Procedia Computer Science 00 (2018) 1–11

Procedia Computer Science

www.elsevier.com/locate/procedia

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 Semantic Web are the key element for humans to access the data. In the following, we analyze six 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. We analyze the datasets based on a set of metrics including completeness, unambiguity, multilinguality, labeled object usage, and monolingual islands. Comparing a set of differently soured datasets can help data publishers to understand what they can improve and what other ways of collecting and data can be adopted. Overall, the centrally published datasets are more comprehensive, however also in lack of data comprehensive. The community maintained dataset can show the best human accessibility, based on its constraints and its dedicated community. ...

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], browsing [2] and visualisation [3] to question answering [4, 5] and ontology modeling [6].

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. Figure 1 shows an example from Wikidata. The Wikidata triple stating that Berlin is the capital of Germany consists of two items `Q64` and `Q183` connected by property `P1376`. `rdfs:label` is used to attach labels in different languages to each of the three parts of the triple. This makes the Wikidata statement easier to understand by people and helps multilingual applications select the correct label to display for each audience.

☆Fully documented templates are available in the elsarticle package on http://www.ctan.org/tex-archive/macros/latex/contrib/elsarticleCTAN.

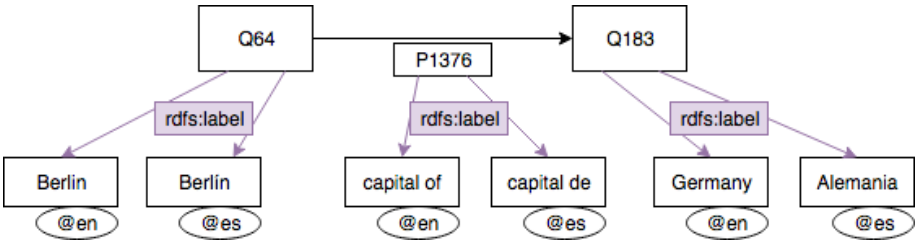

Fig. 1. Example of a labeled triple from Wikidata

Ell et al. have introduced a framework to study the human readability of the Web of data [7]. The framework consists of a method to collect different natural language representations of URIs in a linked data set 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 ...

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:

BTC 2010 The dataset used in [7] was our baseline. We wanted to reproduce the initial findings to calibrate our implementation of the analysis framework. While Ell et al. [7] use a subset of the dataset, we apply them 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 []. Again, we picked datasets in 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 ARE THEY ENRICHED BY THE COMMUNITY, WHY DID WE PICK THIS DOMAIN?

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 can have a more complete picture of the human face of the Web of data and of areas that require further improvement.

ADD MAIN FINDINGS HERE

The remainder of this paper is organised as follows. We will start in Section 2 by introducing the framework by Ell et al. that we used as a starting point for our analysis. We then give details on the seven datasets in our study and the data science methods we implemented to collect, clean and analyse the data

[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

(Section **??**). We outline the results in Section **??** and discuss their implications and the limitations of the study in Section **??**. We frame our contribution in the context of related studies around data quality and multilinguality in Section **??** before concluding with a summary of findings and planned future work in Section **??**.

2. Background

Ell et al. have introduced a framework to analyse label coverage of linked data resources [7], which sets a baseline for investigating the human readability of the web of data.

The framework focuses on non-information resources (NIR), 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, which are discussed in the following.

2.1. Properties

In the first step, the user must compile a list of properties that are used to add human readable labels to NIRs. In [7], 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` [8]. This makes it difficult to use this information automatically - applications need to be aware of the different ways in which the information is expressed [9] and decide which parts to display to the user and how. Based on the list of properties, the user then collects the labels and analyses them.

2.2. Analysis

In the second step, the user computers several metrics:

Completeness. To improve data accessibility, each NIR or entity - in the following we will use the terms interchangeably - in the data should have at least one label. In a general-purpose knowledge base such as Wikidata this requirement goes much further: ideally, labels should be available in each relevant language.

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, where (S_L) denotes the labelled subjects and S all subjects in the triples in the dataset:

Metric	Description
Natural Language URI	Identifying which type of URI are used
Labeling Properties	Identifying properties used for labeling
URI	Type of URI to express a NIR
Completeness	Coverage of entities in terms of labels
Unambiguity	Conflicting labels for one entity
Multilinguality	Diversity in terms of languages
Monolingual Islands	Entities labeled in more than one language
Labeled Object Usage	Usage of labeled and unlabeled objects

Table 2. Metrics used

$$LC = \frac{\sum S_L}{\sum S} \tag{1}$$

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 could not be differentiated as preferred. Therefore, we limit the properties in this tasks to the known properties used for labeling and removed properties used for e.g. description. We checked if entities would use the same property twice or have two conflicting labels. We evaluate unambiguity for the most used language, as using the same property in different languages would be encouraged. We define it as the share of entities that have no duplicated language information in ratio to all entities. Wikidata uses three main labeling properties, to describe an entity in natural language: label, alias, and description. To satisfy multiple ontologies, e.g. labels are described with multiple URIS. Therefore, in Wikidata there is expected redundancy, however, all labeling properties have the same string as value between the different ontologies. Therefore, this redundancy does not indicate an increase in ambiguity. Following [10] we used only one property (e.g. `rdfs:label` for labeling) for each of the three categories in Wikidata's case.

Multilinguality. To be able to cater to various readers of multiple languages, it is needed to be able to access information in multiple languages. Therefore, we measure the multilinguality by the number of languages a resource covers. Additionally we evaluate, which other languages are used and how they are distributed.

3. Data And Methods

3.1. Methods

To understand which properties are used for labeling, we looked at the most used properties in a corpus that refer to a string value[2] and selected the ones used for labeling and description in natural language manually. Following the metrics described earlier, we processed all datasets, to identify their size, completeness in terms of label coverage, unambiguity, and multilinguality. We exclude Ell et al.'s efficient accessibility as it refers to datasets that contain multiple graphs, which is not the case for our datasets beside BTC. To reduce the size of the datasets and make computation easier, we encoded the triples. Each URI is encoded using SHA256, converting them to integer. To calculate completeness and efficient accessibility, we counted number of subject that are labeled. To identify unambiguity, we identified triples with the same language and limited the amount of properties to the ones clearly used for labeling, excluding properties used for

[2]In the case of BTC, we used the properties suggested by Ell et al. [7]

descriptions or aliases. Multilinguality was processed by counting the languages used, by identifying the language tags. From there we identified the number of usage of each language tag and calculated how the languages relate to each other in terms of content size. In Table 3, we describe the different metric we used to gain an insight on the datasets presented. We extended the metrics with the following two measurments to gain a broader insight into the labeling and multilinguality of the datasets.

Labeled Object Usage. Additionally to the metrics described in Section 2, we analyze the reuse of labeled entities. An entity that is used more often as an object has a higher visibility and should therefore be more likely be labeled. For the access of information, the labeling of such high-visible entities is more important than others. We measure how often in average a labeled entity is used as object in the dataset compared to unlabeled entities.

Monolingual islands. We extend the metrics of multilinguality with another aspect important for the language coverage of a dataset: So called *monolingual islands* are discussed in the context of multilingual data [11]. 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. This could occur especially in user-contributed knowledge bases, where a user with knowledge on one topic contributes only in their language, without the translation needed for broader access. Therefore, we measure how many entities are available in multiple languages compared to entities available in a single language.

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 [12]. 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[3]. 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 be 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 manually evaluating whether a dataset uses consistently either opaque or natural language identifier as URI.

4. Results

We investigated seven datasets from different sources towards their label coverage and multilinguality. In the following we present the results.

4.1. 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 generally 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.

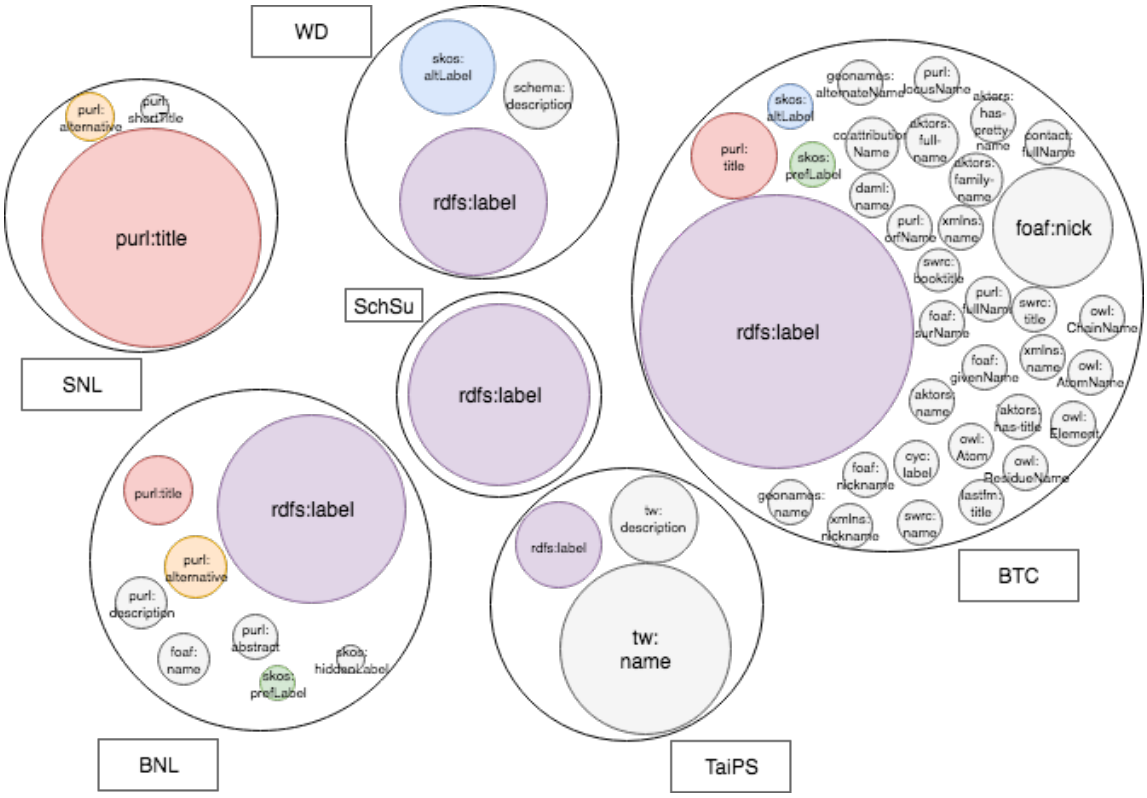

Fig. 2. Labeling properties of the datasets, size based on percentage of usage

4.2. Labeling Properties

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 2, 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.

4.3. Completeness

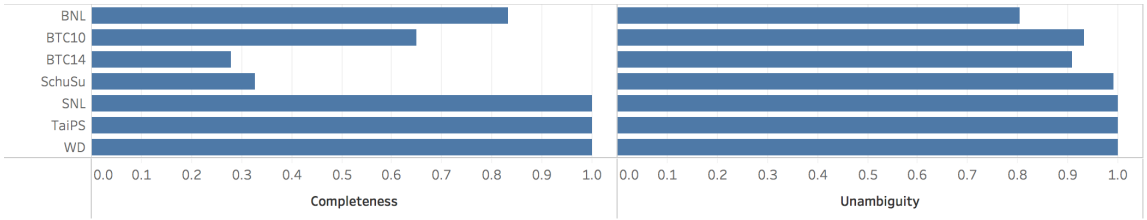

Fig. 3. Results for completeness and unambiguity of subjects over all datasets

Generally, the completeness of labels over all languages is high. While there are missing labels, e.g. in the SchSu there were patterns to detect. We found that especially in the centrally published datasets, there

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

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

Table 3. Number of languages in the different datasets

might be labels missing by type of entity. This makes it worthwhile to investigate the label distribution in Open Data further. For example, in the SchSu dataset, all entities of type School Site[4] and Postal Address[5] 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.

4.4. Unambiguity

While all datasets are relatively unambigious, the datasets using less labeling properties could score better results. Less messy data supports a better access to it and makes it easier for humans and machines to find concepts they are looking for.

4.5. Multilinguality

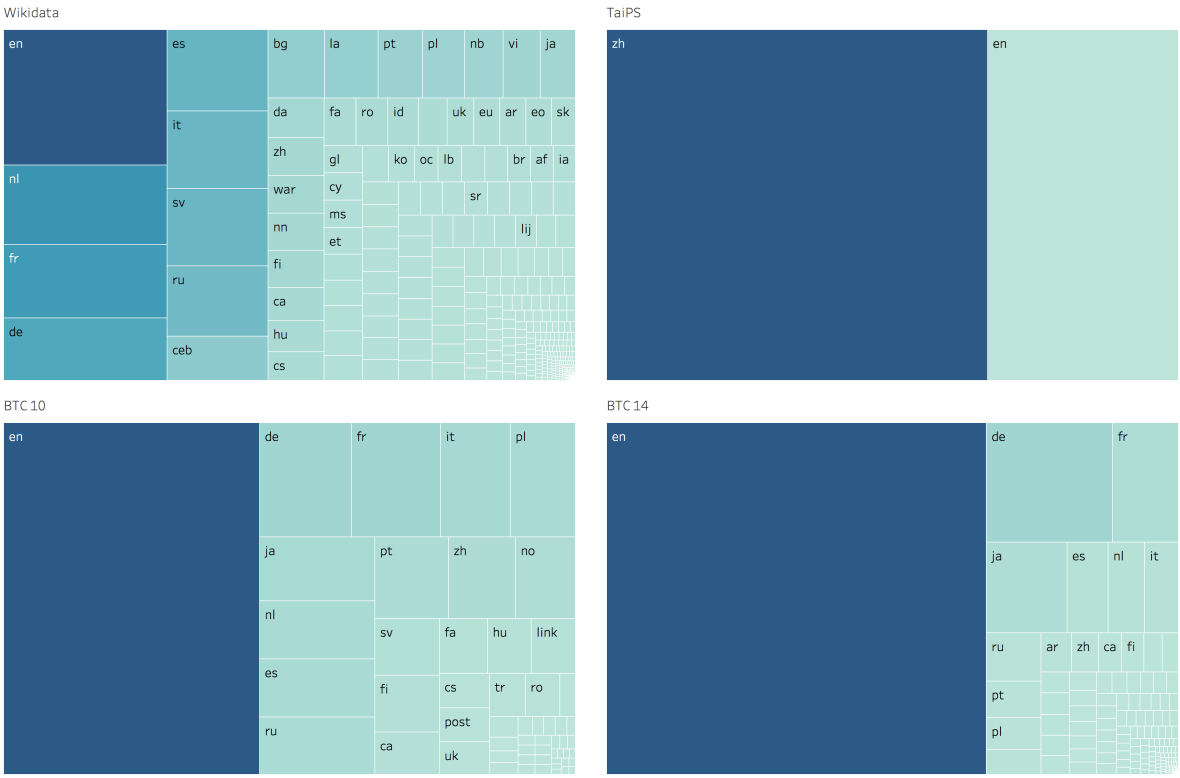

Fig. 4. Language distribution in the four multilingual datasets

Multilinguality in structured data is important, as it makes the data accessible to a wider audience and can be reused for any community. We could detect that even multilingual resources, such as SNL, do not

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

	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

	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

use language tags at all, which needs to be a first step for reuse. This limits the accessibility. English is the main language of all datasets investigated but TaiPS. The multilinguality of the TaiPS dataset is promising, as it suggests that the lack of non-English data can be overcome by the publication of more governmental data of non-English speaking countries. Compared to its British equivalent, the SchuSu dataset, it shows interestingly, that a country with non-Latin script language becomes part of the multilingual semantic web, providing translations in English. From the multilingual datasets, Wikidata is the datasets with the biggest variety- more languages are covered than by any of the other datasets, and the small number of share of English suggests that other languages are better covered, supporting the findings of [10].

4.5.1. Monolingual Island

To understand how actual multilingual the dataset is, we wanted to check whether entities are available in multiple languages, or they are mainly in a single language. 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. A very low number of entities are available in a variety of languages. 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. The Taiwanese dataset, TaiPS, provides a comprehensive coverage of the languages provided. 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, are labeled in Chinese as well, therefore there is no divide between the languages. 50% of the dataset is labeled only in Chinese.

4.6. Labeled Object Usage

To understand, whether labeling of objects has an impact on their usage, and to find whether highly used objects are labeled, we looked at how often labeled and unlabeled objects are used in average. Generally, we can not see a higher usage of labeled entities as objects. To the contrary, labeled entities are less used as objects in most datasets as visible in Table 5. Most prominently is TaiPS. The average usage for unlabeled object seems extraordinarily high in this dataset. That's because there are only five unlabeled objects[6]. These entities can be described as classes, that indicate of which type an entity is and therefore are highly reused. The only dataset that has a higher reuse of labeled objects is BTC14. As this data is naturally grown and extracted from online sources, it indicates that a variety of datasets actually take more care of

[6]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>

the labeling of widely used objects, which is promising. We can see here the importance of labeling high reused entities in an extreme case: almost all entities will use at least one of those classes, but they are not accessible to humans. In Wikidata every entity (including all entities used as objects) is labeled, therefore it is not comparable with the other datasets in this case.

5. Discussion

RECOMMENDATIONS FOR DATA PUBLISHERS BASED ON RESULTS

1. all entities should be labeled
2. generally: more coherent usage of labeling properties for the web at large
3. centrally published datasets: more translation, clear marking of language codes, more languages covered
4. centrally published datasets: special care for labeling of highly reused entities in the own dataset
5. conclusion: maybe learn from differently published dataset, more interlinking between dataset (see lod cloud) will provide better results, make use of existing resources.

We compared seven datasets of different sources, based on the metrics introduced by Ell et al. [7]. Based on our results, we can draw recommendations for publisher of data:
Looking at different ways to publish data, we can say overall, the centrally published datasets are more coherent than the one scraped from different web sources.

All entities should be labeled. We emphasize on the fact, that all entities should be labeled. This is not currently the case, even for centrally published datasets. A constraints in the way new entities are published, such as in Wikidata, seems to support a more coherent labeling of entities overall. An emphasize by these data publishers should be to label every part of the dataset. This is particularly visible in SchuSu. Similarly, each entity, even if not used as a subject in statements, should be labeled. This is particularly important, when it is heavily reused. There is improvement to be seen in the open data datasets. We can see that in the real world, such as the BTC corpora, this difference in usage of labeling becomes more apparent: more reuse leads to more labeling and vice versa.

Labeling properties should be limited in number and coherent. A limited amount 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 can see that the more variance we have in the data, such as in BTC, the more labeling properties will be introduced. While the current state is not ideal, it is promising that the labeling properties overlap between all datasets to some extend. However, to satisfy the option of working with multiple knowledge bases, a mapping is still needed. Encouraging for future work is the fact that the labeling properties stayed consistent the same– rdfs:label is still one of the most used labeling properties. A smaller amount of labeling properties is already visible in expert maintained datasets.

More Languages does not mean better coverage. Multilinguality is an important for access of different communities to the same datasets. Particularly community maintained datasets, such as Wikidata, could score high here. It might be considerable to learn from the community translation efforts. Additionally, datasets published in non-English countries can be multilingual, which shows it is possible for datapublishers to add more languages to their dataset. However, particular care should be taken to not only cover many languages, but that one entity is actually translated to multiple languages, otherwise a knowledge exchange independent of language is not possible. There is way of improvement on this.

6. Related Work

Multilingual information on the semantic web has been a topic of increasing popularity. [13] Garcia et al. [11] 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. The authors of [14] give an insight on multilingual data on the web and suggest a framework to contribute to more multilingual data. Multiple knowledge bases such as DBpedia [15], YAGO [16], and MENTA [17] have extended their approaches of extracting data from Wikipedia to multiple languages. Completeness of knowledge bases in the semantic web is in important topic to understand, how to improve them, and what kind of information is missing. The authors of [18] gain an insight of the completeness by learning rules for what makes entities complete. While they focus on general statements, Ell et al. [7] develop a set of metrics, that focus on language information in the web of data. While they 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 language information on the semantic web. Similarly, in [10], we analyzed Wikidata in regards to its multilingual content.

7. Conclusion

We compared six recent datasets to find the state of labels and multilinguality in the web of data. We could find that there is still a big lack of labels on the web, especially when it comes to non-English content. This lack of multilingual content can be observed between all datasets, independent of their source. It reflects the maldistribution of languages in content online overall. In future work it is to investigate whether a combination of different sources of linked data can improve the coverage of languages.

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. URL `https://doi.org/10.4018/jswis.2009081903`.

[2] 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.

[3] 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`.

[4] 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.

[5] 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. URL `https://doi.org/10.3233/SW-160247`.

[6] 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`.

[7] 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`.

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

[9] 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`.

[10] 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`.

[11] 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. URL https://doi.org/10.1016/j.websem.2011.09.001.

[12] 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.

[13] 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. URL https://doi.org/10.1007/978-3-662-43585-4.

[14] 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. URL http://doi.acm.org/10.1145/2479787.2479867.

[15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi: 10.3233/SW-140134. URL https://doi.org/10.3233/SW-140134.

[16] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015. URL http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf.

[17] Gerard de Melo and Gerhard Weikum. MENTA: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1099–1108, 2010. doi: 10.1145/1871437.1871577. URL http://doi.acm.org/10.1145/1871437.1871577.

[18] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. Predicting completeness in knowledge bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, pages 375–383, 2017. URL http://dl.acm.org/citation.cfm?id=3018739.