

# Challenges of Linking Organizational Information in Open Government Data to Knowledge Graphs

Jan Portisch<sup>1</sup>[0000-0001-5420-0663], Omailma Fallatah<sup>2</sup>[0000-0002-5466-9119],  
Sebastian Neumaier<sup>3</sup>[0000-0002-9804-4882], Mohamad Yaser  
Jaradeh<sup>5</sup>[0000-0001-8777-2780], and Axel Polleres<sup>3,4</sup>[0000-0001-5670-1146]

<sup>1</sup> Data and Web Science Group, University of Mannheim, Germany  
`jan@informatik.uni-mannheim.de`

<sup>2</sup> Information School, The University of Sheffield, Sheffield, UK  
`oafallatah1@sheffield.ac.uk`

<sup>3</sup> Vienna University of Economics and Business, Austria  
`{sebastian.neumaier, axel.polleres}@wu.ac.at`

<sup>4</sup> Complexity Science Hub Vienna, Austria

<sup>5</sup> L3S Research Center, Leibniz University Hannover, Germany  
`jaradeh@l3s.de`

**Abstract.** Open Government Data (OGD) is being published by various public administration organizations around the globe. Within the metadata of OGD data catalogs, the publishing organizations (1) are not uniquely and unambiguously identifiable and, even worse, (2) change over time, by public administration units being merged or restructured. In order to enable fine-grained analyzes or searches on Open Government Data on the level of publishing organizations, linking those from OGD portals to publicly available knowledge graphs (KGs) such as *Wikidata* and *DBpedia* seems like an obvious solution. Still, as we show in this position paper, organization linking faces significant challenges, both in terms of available (portal) metadata and KGs in terms of data quality and completeness. We herein specifically highlight five main challenges, namely regarding (1) temporal changes in organizations and in the portal metadata, (2) lack of a base ontology for describing organizational structures and changes in public knowledge graphs, (3) metadata and KG data quality, (4) multilinguality, and (5) disambiguating public sector organizations. Based on available OGD portal metadata from the *Open Data Portal Watch*, we provide an in-depth analysis of these issues, make suggestions for concrete starting points on how to tackle them along with a call to the community to jointly work on these open challenges.

**Keywords:** Open Data · Dataset Evolution · Entity Linking · Knowledge Graphs · Knowledge Graph Evolution.

## 1 Introduction

Open Data from public administrations, also called *Open Government Data* (OGD), provides a rich source of structured data that has become a key com-

ponent of an evolving Web of Data. The key factors for the success of OGD initiatives are on the one hand the incentives for publishing organizations to demonstrate transparency or compliance to regulations, but on the other hand also the availability of agreed standards and best practices for publishing OGD: de facto publishing standards for metadata on OGD portals such as *DCAT* or, more recently, Schema.org’s dataset vocabulary [3], as well as widely used open publishing software frameworks such as *CKAN* or *Socrata*, provide technical means to publish structured data along with descriptive metadata. There are over 250 (governmental) portals worldwide relying on these software frameworks for describing and publishing datasets [16]. Yet, as more and more data is becoming available, findability, as well as quality and trust are of utmost importance in order to utilize the data. While in terms of findability, metadata about the temporal and geo-spatial scope of datasets are most relevant [10,15], *provenance* information has to be known to assess the trustworthiness of OGD. This is usually done in the form of giving a speaking label of the publishing body in the metadata. For instance, ”European Commission” is mentioned as a publisher of 12,448 datasets on <https://data.europa.eu/>; an organization that can be uniquely referenced also in existing knowledge graphs (KGs) such as DBpedia (`dbr:/European_Commission`)<sup>6</sup> or Wikidata (Q8880), however, such links are not (yet) explicit. Moreover, in other cases, different publishing organizations within the metadata have non-descriptive names such as ”Parlament”<sup>7</sup> (on <https://data.gv.at/>), which only in the context of the portal itself make sense.<sup>8</sup> Apparently, the publisher here actually refers to the *Austrian Parliament*. Alternatively, in other cases in addition, the contact information (e.g. an e-mail address or URL) found in the metadata can provide additional context on the publishing organization.

Summarizing, organizational information is usually not yet standardized in OGD portals by means of unique identifiers. Notably, this problem is aggravated by the fact that public bodies – just as any institution – are affected by *organizational changes*, that is, for instance ministries are being merged or restructured, and, therefore, the publishers may change over time and across different versions of datasets. Overall, this means that, while several qualitative comparisons of Open Data initiatives exist on a country level<sup>9</sup>, tracking the success of Open Data policies on the level of publishing organizations, or, respectively, tracking the development of these organizations in terms of mergers and re-structuring is hardly possible at the moment.

We argue that unambiguously linking Open Data publishers to URIs in public KGs would both increase findability of datasets (e.g. queries for datasets by statistical offices located in the European Union would be possible) as well as

<sup>6</sup> URL prefixes such as `dbo:`, `dbp:`, `wdt:`, or `schema:` can be referenced in `prefix.cc`.

<sup>7</sup> German writing of the English word ”parliament”.

<sup>8</sup> As <https://data.gv.at/> is the Austrian national data portal, the label ”Parlament” refers to the Austrian parliament.

<sup>9</sup> cf. for instance <http://opendatamonitor.eu> or <http://europeandataportal.eu/dashboard>

make it easier for data consumers to trust in the data, given reliable provenance information. Additionally, advanced queries for dataset monitoring and analyses would become possible. Lastly, even changes in organizations (such as mergers and renamings) would be less confusing for dataset users as long as the organizations still remain correctly linked. We therefore believe that linking Open Government Data and metadata with entities found in open KGs could be a solution to the stated ambiguity problems. Yet, to the best of our knowledge, neither the coverage of OGD publishing organizations, nor the specific challenges of this organizational linking problem have been investigated so far. The focus of the present position paper is therefore to study the feasibility and main open research problems for providing working solutions in this area. More concretely, we identify five challenges that are yet to be solved when linking organizations of public datasets to knowledge graphs, for each of which, we discuss potential solutions to be applied by Open Data publishers, the knowledge graph community, or – where possible – through automated linking approaches.

The rest of the paper is structured as follows: Section 2 provides an overview of the most important related work to our target contributions. In Section 3 we analyze and briefly show the (non-)performance of state-of-the-art entity linking systems. This analysis is based on a gold standard that we created to further analyze and motivate underlying issues; subsequently, the identified main challenges in our opinion primarily responsible for this poor performance are discussed one by one in Section 4, whereupon Section 5 discusses possible directions and starting points to tackle them. We conclude in Section 6 with an outlook and call for future work.

## 2 Background and Related Work

Our analysis and observations are mainly based on the data gathered by the *Open Data Portal Watch* (ODPW) project<sup>10</sup>. ODPW provides a large collection of Open Data metadata which has been compiled in order to monitor the quality of OGD portals: the ODPW project is regularly collecting metadata from over 250 portals world-wide, providing access to metadata dumps as weekly snapshots as well as various quality metrics [16,14] per portal, i.e., typically at country-level. However, a more fine-grained analysis of Open Data quality, as well as analysis of Open Data *on the level of single publishing organizations* is not yet supported, for the reasons we will outline in the following sections.

As for other related work on connecting Open Data to KGs, in [7] the authors propose a system to integrate user-generated mappings of attributes into an existing Open Data ecosystem; however, this system did not yet allow links to public KGs. Moreover, the temporal aspects of changes, that we will focus upon herein, were not covered there. The system was part of the former EU Open Data Portal<sup>11</sup> but is currently not available there anymore.

<sup>10</sup> <https://data.wu.ac.at/portalwatch>

<sup>11</sup> <https://data.europa.eu/euodp/de/data>

Tygel et al. [19] present a system to link datasets from different Open Data portals by extracting the tags and keywords from metadata descriptions: the tags get reconciled using automated translations and similarity measures, and re-published using unique URIs and meta-information for the reconciled tags. Again, specifically, links to organizations and temporal changes were not taken into account in this approach. However, the approach tries to solve multilinguality-issues using automated machine translation, which as we will discuss below, are also relevant in our context.

Overall, we observe that so far not much work has been carried out in terms of matching organizational information from OGD datasets to KGs. The most prominent recent contribution is the Google Dataset Search [4] service which offers a dedicated search engine for public datasets. To do so, Google links the identified datasets to their internal knowledge graph, in particular by partially mapping the publishing organizations. While no details about the actual matching approach and its coverage are provided, as a main challenge (besides data quality) they identify the ambiguity of organization names which are tackled by considering the website context for the mappings.

Related to our addressed challenge of linking organisations in the context of OGD is the heterogeneity of academic/research organisations; this is addressed by the EU-funded project RISIS.<sup>12</sup> The goal of this project is to provide a comprehensive register of public-sector research and higher education organizations in European countries. Each entry in the register provides a stable identifier and a set of characteristics of the entities, such as the website, country, and the entity type. While the register is a valuable source in the domain of research organisations, there is no coverage in the domain of OGD.

Notably, there are already existing standards and vocabularies which aim to solve the problem of heterogeneous metadata and missing links to publishing organizations. For instance, the *Semantic Government Vocabulary* (SGoV) [11], the *Data Catalog Vocabulary* (DCAT) [12], etc. – for publishing OGD. Yet, in practice, where these vocabularies are used (or mapped to, as in [14]) the respective attributes to link to publishing organizations are rather linking to (ambiguous) string labels than to URIs.

### 3 On the Performance of Current Entity Linking Systems

In order to analyze the challenges of linking organizational information to knowledge graphs' entities in depth, three methods have been applied which are presented in the following.

#### 3.1 Analysis of the ODPW Database

To assess the linking problem quantitatively, we focus on metadata from the Open Data Portal Watch data base [16], accessible via a public API. The data

<sup>12</sup> <http://risis.eu/orgreg/>

base consists of weekly metadata crawls from 252 data portals starting in early 2016. As of November 2019, when we conducted this analysis, the data covered 2,552,114 individual datasets. Under scrutiny for this work are the organizational metadata details and changes of those over the observed time frame. All statistical figures concerning organizational metadata in public datasets given in this paper refer to this metadata corpus unless stated otherwise.

### 3.2 Gold Standard for Change Analysis and Linker Evaluation

From the corpus, we created links from randomly chosen ODPW datasets by manually assigning the publishing organizations in terms of their existing Wikidata and DBpedia entities. We linked 200 distinct organizations of 174 distinct datasets. A match was only added to the gold standard when at least a link to Wikidata could be found. Each link was checked by at least two authors of this paper and only added if there was agreement concerning the link. The annotated instances in the final gold standard are from 57 different data portals and cover publishing organizations in different parts of the world. Notably, out of these 200, only 72.5% of the organizations could be manually matched to DBpedia which suggests a lower organization coverage for public administration institutions compared to Wikidata.

The gold standard also covers organizational changes in the datasets in terms of updates on the `dcat:publisher` property:<sup>13</sup> for 26 datasets direct changes can be observed in terms of updated label for the `dcat:publisher`; that is, 26 out of the 200 linked instances potentially reflect an organizational change, some of which could indeed be mapped to different organizations (whereas others only indicate a refinement or correction). The gold standard is publicly available on GitHub<sup>14</sup> under the CC-BY license. It can also be used to evaluate linking systems on their ability to match organizational entities.

### 3.3 Evaluation of Current Matching Systems

To make the point of limited usability of currently available entity linkers “off-the-shelf”, we evaluated multiple state-of-the-art linkers (on the `dcat:publisher` label information only) and also implemented a naïve baseline entity linker based on *term frequency - inverse document frequency* ( $TF/iDF$ ) by comparing whole metadata descriptions with DBpedia and Wikidata abstracts. For each target entity, i.e. an entity of a public KG, a document is built consisting of its labels, alternative forms, as well as its `dbo:abstract` in the case of DBpedia, and `schema:description` in case of Wikidata. The linker produces a one-to-many mapping by ranking  $TF/iDF$  matches in decreasing order according to their

<sup>13</sup> The ODPW metadata already maps different schemata uniformly to DCAT, cf. [17].

<sup>14</sup> <https://github.com/YaserJaradeh/LinkingODPublishers/blob/master/GoldStandard.csv>

similarity scores. To obtain a one-to-one mapping, only the top-1 match is considered for the evaluation. The linker is implemented in Python and is available on GitHub<sup>15</sup>.

In order to give a quick overview on the performance of current entity linking systems, seven state-of-the-art systems as well as our introduced baseline linker have been run on the datasets of our manually created gold standard. While linking to Wikidata achieves generally better results due to a higher concept coverage, the performance scores even for the best linking systems are clearly too low for a fully automated approach.

**Table 1.** Performance of different linking systems for the two target KGs Wikidata and DBpedia. The best F1 scores of each KG are represented in bold print. Most systems evaluated here are tailored to one specific target graph. The symbol (-) indicates that the system does not work on the stated knowledge graph.

	DBpedia			Wikidata		
	P	R	F1	P	R	F1
<b>Exact Matching</b>	0.071	0.059	0.063	0.099	0.102	0.1
<b>DBpedia Spotlight</b> [13]	0.214	0.223	0.217	-	-	-
<b>Text Razor</b> [1]	-	-	-	0.214	0.206	0.207
<b>EARL</b> [6]	0.204	0.2	0.201	-	-	-
<b>TagMe</b> [9]	0.055	0.067	0.06	-	-	-
<b>Meaning Cloud</b> [2]	0.105	0.104	0.103	-	-	-
<b>FALCON</b> [18]	0.266	0.254	0.258	-	-	-
<b>Open Tapioca</b> [5]	-	-	-	0.432	0.42	0.423
<b>Simple TF-IDF Linker</b>	0.39	0.373	<b>0.378</b>	0.621	0.587	<b>0.596</b>

This small experiment clearly demonstrates that relying solely on exact string matching techniques, or comparing abstracts with metadata descriptions achieves poor results as shown in Table 1. We argue that the problem is indeed not purely based in the non-suitability of matching techniques themselves but fundamentally related to open challenges brought by the nature of organizational data and their representation in KGs.

Though a simple TF/iDF approach is used for the naïve linker, it is still outperforming other baselines. One reason is that this simple linker only searches a part of the entire search space, namely the collection of organizations found in the knowledge graph. Other general purpose tools (e.g. DBpedia Spotlight) try first to find out what the entity type is and perform the actual linking afterwards. Therefore, in the case of the simple linker, the search space is more restricted and the disambiguation process is more accurate.

<sup>15</sup> <https://github.com/YaserJaradeh/LinkingODPublishers/tree/master/Scripts>

## 4 Challenges

In the following section we identify fundamental and in our opinion open challenges that complicate automated linking of organizations to KGs; each identified challenge will be illustrated and where possible quantitatively analyzed on our corpus or, respectively the KGs under consideration (DBpedia and Wikidata).

### 4.1 Challenge 1: Temporal Changes in Organizations

**Challenge Statement** Organizations change over time due to mergers, splits, or renamings. On the other hand, the metadata of open datasets also changes over time. Both types of changes complicate the automated linking process.

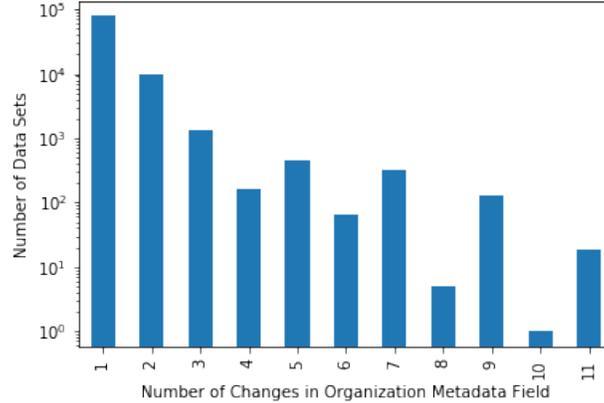
**Challenge Analysis** While the actual change in organizational structures is not in itself problematic, there are many consequences which affect the linking process. For instance, information about a change might not be reflected in the dataset’s metadata and/or in public KGs, or respectively be reflected asynchronously: there is latency in terms of both when/whether the change is updated in the metadata and the KGs. Furthermore, it is not clear how such temporal changes shall be reflected in public KGs (see Subsection 4.2).

In terms of temporal changes on the metadata level, we analyzed changes of individual dataset publishers per dataset and data portal over time. In order to take into account the sheer size of our metadata corpus and to increase the performance of our analysis, a heuristic was applied: only datasets where the organization label of the first occurrence is different from the organization label in the last occurrence were considered in this statistical evaluation. In total, 109,280 organizational changes could be identified in this way.

In a second step, the distribution of the number of organizational changes on a per dataset basis was analyzed. The maximum number of organizational changes is 11, meaning the organization of a single dataset on a data portal was changed 11 times. Figure 1 shows the distribution of changes. It can be seen that the distribution follows a power law: while 4 and more changes are relatively unlikely, datasets with one change clearly dominate the distribution with 78,378 occurrences.

It was further analyzed how many organizational changes cause the changes on dataset level, i.e. whether there are bulk changes that propagate across different datasets and portals (e.g. through “harvesting portals” that import metadata from other portals), and how these changes are distributed. The fact that there are only 33,879 distinct organization labels in the dataset but that there are more than 90,000 changes on a per-dataset level, indicates that bulk changes occur. We found that these roughly 90,000 changes on dataset level are caused by 12,489 individual changes of organization labels. It is important to note that the number of changes here does not necessarily reflect changes on distinct organizations – multiple renamings of the same organization are also counted (which can likewise be seen in Table 2). Figure 2 shows the distribution of individual organization label changes. Although there are few changes that are heavily picked up

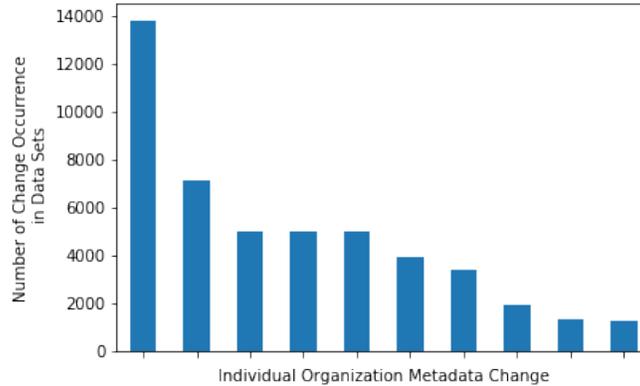
in the datasets, the distribution is more linear compared to the one in Figure 1. Table 2 displays the ten most frequent label changes on ODPW.



**Fig. 1.** The distribution of organizational changes for individual datasets. On the X-axis, the number of times an organization has been changed for a dataset on a particular data portal is shown while the overall frequency of such a change is reflected on the Y-axis. Note that the Y-axis is log-scale.

Overall, in our analysis of the Open Data Portal Watch data base, we qualitatively identify six reasons for organizational metadata changes – only two of them (I and II) being an actual change of the organization:

- I **Renaming:** Organizational changes due to actual renamings of the publishing institution such as from *Department for Communities and Local Government* to *Ministry of Housing, Communities and Local Government*.
- II **Structural Changes:** Changes due to structural transitions of the publishing organization such as mergers, divisions, or other restructurings. An example here would be the *Department of Energy and Climate Change* that was merged with the *Department for Business, Innovation and Skills* to form the *Department for Business, Energy and Industrial Strategy*.
- III **Specialization:** Changes to further specify which organization is meant such as changing the label *Department of Education* to *Department Of Education (Northern Ireland)*. In this category also fall changes that further define which part of the organization was involved in the dataset creation or provision such as from *Bristol City Council* to *Bristol City Council - Sustainability Team*.
- IV **Generalization:** Changes that generalize the authorship – most likely in order to make the publisher easier to find and to identify. An example here would be a change from *Martin Farrell* to *West Sussex County Council*.
- V **Editorial/Error Correction:** Changes due to corrections e.g. from *Ordnance Survey* to *Ordnance Survey*.



**Fig. 2.** The number of datasets in which an individual change, such as an organizational renaming, occurs. Each bar represents an individual change (cf. Table 2), e.g., the first one refers to the change from “NSGIC GIS Inventory (aka Ramona)” to “NSGIC GIS Inventory” and was propagated to almost 14000 individual datasets.

VI **Other:** Changes that we could not further classify, such as from *Frederick Manby*, *Noah Linden* to *Science* or from *Daryl Beggs*, *Ruth Oulton*, *Benjamin Lang* to *Benjamin Lang*, *Daryl Beggs*, *Ruth Oulton*, neither of which are mappable to an actual organization.

#### 4.2 Challenge 2: Lack of Consistently used Base Ontology for Modeling Organizations in Knowledge Graphs

**Challenge Statement** Organizations change over time – however, this change cannot be sufficiently expressed in current knowledge graphs’ ontologies or, respectively, existing capabilities of the vocabulary are not broadly used. Additionally, the vocabulary is insufficient in terms of expressing the relation between a geographic area and its governing body.

**Challenge Analysis** Wikidata offers specific properties for capturing different types of organizational changes. For example, *Also Known As* (`skos:altLabel`), *Official Names* (`wdt:P1448`), *Replaces* (`wdt:P1365`), *Replaced By* (`wdt:P1366`), and *Follows* (`wdt:P155`). Wikidata’s model also offers a property named *The Point In Time* (`wdt:P585`) which indicates the date from when a fact considered **true**. However, in the majority of the studied instances, these properties were not used to reflect organizational changes. In only 50% of all cases, an organizational change was reflected in Wikidata – mainly by using *Also Known As* and *Official Names* properties. For instance, the *Department for Environment and Water* (`wdt:Q5260295`) formally known as *Department of Environment, Water and Natural Resources*, is correctly listed in Wikidata using the property *official names* along with the *End Time* property (`wdt:P582`). However, other captured

changes are only annotated as *Also Known As* without any further details. A case in point is *Department of Finance, Services and Innovation* (`wdt:Q17004340`) previously known as *Department of Finance and Services*. While on the level of countries and states, the relation to governing or administrative bodies is clear through multiple concepts used with multiple properties (e.g. *office held by head of state* (`wdt:P1906`) or *applies to jurisdiction* (`wdt:P1001`)), this is not true anymore for smaller areas such as cities. Here, *applies to jurisdiction* (`wdt:P1001`) is typically used. In many cases, there is no distinction made between the area and the governing body.

DBpedia’s vocabulary also offers properties to capture organizational changes, such as `dbp:preceding`, `dbp:replace`, `dbp:predecessor`, `dbp:successor`, and property `dbp:merger`. However, such properties are not used widely to reflect actual changes. Furthermore, DBpedia lacks the temporal dimension that can be easily expressed in Wikidata. For instance, there is a property `dbo:mergerDate` – however, this property’s `rdfs:domain` is `dbo:Place` which is, in fact, disjoint with `dbo:Organisation`’s parent class `dbo:Agent` and consequently cannot be used to express the temporal details of an organizational merger. To quantify these statements, in the sampled organizational changes, only 34% of the conducted organizational changes were reflected in DBpedia entities. Moreover, similar to Wikidata, changes are rarely supported with details that describe the change. For instance, *London Fire and Emergency Planning Authority*<sup>16</sup> which replaces *London Fire and Civil Defence Authority*<sup>17</sup>, is captured by `dbp:predecessor` and `dbp:successor` properties in both instances – however, without any timeline information. In terms of the relation between geospatial areas and governing bodies on DBpedia, it is available to some extent in the vocabulary through multiple properties such as `dbp:governingBody`, but the vocabulary is rarely used.<sup>18</sup> Similar to Wikidata, in many cases no distinction is made between areas and the governing body. We note even that the DBpedia entities themselves are inconsistent in this regard, such as `European Union` both being typed as `dbo:Country` and `dbo:Organisation`, two classes labelled as `owl:disjointWith`.

### 4.3 Challenge 3: Metadata Quality

**Challenge Statement** Varying metadata quality among data portals complicates the automated linking process.

**Challenge Analysis** Poor metadata quality in provenance information is a major issue when linking organizations to unique KG entities. This issue has also been addressed by Google [4] and could be confirmed by our analysis of the ODPW data base: as an example, Table 2 displays the most frequent organization

<sup>16</sup> `dbr:London.Fire.and.Emergency.Planning.Authority`

<sup>17</sup> `dbr:London.Fire.and.Civil.Defence.Authority`

<sup>18</sup> A SPARQL query for `dbp:governingBody` resulted in  $\sim 6,000$  usages with only 930 distinct objects over all of DBpedia.

metadata changes on ODPW. It can be noted that from the 10 most frequent changes, at least 4 are not meaningful, and in fact losing semantics (marked in *italic*, i.e. not indicating any derferenceable publishing organization, but rather generic departments names or individual authors (potentially raising additional privacy problems)).

**Table 2.** Top 10 organization changes by label together with the number of occurrences of the change within datasets listed in ODPW.

Old Organization Label	New Organization Label	Frequency
NSGIC GIS Inventory (aka Ramona)	NSGIC GIS Inventory	13,793
Geoscience Australia	<i>Corp</i>	7,111
Daryl Beggs, Ruth Oulton, Benjamin Lang,	<i>Benjamin Lang, Daryl Beggs, Ruth Oulton,</i>	5,007
Daryl Beggs, Ruth Oulton, Benjamin Lang,	<i>Engineering</i>	5,007
Benjamin Lang, Daryl Beggs, Ruth Oulton,	<i>Engineering</i>	5,007
Ivan Begtin	Федеральная служба статистики	3,359
Archive bot	Национальный цифровой архив России	1,925
Senatsverwaltung für Gesundheit und Soziales	Senatsverwaltung für Gesundheit und Soziales Berlin	1,298
Senatsverwaltung für Gesundheit und Soziales Berlin	Senatsverwaltung für Gesundheit und Soziales	1,273
PAT S. Statistica	ISPAT	1,121

#### 4.4 Challenge 4: Multilinguality

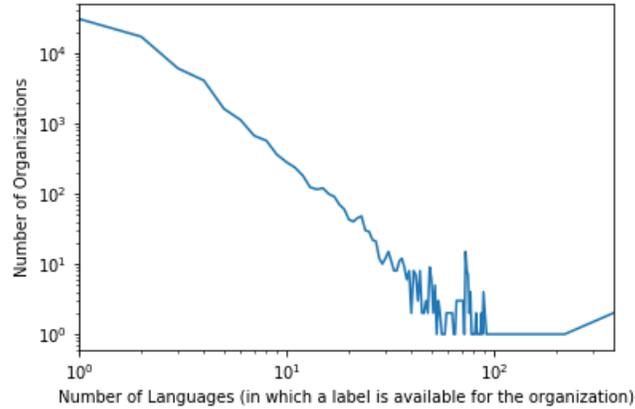
**Challenge Statement** As public dataset providers are spread around the world, different language identifiers further complicate the linking process. For example, the Chinese Central Bank is called *People’s Bank of China* in English, *Chinesische Volksbank* in German, and 中国人民银行 in Chinese.

**Challenge Analysis** The analysis of the ODPW data base showed that organizational labels are typically stated in the language where the publishing institution resides and that translations are often not given. An exception here is the European Data Portal ([europeandataportal.eu](http://europeandataportal.eu)): this portal harvests datasets from all member states, and the labels are automatically translated to English. However, automated translations do not necessarily correspond to the correct labels in other languages. DBpedia is not entirely multilingual in a sense that multiple labels are given in various languages for organizations in all cases: instead, there are dedicated DBpedia versions for multiple languages. Wikidata is more aligned in this regard: it is possible to define multiple labels in any given language. The People’s Bank of China, for instance, can also be found using its Chinese label. Even though multilingual labels can be defined on Wikidata, this is often not sufficient for our case. For example, the concept *Russian Federal State Statistics Service* (Федеральная служба статистики) *does* exist on Wikidata<sup>19</sup> – but there is no Russian label defined for it as of April 2020. Yet,

<sup>19</sup> <https://web.archive.org/web/20190403150124/https://www.wikidata.org/wiki/Q2624680>

this label appears more than 3,000 times in the ODPW data base as publisher.

Figure 3 shows the distribution of languages in which a label is given for Wikidata entities typed as organization on a logarithmic scale. As it can be seen, the distribution follows a power-law: For most organizations labels are defined only in a single language.



**Fig. 3.** Distribution of the number of languages in which labels exist for Wikidata concepts of type *Organization*. The axes are logarithmically scaled. The distribution follows a power-law.

#### 4.5 Challenge 5: Disambiguating Public Sector Organizations

**Challenge Statement** While companies can mostly be linked to one named entity without too much effort, this task is harder for public bodies.

**Challenge Analysis** The disambiguation problem is two-sided: (i) When only states and cities are quoted as originator of a dataset, there is ambiguity in terms of the actual concept that is referred to in the KG which may hold multiple entities for a particular label. For example: Does *New York* refer to the city of New York, the state of New York, or some particular administrative body of New York City? Wikidata contains entities for all three cases but the disambiguation is complicated without further context. (ii) Similarly, given a concept in a KG, it can be hard to link it to dataset publishers due to the ambiguities – in particular when acronyms are used; this has also been pointed out by Google [4]. Also institution names common in several countries (e.g. “*Statistics Office*” could be hard to disambiguate, though the portal context of nationally operated OGD portals may help here.

## 5 Towards Solutions for Linking Organizations

In order to tackle the aforementioned challenges we outline possible solutions paths and starting points below.

### 5.1 Challenge 1: Temporal Changes in Organizations

The fact that organizations change over time is a given. It is, therefore, important to acknowledge this and improve current technologies, including publicly available KGs and data portals, to better and more consistently represent such changes. For example, on data portals, the metadata should be timestamped so that it is clear as of which date the information is valid. It may, in addition, help not to only store the most recent version but to keep a history of organizational labels and metadata since a change may not yet be reflected in the target to which the organization is mapped.

### 5.2 Challenge 2: The Lack of a Base Ontology for Public Knowledge Graphs

At least as important as the design of a capable base ontology is the application. Our analysis showed, for instance, that the existing capabilities of the Wikidata and DBpedia vocabulary to reflect organizational changes are rarely exploited. Therefore, in order to automate and maintain mappings of such information, efforts to semantically represent organizational changes within KGs are required. This can be done by promoting currently available properties that express changes in organizations e.g. in the form of best practices for editors. At the same time, existing ontologies can be extended to better capture organizations and administrative units.

### 5.3 Challenge 3: Metadata Quality

Improving the quality of metadata across different open datasets will significantly improve the linking quality. As mentioned earlier, while some of the observed changes are meaningful, such as “NSGIC GIS Inventory (aka Ramona)” to “NSGIC GIS Inventory” or “Ivan Begtin” to “*Федеральная служба статистики*” (*Russian Federal State Statistics Service*), there is a concerning number of random changes occurring across datasets such as “Corp”. One way to improve the metadata that is provided is to refine information extraction methods used to create such datasets. Furthermore, dataset publishers should be urged and motivated to keep their metadata as current and as accurate as possible. Moreover, as we are aware that the data quality will not improve instantly, automated linking systems need to be able to handle a certain amount of noise. As our analysis showed, the metadata changes rather frequently (this is also observed by Google [4]) – hence, it is important to monitor meaningful

changes on regular basis. Dataset harvesters, such as ODPW, can be beneficial for this purpose by regularly retrieving metadata and detecting potential changes.<sup>20</sup>

#### 5.4 Challenge 4: Multilinguality

Multilinguality can be addressed by exploiting the multilingual capabilities of the Semantic Web. The problem is less pronounced for Wikidata compared to DBpedia. An active promotion of multilingual content (in KGs as well as on knowledge portals) can help in overcoming multilingual issues. For example, if multiple organization labels in multiple languages for the same organization would be available in the dataset metadata as well as in KGs, an overlap which might lead to a match becomes more likely. In addition, the use of dictionaries, such as *WordNet* [8] or *Wiktionary*, may help in some cases. For DBpedia, interlanguage links could be exploited to allow for multilinguality to a certain extent.

#### 5.5 Challenge 5: Disambiguating Public Sector Organizations

One of the most pronounced problems is the disambiguation of labels. Here, the context has to be very broad to also include, for instance, the local top level domain (indicator for the country), contact e-mail addresses, and data portal URLs. Specialized linking systems are required for this task as current generic solutions fail to successfully disambiguate organizations. We believe that exploring the context of the dataset during the linking process can provide more accurate predication of the linked resource. Information such as the portal ID or URL provides more indications of the organization's context such as the country. For example, utilizing the country of an organization which is often found as part of portal ID can improve linking accuracy: if two organizations have the same name but are coming from different countries (e.g. *Ministry of Education*) the broader context helps to disambiguate them.

#### 5.6 Across Challenges 1-5: Enabling a Community-Driven Linking Process

As outlined above, a fully-automated linking process is as of now not available. Therefore, we argue that while the community is working on improving the boundary conditions, a manual lookup service is required that allows a data science community as well as dataset publishers to annotate organizational links on a dataset level together with a community effort. This can be applied by allowing humans to use a voting function to ensure a high linking quality. We believe that such a service will not only improve the linking quality but will also

---

<sup>20</sup> Note that to a certain extend, up-to-date metadata is available e.g. through the ODPW data base that was also used for our analysis: <https://data.wu.ac.at/portalwatch/data>.

help in maintaining the linking results overtime. This service might look similar to [www.prefix.cc](http://www.prefix.cc), where publishers can be quickly looked up given a dataset URL or a data portal URL together with a label. As the same unique labels are used on many portals, the service could transitively reason organizational links for datasets not yet annotated which can be up or down voted by the user community. Over time, a larger gold standard could be created to improve and fine-tune existing linking systems.

## 6 Conclusion and Outlook

In this paper, we discussed the need and open problems of linking organizations of public datasets to their corresponding entities in public KGs such as Wikidata and DBpedia. In order to understand the current state of this issue, we created a gold standard mapping of open dataset organizations to KGs entities. We evaluated the performance of different current entity linking approaches including our own simple approach. As the results of the automated linking approaches were disappointing, we outlined five major challenges to be addressed. This includes (1) the temporal changes that happen on a regular basis in the organizations themselves and therefore in the metadata of open datasets. Our analysis also shows (2) that KGs are not fully using their existing capabilities to express organizational changes and we also address shortfalls in the existing vocabulary. We have also addressed (3) metadata quality issues and (4) multilinguality aspects within the linking process. Lastly, we found that (5) the disambiguation of public sector organizations is a hard task. We provide directions in terms of how these challenges can be addressed in the future. For future work, we aim to explore the idea of a community-driven effort in order to improve linking quality and to maintain that linking over time.

**Acknowledgements.** The authors thank Vincent Emonet, Paola Espinoza-Arias, and Bilal Koteich who contributed preliminary analyses regarding the challenges addressed in this paper. We also thank the organizers of the International Semantic Web Summer school (ISWS) 2019: the idea for this paper origins in discussions at the school.

## References

1. Extract meaning from your text., <https://www.textrazor.com/>
2. Text analytics – meaningcloud text mining solutions (Mar 2016), <https://www.meaningcloud.com/>
3. Assaf, A., Troncy, R., Senart, A.: HDL - towards a harmonized dataset model for open data portals. In: Workshop on Using the Web in the Age of Data (USEWOD '15) co-located with (ESWC 2015). pp. 62–74 (2015)
4. Brickley, D., Burgess, M., Noy, N.F.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference, WWW. pp. 1365–1375. ACM (2019)

5. Delpuch, A.: Opentapioca: Lightweight entity linking for wikidata. CoRR **abs/1904.09131** (2019), <http://arxiv.org/abs/1904.09131>
6. Dubey, M., Banerjee, D., Chaudhuri, D., Lehmann, J.: EARL: joint entity and relation linking for question answering over knowledge graphs. In: The Semantic Web - ISWC 2018 - 17th Int'l Semantic Web Conference. pp. 108–126 (2018)
7. Ermilov, I., Auer, S., Stadler, C.: User-driven semantic mapping of tabular data. In: I-SEMANTICS 2013. pp. 105–112. ACM (2013)
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts (1998)
9. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM. pp. 1625–1628 (2010)
10. Kacprzak, E., Koesten, L., Ibáñez, L.D., Blount, T., Tennison, J., Simperl, E.: Characterising dataset search - an analysis of search logs and data requests. *J. Web Semant.* **55**, 37–55 (2019)
11. Kremen, P., Necaský, M.: Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary. *J. Web Semant.* **55**, 1–20 (2019)
12. Maali, F., Erickson, J.: Data Catalog Vocabulary (DCAT). W3C Recommendation (Jan 2014), <http://www.w3.org/TR/vocab-dcat/>
13. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: 7th Int'l Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011. pp. 1–8 (2011)
14. Neumaier, S.: Semantic Enrichment of Open Data on the Web. Ph.D. thesis, Vienna University of Technology (2019)
15. Neumaier, S., Thurnay, L., Lampoltshammer, T.J., Knap, T.: Search, filter, fork, and link open data: The adequate platform: data- and community-driven quality improvements. In: Companion of the The Web Conference 2018 on The Web Conference 2018. pp. 1523–1526 (2018)
16. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. *J. Data and Information Quality* **8**(1), 2:1–2:29 (2016)
17. Neumaier, S., Umbrich, J., Polleres, A.: Lifting data portals to the web of data. In: 10th Workshop on Linked Data on the Web (LDOW2017) (Apr 2017)
18. Sakor, A., Mulang, I.O., Singh, K., Shekarpour, S., Vidal, M., Lehmann, J., Auer, S.: Old is gold: Linguistic driven approach for entity and relation linking of short text. In: Proceedings of the 2019 NAACL-HLT 2019. pp. 2336–2346 (2019)
19. Tygel, A., Auer, S., Debattista, J., Orlandi, F., Campos, M.L.M.: Towards cleaning-up open data portals: A metadata reconciliation approach. In: 10th IEEE International Conference on Semantic Computing, ICSC 2016. pp. 71–78 (2016)