility, unambiguity, and multilinguality. If a dataset contains multiple graphs, efficient accessibility measures the extent to which entities are accessible in the same graph. As all datasets but BTC do not contain multiple graphs, we excluded efficient accessibility in the following.

The metrics are applied on a sample of the BTC 2010 corpus.

They find in their analysis that only 38.2% of NIRs are labeled. Unambiguity results to 98%. English is the most used language, followed by German and French. They conclude with suggestions, we support in the following: all entities should be labeled, all entities should be labeled multilingual, avoid using other label properties but `rdfs:label`, and do not provide more than one preferred label.

3. Datasets

We work with seven datasets of different natures. An overview can be found in Table 1.

Billion Triple Challenge (BTC10 and BTC14). The Billion Triple Challenge (BTC) contains data crawled from the web. The dataset of the Billion Triple Challenge (BTC) of 2010 [10] was originally used by Ell et al. [8] to analyze the coverage of labels on the web of data. We add the version of 2014 to our analysis. BTC was set up to have one comprehensive collection of data on the web of data. The data is collected using semantic web crawlers, such as SWSE [11], from a big variety of web sources based on a list of seeds[2]. The data is provided as RDF N-Quads[3], in the form: `subject predicate object graph.` where the graph denotes the web source of the statement.

Wikidata (WD). Wikidata is a knowledge base created in 2012 originally to support Wikimedia projects such as Wikipedia. Wikidata is built by a community of volunteers and bots [12] and regarded as a secondary source of information [13]. The knowledge base is inherently multilingual such that each entity has an opaque URI that is not dependent on language information. Hence, editors are encouraged to add data, automatically or manually, in any of the over 400 languages Wikidata supports. An interface is provided that lets contributors work in their native language, attracting a more diverse audience. Compared to the other data sources, this one is the only one that is crowdsourced, and not curated by experts but by a large community of users. The data of Wikidata is provided in the formats JSON, XML and NT. We worked with the NT dump mainly.

Governmental Data. We selected two sets of data published as RDF from two different sources, the British and the Taiwanese open data portals. This gives us an insight in how natural language structured data is published in the context of governmental institutions.

Taiwan Open Data: Public Services Address List (TaiPS) We chose a dataset of a country that is neither English speaking nor in the western hemisphere in order to broaden our insight. According to the Open Data Index of the Open Knowledge Foundation of 2015[4], Taiwan ranked first place in terms of open governmental

[2] `http://km.aifb.kit.edu/projects/btc-2014/`
[3] `https://www.w3.org/TR/n-quads/`
[4] `http://2015.index.okfn.org/place/taiwan/`

Metric	Description	
Labeling Properties	Properties used for labeling	
Natural Language URI	Type of URI to express a NIR	
Completeness	Coverage of entities in terms of labels	$LC = \frac{\lvert S_L \rvert}{\lvert S \rvert}$
Unambiguity	Conflicting labels for one entity	$U = \frac{\lvert S_U \rvert}{\lvert S \rvert}$
Multilinguality	Diversity in terms of languages	$ML = \frac{\sum_{j=1}^{z_l} t_j^l}{\sum_{j=1}^{Z} t_j}$
Monolingual Islands	Entities labeled in more than one language	$MI = \frac{\lvert S_{Multi} \rvert}{\lvert S_{Mono} \rvert}$
Labeled Object Usage	Usage of labeled and unlabeled objects	$LOU = \frac{\sum_j^{\lvert O_l \rvert} n_{o_j}}{\lvert O_l \rvert}$

Table 2. Metrics used

data. The official language of Taiwan is Standard Mandarin Chinese. The chosen dataset is a list of public services[5], such as the *Taipei City Government Disaster Response Center*.

UK Open Data: Schools in Surrey (SchSu) To learn about datasets that are published by an English speaking country, we chose a set about Schools in Surrey[6] from the UK open data portal. It contains information about schools for the community, such as type of school, addresses, and contact information.

GLAM data. GLAM data, similar to governmental data, is published centrally by one organization. However, in this case it is published by cultural institutions, such as libraries. In the case of institutes in the GLAM field, we chose two datasets from comparable sources: The Swiss National Library and the British National Library. Both datasets contain bibliographic data and are used in these institutions.

Swiss National Library (SNL) Given Switzerland has three official languages (German, French and Italian), it is to be expected that data published by the Swiss open data portal would support multiple languages. The Swiss National Library published bibliographic data in 2016[7].

Linked Open British National Bibliography (BNL) Again, we wanted to compare the dataset to one originating in an English speaking country. We expect to gain a benefit to compare two similar sources, to understand whether a standard can be found already. Just as the Swiss library, the Linked Open British National Bibliography publishes bibliographic data in structured format[8].

4. Framework

Our framework consists of two main steps: First, we identify labeling properties for each dataset. Then we apply the metrics described below. We process all datasets, to analyze their natural language URIs, completeness, unambiguity, mulitilinguality, monoglongual islands and labeled object usage. In Table 2, we describe the different metric we used to gain an insight on the datasets presented. To reduce the size of the BTC datasets and make computation lighter, we encode the triples. Each URI is encoded using SHA256, converting them to integer. The code for the analysis can be found at `https://github.com/luciekaffee/metrics-label`.

Properties. In the first step, we compile a list of properties that are used to add human readable labels to NIRs. In [8], this list consists of 36 properties, which have been curated manually based on data from the BTC 2010 corpus, including `rdfs:label`, as well as several other properties in commonly used vocabularies such as FOAF, SKOS and Dublin Core. Most datasets use several properties to attach textual information to NIRs besides the recommended `rdfs:label` [14]. This complicates automatic reuse of this information

[5] `https://data.gov.tw/dataset/8666`

[6] `https://data.gov.uk/dataset/schools2`

[7] `https://opendata.swiss/en/dataset/bibliografische-daten-rdf/resource/38b3d882-12d6-478a-8fbd-7beeccc29046`

[8] `https://data.gov.uk/dataset/the-linked-open-british-national-bibliography/resource/84bdfae5-e8c4-47b0-8e43-49d108fd8440`

– applications need to be aware of the different ways in which the information is expressed [15] and decide which parts to display to the user and how. Based on the list of properties, we then collect the labels and analyses them. To understand how the set of chosen datasets are labeled, we look at the most used properties in a corpus that refer to a string value[9]. From this set of properties, we select the ones used for labeling and description in natural language manually. Wikidata uses three properties, to describe an entity in natural language: label, alias, and description. To satisfy multiple ontologies, these three concepts are covered with multiple ontologies, e.g. `rdfs:label` and `schema:name`. Therefore, in Wikidata there is expected redundancy. However, all labeling properties have the same string as value between the different ontologies. Following [16] we used only one property (e.g. `rdfs:label` for labels) for each of the three categories in Wikidata's case. After collecting the labeling properties, we count the occurrence of each labeling property in the whole dataset.

Natural Language URI. Each entity in the semantic web has a unique ID, which it can be identified with, so called URIs. In their functionality they clearly differ from labels [17]. While labels are a way of humans to interact with the data in natural language, URIs are supposed to be identifier and references to concepts that ideally do not have to change. The authors encourage the usage of opaque URIs, that is language independent identifier[10]. Opaque URIs can contain any form of ID, that is not a word from any natural language, such as a numeric value. They should be independent from the actual content of an entity. The authors argue those will prevent a bias towards the English or any other language and is a better choice for ontologies which will support descriptions of the concepts in multiple languages. Additionally, if names of concepts are amended it is impossible to change a descriptive URI due to conventions, while an opaque URI never has to change. In our metrics, we give Natural Language URIs consideration, by extracting a sample of URIs and manually evaluating whether a dataset uses consistently either opaque or natural language identifier as URI.

Completeness. To improve data accessibility, each NIR or entity in the data should have at least one label. Considering a dataset consisting of triples made of subjects, predicates, and objects, label completeness LC is defined as the ratio of subjects that have at least one label. Let S be the set of all the unique entities and S_L the set of entities that have at least one label, we compute LC as follows:

$$LC = \frac{|S_L|}{|S|} \ , \tag{1}$$

where $|S_L|$ and $|S|$ denote the cardinality of those two sets such that $|S_L| \leq |S|$. The metric takes into account any property identified in the previous step, as we assume that any natural language representation of a resource is useful for human data interaction. The metric does not differentiate between languages. Each label, English or otherwise, with or without a language tag, is considered.

Unambiguity. If a user wants to access an entity, a system has to decide which natural language label should be displayed. This differs from the previous task as we only want to understand how often the same entity has different labels that can not be differentiated as preferred. Therefore, we limit the properties in this tasks to properties used for labeling and excluded properties used for e.g. description. The subset was created manually, by examining the name or, if available, description of each property. We define unambiguity as a resource having only one labeling property per entity, making accessing the label for the entity e.g. for querying simple. Further, each entity $s_j \in S$ in the dataset D should have only one label, therefore each labeling property can not be used more than once per entity.

We evaluate unambiguity for the most used language. Labeling an entity in multiple languages should not be punished by the metric. We define unambuguity U as the portion of entities that have no duplicated language information compared to number of all entities in S. Formally, let S_U be the set of entities have no duplicated language information, we compute:

$$U = \frac{|S_U|}{|S|} \ , \tag{2}$$

where $|S_U|$ is the cardinality of S_U.

[9] In the case of BTC, we used the properties suggested by Ell et al. [8]

[10] This follows also the recommendations of `http://www.w3.org/Provider/Style/URI`

Multilinguality. To be able to cater to various readers of multiple languages, it is necessary to provide information in multiple languages. We measure multilinguality of the dataset in two steps. First, multilinguality of a dataset D is measured by the number of languages the entities cover overall, L_D.

Further, we measure how the languages of the labels are distributed, for each language $l \in D$. Let $T^l = \{t^l_1, t^l_2, \ldots, t^l_{z_l}\}$ be the set of the z_l language-defined triples in language l and let $T = \{t_1, t_2, \ldots, t_Z\}$ be the set of all language-defined triples such that $T^l \subseteq T$, we compute:

$$\frac{\sum_{j=1}^{z_l} t^l_j}{\sum_{j=1}^{Z} t_j} \quad , \tag{3}$$

where Z is the total number of triples in D.

Monolingual islands. We extend the metrics of multilinguality with another aspect important for the language coverage of a dataset: so-called *"monolingual islands"*, which are discussed in the context of multilingual data [18]. They describe the phenomena when data is published in mainly one language and not interlinked to sources that provide information on the same concept in another language. In terms of multilinguality, it is not only important to measure how many languages a dataset covers, but also how well information between those languages is connected. For example, a dataset could focus on one topic in one language and another topic being covered only by a second language. In this scenario, a person not able to understand both languages would only have access to a subset of the content. Therefore, we measure how many entities are available in multiple languages. Let S_{Multi} be the set of all entities available in multiple languages and S_{Mono} the set of all entities labeled only in one language, we compute Monolingual Islands MI:

$$MI = \frac{|S_{Multi}|}{|S_{Mono}|} \tag{4}$$

where $|S_{Multi}|$ is the cardinality of S_{Multi}.

Labeled Object Usage. Labeled object usage measures the reuse of labeled entities. We assume that an entity that is used more often as an object has a higher chance to be seen by a user, e.g. as a result of a query. Therefore, we describe this entity as having a *high visibility*. Entities with a higher visibility should be labeled more, as they are more likely to be seen by a user and have a higher impact on the dataset. We measure how often in average a labeled entity E_{Lab} is used as object in the dataset compared to unlabeled entities E_{UnLab}.

Let O be the set of all the unique entity-type objects in all triples in dataset D. Furtheremore, let O_l and O_u be the set of set of unique labeled and unlabeled objects respectively, such that $|O| = |O_l| + |O_u|$.

For each entity in $o_j \in O_l$, we identify its number of occurrences as object in the triples of D, n_{o_j}, and we compute:

$$\frac{\sum_{j}^{|O_l|} n_{o_j}}{|O_l|} \tag{5}$$

The same is done for the set of unlabeled entities in O_u.

5. Results

5.1. Labeling Properties

After collecting all labeling properties, we measure their occurrences in the datasets. Overall, the labeling property common to most of our investigated datasets and widely used in them, is `rdfs:label`. As can be observed in Figure 1, only SNL does not use `rdfs:label` but `purl:title`, the equivalent in the purl ontology. Given the bigger number of labeling properties overall in BTC, it is the one that shares the most labeling properties between all datasets. Some datasets use labeling properties outside the standard ontologies. For example, TaiPS' most used labeling property is `http://linked-data.moi.gov.tw/ontology/moi/name`, a property introduced by this dataset.

5.2. Natural Language URIs

In our datasets, Wikidata, SNL, and BNL use completely opaque URIs. In the BTC corpora, due to their nature of being from different sources, the URIs make use of both, but in our investigated sample utilize English keywords in the URI. The TaiPS dataset uses unique, opaque identifier for each public service, however, they use the type in English in the URI, such as `http://linked-data.moi.gov.tw/resource/FireAgency/00028` for the *Qidu Branch*. The same pattern of URIs can be observed in the school dataset (SchSu), where the type of resource (e.g. school or address) is displayed as part of the URI. As we show in Section 5.5, the datasets using opaque URIs do not necessarily provide a better cover of languages. While SNL and BNL use opaque URIs, they are not mutlilingual. Additionally, we find that datasets of the same type of publishing confirm to one way of designing URIs.

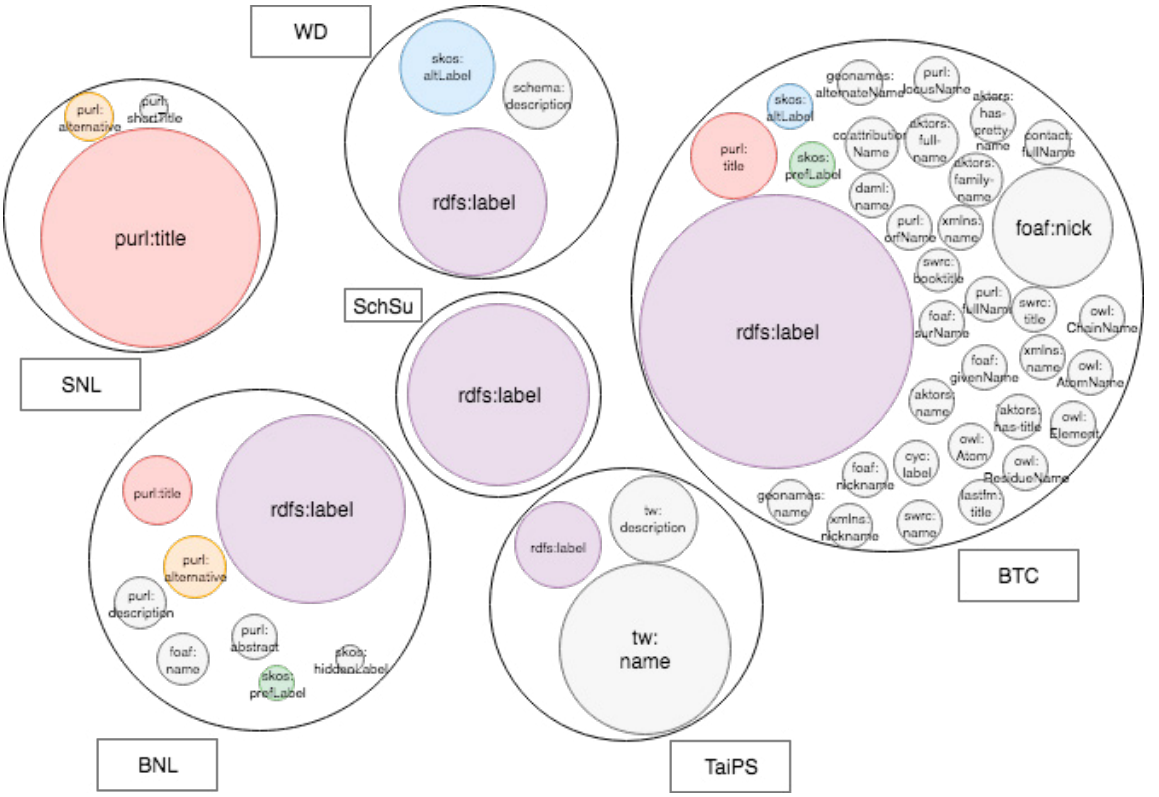

Fig. 1. Labeling properties of the datasets, diameter size based on percentage of usage

5.3. Completeness

Generally, the completeness of labels over all languages is high in most datasets as visible in Figure 2. If there are missing labels there were patterns to detect. A manual investigation shows that especially in the centrally published datasets, there might be labels missing by type of entity. This makes it worthwhile to investigate the label distribution in Open Data further. For example, SchSu is the dataset with the lowest completeness. All entities of type School Site[11] and Postal Address[12] have no label. Wikidata could score a high coverage. This can be attributed to the fact that the whenever a new item or property is created, it has to be connected to at least either one label, description or alias in any language. Comparing BTC10 and BTC14 gives an interesting insight: The new version of the BTC corpus covers fewer entities with labels. We

[11] `http://data.surreycc.gov.uk/def/assets/SchoolSite`
[12] `http://schema.org/PostalAddress`

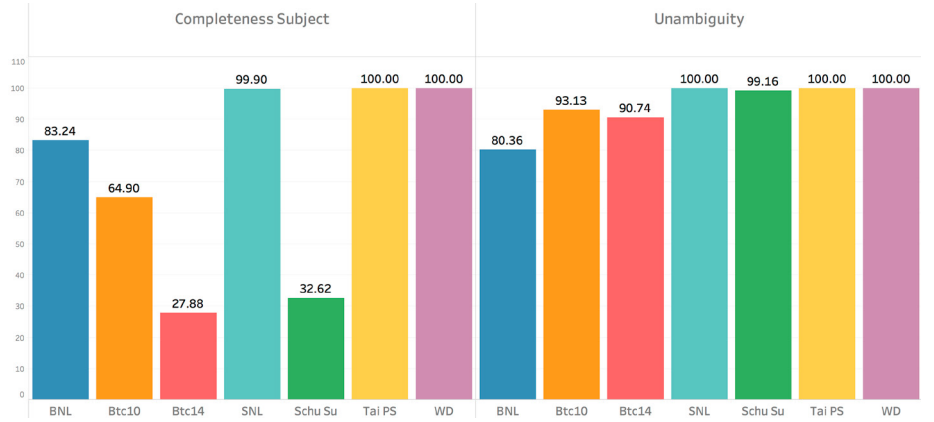

Fig. 2. Results for completeness and unambiguity of subjects over all datasets

	BTC10	BTC14	WD	SchuSu	TaiPS	BNL	SNL
# Languages	55	183	424	NA	2	1	NA

Table 3. Number of languages in the different datasets

assume that the decrease of coverage is due to the increase of the dataset size. While the dataset increased by 28.97% in terms of triples, and 61.2% increase of unique subjects, the number of unique subjects that are labeled decreased by 30.75%. Furthermore, datasets of the same publishing type show different results: SchSu compared to TaiPS has a worse coverage, similar to BNL compared to SNL.

5.4. Unambiguity

While all datasets are relatively unambigious, the datasets using fewer labeling properties could score better results. All unambigious datasets (Wikidatata, TaiPS, and SNL) have a low number of labeling properties (1, 3, and 3). In contrast, BTC14 and BTC10 uses 36 properties and 9.3% and 6.9% respectively of entities are ambiguous. An unambigous ontology increases unambiguity of entity labels and supports machines and humans to identify the data correctly.

5.5. Multilinguality

Multilinguality in structured data is important, as it makes the data accessible to a wider audience and can be reused for any community. Even resources published by institutions from multilingual countries, such as SNL, do not use language tags at all. While the data of SNL uses labels, none of the labels are marked with a language tag, as would be the standard. The URI `http://purl.org/dc/terms/language` is used to indicate languages of an entity. As this should be used to indicate languages of the concept described by the entity, we exclude it from the following measurements that focus on languages of labels. Using only identifier that state the entities language, but not the labels language, limits accessibility and reuse.

The datasets cover a varying amount of languages (from 1 to 424 languages, Table 3). English is the main language of all datasets but TaiPS. TaiPS is an interesting case, as it indicates that non-English data publishers can be willing to publish translated datasets. A dataset published in a non-Latin script language can become part of the multilingual semantic web, providing translations to English.

Table 3 shows the distribution of languages in the datasets. Wikidata is the datasets with the biggest variety- more languages are covered than by any of the other datasets, and English has a smaller share of content (11%). Even datasets that cover a similar high number of languages (BTC10 and BTC14) have a high share of English information (44.7% and 66.4%). A better distribution should be encouraged to serve more communities.

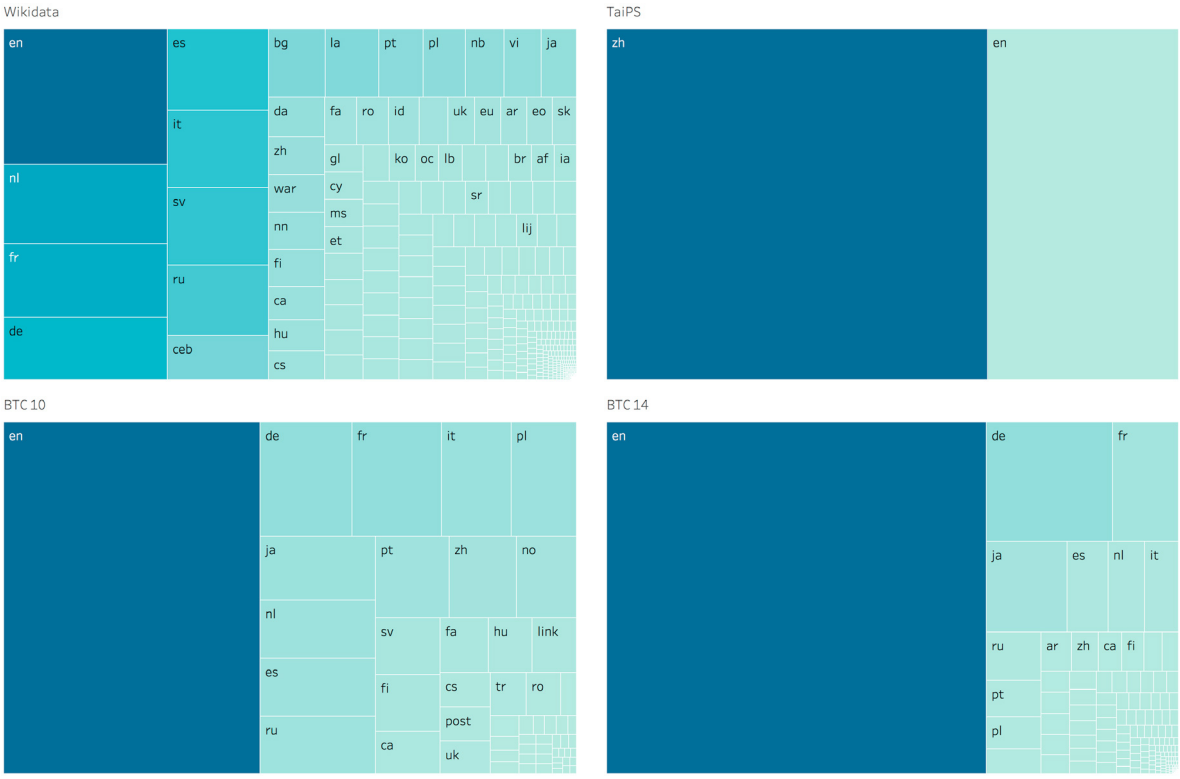

Fig. 3. Language distribution in the four multilingual datasets

	BTC10	BTC14	WD	SchSu	TaiPS	BNL	SNL
1	0.99	0.93	0.58	–	0.5	1	–
2	0.004	0.043	0.17	–	0.5	–	–
2-5	0.006	0.05	0.27	–	–	–	–
5-10	0.002	0.008	0.09	–	–	–	–
>10	0.0008	0.005	0.08	–	–	–	–

Table 4. Share of entities having labels in multiple (1, 2, 2-5, 5-10, over 10) languages

5.6. Monolingual Island

In order to be truly multilingual, each entity in a dataset should be available in multiple languages. We find that in both BTC10 and BTC14 the vast majority (99% and 93% of all labeled entities respectively) are only available in one language, as visible in Table 4.Wikidata however shows a very different image: only over half of the entities are available in one language. A big share of entities (27%) are available in between two to five languages. In the Taiwanese dataset, TaiPS, 50% of the dataset is labeled only in Chinese. Given the dataset is from a primarily Chinese speaking country, it is easy to follow why Chinese is the dominant language in this dataset. However, all entities that are labeled in English (50%), are labeled in Chinese as well. This measurement can not be applied to SchuSu and SNL as both are not multilingual.

5.7. Labeled Object Usage

Frequently used objects should be labeled due to their higher visibility. Therefore, we investigate how often labeled and unlabeled objects are used on average. Contrary to our intuition, overall entities that are labeled are not more likely to be used as objects. Labeled entities are less used as objects in most datasets as visible in Table 5. This effect applies foremost to TaiPS. The average usage for unlabeled object seems

	BTC10	BTC14	WD	SchSu	TaiPS	BNL	SNL
Labeled	11.5	18.2	NA	2.1	1.0	2.9	3.7
Unlabeled	13	7.3	NA	3.1	1071.2	8.9	20.9

Table 5. Average usage of labeled and unlabeled objects

extraordinarily high in this dataset, this is because there are only five unlabeled objects[13]. These objects can be described as classes, that indicate of which type an entity is. Classes are by nature highly reused, as each entity should be assigned as part of a class. In the case of unlabeled classes, we can see the importance of labeling high reused entities in an extreme case: almost all entities will use at least one of those classes, but their classes are not accessible to humans. In Wikidata every entity (including all entities used as objects) is labeled. The only dataset that has a higher reuse of labeled objects is BTC14. The information collected for BTC14 is based on data extracted from the web of data, and is a collection of multiple sources. It is promising higher reused items are likely to be labeled in such a dataset constructed from distributed sources.

6. Discussion

We analysed seven datasets of different sources with a framework that combines different metrics to assess the human accesibility of a dataset in terms of labels. Following our results, we draw recommendations for data publishers. It gives us an insight on what different datasets have to improve. To compare the different data publishing types in detail, a bigger variety of datasets for each publishing type should be considered and evaluated based on our framework. Overall, the community maintained dataset (Wikidata) is the most diverse and comprehensive dataset in terms of labels. Even between datasets that are published in the same way, such as governmental data, we find no coherent style of publishing natural language data. This should be improved in future work by publishing the data complying standards with the help of e.g. clear guidelines for labeling.

All entities should be labeled. We emphasize on the fact, that all entities should be labeled. This is not currently the case. While datasets that are centrally published can be expected to follow standards easier, even the investigared centrally published datasets do not follow this suggestion. Wikidata enforces labeling by enforcing it with publication of a new entity. Such a constraint seems to support a more better coverage in labeling of entities overall. A high coverage of labels is particularly important when the entity is heavily reused.

Labeling properties should be coherent and limited in number. A limited number of labeling properties makes it easier to differentiate which is the preferred label for an entity. Even if the property is not standardized, it reduces ambiguity. We see that the more variance we have in the data, such as in BTC, the more labeling properties will be introduced. Using a standard property in labeling makes the data more accessible. While the current state is not ideal, it is promising that the labeling properties overlap between all datasets to some extent. However, to be able to reuse data of multiple knowledge bases, a mapping between labeling properties is still needed. Encouraging for future work is the fact that rdfs:label is still one of the most used labeling properties as in Ell et al.'s study. A smaller amount of labeling properties is already visible in expert maintained datasets.

More languages does not mean better coverage. Multilinguality allows different communities to access the same datasets. Particularly community maintained datasets, such as Wikidata, could score high here. Wikidata's community translates labels for one entity in different languages [16], and can therefore draw from

[13] Unlabeled objects in TaiPS are `<http://linked-data.moi.gov.tw/ontology/moi/FireAgency>`, `<http://linked-data.moi.gov.tw/ontology/moi/Address>`, `<http://linked-data.moi.gov.tw/ontology/moi/HouseholdRegistration>`, `<http://linked-data.moi.gov.tw/ontology/moi/PoliceAgency>`, and `<http://linked-data.moi.gov.tw/ontology/moi/ImmigrationAgency>`

community knowledge about different languages. This way of a crowdsourced translation can be applicable to different data sources. Additionally, datasets published in non-English countries can be multilingual. This indicates an incentive to make the information accessible in multiple languages that should be investigated further. However, particular care should be taken to not only cover many languages, but that each entity is actually translated to multiple languages. Otherwise, a knowledge exchange independent of language is not possible.

7. Related Work

Labels are fundamental to make data accessible. They enable humans to understand the data they work with. To gain an understanding of how well labels are supported, and what to improve on the current state, we developed a framework, based on the metrics suggested by Ell et al. [8]. While Ell et al. test their metrics on only one dataset to gain an insight, we extend their research by evaluating a more recent, larger set of diverse data sources to gain an understanding of the state of human accessible information in the form of labels. Manaf et al. [19] survey OWL ontologies towards their label usage. They focus on the frequency of labels and meaningful labels, similar to our completeness metric. They conclude that most ontologies do not use labels. Similar to previous work, Färber et al. [20] compare five knowledge bases' quality. Their metrics, *Ease of Understanding* and *Labels in multiple languages*, investigates the use of labels in those knowledge bases. They show that the investigated knowledge bases, similar to our results, have a high completeness in terms of labels, and are mainly English. Wikidata is the most diverse in terms of multilinguality.

An important aspect in regards to labels is multilinguality [21]. In [16], we analyzed Wikidata in regards to its multilingual content. Garcia et al. [18] suggest while the semantic web can be a resource of multilinguality, there is a lack of services provided to support a fully multilingual web still. As they suggest in their work, most content is still mainly monolingual. One of the important aspects to shape the future of the multilingual web is to set standardized guidelines to follow for resources on the semantic web. Clear guidelines help data publishers to consider a multilingual layout in the developing of their KB as suggested by Gómez-Pérez et al. [22]. The authors give an insight on multilingual data on the web and suggest a framework to contribute to more multilingual data.

8. Conclusion

We compare seven recent datasets to find the state of labels and multilinguality in the web of data. We show that datasets of different sources have different advantages and disadvantages in terms of label coverage. While entities are generally widely labeled, most datasets are not multilingual. Even if datasets cover mutliple languages, entities are likely to be labeled only in one language, which limits access. The widest used labeling property is `rdfs:label`. This is promising, as it supports easy automated access to the labels and decreases the need for ontology alignment between different datasets and their labeling properties. Additionally, a limited number of labeling properties reduces ambiguity. Furthermore, special care should be given to widely used entities, as they are currently more likely to be unlabeled.

It is often needed to combine different data sources to collect all information needed for a task. A combination of different sources of linked data might improve the accessibility for humans, which is a direction for future work. However, it is still a long way for the data to become easily accssible for humans.

Acknowledgements

This research is supported by funding received from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN).

References

[1] Gong Cheng and Yuzhong Qu. Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. *Int. J. Semantic Web Inf. Syst.*, 5(3):49–70, 2009. doi: 10.4018/jswis.2009081903.

[2] Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon S. Hare, and Elena Simperl. Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, pages 319–334, 2018. doi: 10.1007/978-3-319-93417-4_21. URL `https://doi.org/10.1007/978-3-319-93417-4\_21`.

[3] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd international semantic web user interaction workshop*, volume 2006, page 159. Athens, Georgia, 2006.

[4] Jirí Helmich, Jakub Klímek, and Martin Necaský. Visualizing RDF Data Cubes Using the Linked Data Visualization Model. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 368–373, 2014. doi: 10.1007/978-3-319-11955-7_50. URL `https://doi.org/10.1007/978-3-319-11955-7_50`.

[5] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, pages 1–41, 2017.

[6] Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Survey on challenges of Question Answering in the Semantic Web. *Semantic Web*, 8(6):895–920, 2017. doi: 10.3233/SW-160247.

[7] Silvio Peroni, David M. Shotton, and Fabio Vitali. Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation. *Int. J. Semantic Web Inf. Syst.*, 9(1):21–44, 2013. doi: 10.4018/jswis.2013010102. URL `https://doi.org/10.4018/jswis.2013010102`.

[8] Basil Ell, Denny Vrandecic, and Elena Paslaru Bontas Simperl. Labels in the Web of Data. In *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, pages 162–176, 2011. doi: 10.1007/978-3-642-25073-6_11. URL `https://doi.org/10.1007/978-3-642-25073-6_11`.

[9] Pedro A. Szekely, Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, and Georgina Goodlander. Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, pages 593–607, 2013. doi: 10.1007/978-3-642-38288-8_40.

[10] Andreas Harth. Billion Triples Challenge data set. Downloaded from http://km.aifb.kit.edu/projects/btc-2010/, 2010.

[11] 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. doi: 10.1016/j.websem.2011.06.004. URL `https://doi.org/10.1016/j.websem.2011.06.004`.

[12] Alessandro Piscopo, Chris Phethean, and Elena Simperl. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I*, pages 305–322, 2017. doi: 10.1007/978-3-319-67217-5_19. URL `https://doi.org/10.1007/978-3-319-67217-5_19`.

[13] Alessandro Piscopo, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl. Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, pages 542–558, 2017. doi: 10.1007/978-3-319-68288-4_32. URL `https://doi.org/10.1007/978-3-319-68288-4_32`.

[14] Dan Brickley and Ramanathan V Guha. RDF vocabulary description language 1.0: RDF schema. 2004.

[15] Muhammad Saleem, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel-Cyrille Ngonga Ngomo. A fine-grained evaluation of SPARQL endpoint federation systems. *Semantic Web*, 7(5):493–518, 2016. doi: 10.3233/SW-150186. URL `https://doi.org/10.3233/SW-150186`.

[16] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration, OpenSym 2017, Galway, Ireland, August 23-25, 2017*, pages 14:1–14:5, 2017. doi: 10.1145/3125433.3125465. URL `http://doi.acm.org/10.1145/3125433.3125465`.

[17] Elena Montiel-Ponsoda, Daniel Vila-Suero, Boris Villazón-Terrazas, Gordon Dunsire, Elena Escolano Rodriguez, and Asunción Gómez-Pérez. Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications, DC 2011, The Hague, The Netherlands, September 21-23, 2011*, pages 105–115, 2011. URL `http://dcpapers.dublincore.org/pubs/article/view/3626`.

[18] Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John P. McCrae. Challenges for the multilingual Web of Data. *J. Web Sem.*, 11:63–71, 2012. doi: 10.1016/j.websem.2011.09.001.

[19] Nor Azlinayati Abdul Manaf, Sean Bechhofer, and Robert Stevens. A Survey of Identifiers and Labels in OWL Ontologies. In *Proceedings of the 7th International Workshop on OWL: Experiences and Directions (OWLED 2010), San Francisco, California, USA, June 21-22, 2010*, 2010.

[20] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129, 2018. doi: 10.3233/SW-170275.

[21] Paul Buitelaar and Philipp Cimiano, editors. *Towards the Multilingual Semantic Web, Principles, Methods and Applications*. Springer, 2014. ISBN 978-3-662-43584-7. doi: 10.1007/978-3-662-43585-4.

[22] Asunción Gómez-Pérez, Daniel Vila-Suero, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado de Cea. Guidelines for Multilingual Linked Data. In *3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, Madrid, Spain, June 12-14, 2013*, page 3, 2013. doi: 10.1145/2479787.2479867.