Prerequisites for Large-Scale Integration of Industry-Relevant Linked Data

Kay Müller, Martin Brümmer, Markus Ackermann, Sebastian Hellmann

AKSW/KILT, Leipzig University & DBpedia Association
{kay.mueller,bruemmer,ackermann,hellmann}@informatik.uni-leipzig.de
http://aksw.org/Groups/KILT.html

Abstract. DBpedia represents the nexus of the LOD cloud, providing a huge cross-domain linked dataset. Due to the nature of its source datasets, data related to the industry domain such as data about products and supplying companies is practically limited. Since companies compete on a global scale, the faster a company can make informed decision, the higher are the chances that the company can get a competitive advantage over its competitors. Hence, extending DBpedia with company relevant data is of essence for the commercial exploitation of DBpedia. Successful integration of such data into DBpedia depends on a careful assessment of its availability, structure and terms of usage. Large-scale integration of heterogeneous, regularly updated company datasets from different sources depends on a scalable and extensible data integration framework that supports features like versioning, linking and curation. This paper describes an evaluation of these prerequisites.

Keywords: Company Data, Linked Data, Entity Resolution, DBpedia

1 Introduction

Our society is overwhelmed by the amount of structured and unstructured information that is available online. Consumers can choose from a massive amount of available products which are offered by many companies. With the help of consumer product review and price comparison websites, the user can narrow down the search to a small number of products according to his/her requirements. Shaped by this experience, users are expecting the same kind of information and services in the industry domain. Since companies compete against each other on a global scale, the faster a company can make an informed decision about products and suppliers, the higher are the chances that the company can get a competitive advantage over its competitors. A dataset that covers companies, products and branches would have many beneficiaries and applications:

- Industrial research on competition
- Data journalism for business-related articles.
- Government agencies for data enrichment and fraud detection.
- End-user applications & services.

DBpedia can help to realize these advantages by integrating company data into a large, cross-domain linked open dataset. This paper aims at providing the basis and an overview of existing company and related datasets. Furthermore the paper discusses current challenges and it proposes guidelines which can be used to integrate different company datasets.

The rest of the paper is organized as follows. Section 2 summarizes the available government and non-government company dataset providers. In Section 3 challenges regarding a common company URI are discussed. Section 4 available company RDF vocabularies are presented. Then Section 5 describes guidelines for a company dataset fusion workflow which can be used to automate the process. Finally, Section 6 concludes the paper and presents our future directions and discusses related as well as open research questions.

2 Survey of Available Company Data Sources

This chapter aims to create an overview of existing company data providers. In the context of this paper the term company data is data which helps to describe a company entity and its services. Examples of information that can be used to describe a company are a company name, a company address, the URL of the company website, the domain a company works in, what products a company produces and much more.

Our extensive survey started on October 2015 and included a lot of iterations, ending on December 2015.

2.1 Government Company Data

Only three open data providers have created surveys on how company registers are handled by governments all over the world:

- **Global Open Data Index:**[1] A crowdsourced platform which surveys how governments publish government data.
- **Open Data Index:**[2] A crowdsourced platform which surveys how governments publish government data in 14 different categories of which one is company registers.
- **OpenCorporates:**[3] A company dataset which aims to create a company ID for each company in the world. Up to now it has covered more than 95 million companies from all over the world.

On October 2015, the Global Open Data Index showed that only four countries opened up their company registers to the public for free: United Kingdom, India, Norway and Romania. Two months later (December 2015), the same data provider reported nine countries, including the governments of Australia, Indonesia, Republic of Moldova, Uruguay and Taiwan. However, in the meantime,

[1] `http://index.okfn.org/dataset/companies/`
[2] `http://open-data-index.silk.co/explore/T0ENtHf`
[3] `http://registries.opencorporates.com/`

the Indian government has decided to change their licensing model to a non-public one. The way different countries restrict access to company data can be summarized as follows:

- **License**: the license of company data has a big impact on how it can be used in private, research and commercial environments. Some countries restrict further use of company data by using a licensing model which does not allow the data to be used. Others restrict commercial usage only.
- **Fees**: some countries only grant access to the data for money. Some offer a public search interface which can be used to find restricted information and others offer all the data without a charge.
- **Data format**: offering the data in formats, which are not easily machine-readable present a defacto restriction as the time investment to convert data hinders users to users to access the data.
- **Freely searchable**: even though the data might not be offered for free in a machine-readable form, some company data providers offer users the ability to search for basic information like company names online by providing a simple search interface for running basic queries.
- **Updates**: since company information are not static, frequent updates are important. Otherwise the information which is stored in the company dataset can get quickly out of date.

Furthermore, different company datasets differ in the amount of information which is made available. For example, the United Kingdom publishes an updated company dataset via the companies house website[4] once a month as a downloadable dump. The files contain information such as company name, company type, basic accounting and mortgage information, former company names, etc. Another example is the Australian company dataset which does not contain information about whether a company changed its name and when. Furthermore, the Australian company dataset does not contain information about industry domains a company is operating in. The British dataset uses the the Standard Industrial Classification (SIC) system. This information is important if one wants to find companies which work in a certain industry sector.

Table 1 is based on the overview which was created by the Global Open Data Index. The first column shows the result for all 122 reviewed countries. The second column is based on the last row of the first column and gives an overview of all the states which have created a company dataset. The third column represents 19 of the G20 countries, where the 20th country is the European Union (EU). Since some of the EU countries are already covered by the group of G20, it was decided not to evaluate the EU separately.

Table 1 shows that only 62.30% of the governments have created a company dataset. Hence, the second column was created in order to see how countries with an official government company dataset have decided to handle their company data. It is interesting to see that out of all reviewed countries 32% offer at least some of their company data for free. Within the group of countries with an

[4] `http://download.companieshouse.gov.uk/en_output.html`

	122 surveyed countries	**Countries with Data**	**G20 countries (without EU)**
Public License	7.38%	11.84%	15.79%
Free	31.97%	51.32%	26.32%
Machine-readable	17.21%	27.63%	26.32%
Available in Bulk	20.49%	32.89%	31.58%
Frequent Updates	43.44%	69.74%	68.42%
Publicly Available	22.95%	36.84%	26.32%
Data in digital form	40.16%	64.47%	52.63%
Data available online	40.98%	65.79%	42.11%
Data exists	62.30%	100%	89.47%

Table 1: Government company data overview based on the Global Open Data Index

official government company dataset it is more than 50% and within the group of G20 states it is only about 26%. The group of G20 states achieves better results in all categories in comparison to the 122 reviewed countries. Especially in the areas of public licensing and frequent updates. This changes when the group of G20 states is compared to countries which have an official government company dataset. The only category in which the group of G20 states performs better is the handling of public licenses. Overall, it is difficult to obtain company data from government sources, since only about 23% of all reviewed countries have granted access to company data publicly and only about 17% of the company data is available in machine-readable form. Unfortunately, not all company data which is machine-readable is publicly available and according to the used source only about 12% of the government company data is machine-readable and publicly available.

2.2 Non-Government Company Data

Due to the restrictiveness of how company data is made available by different countries, one can find private groups and organizations which offer access to company and product data. Unfortunately, for the Linked Open Data community, many groups have realized the lack of openness in this area as a business model and hence most of the data can only be accessed when being payed for.

Dataset Provider	Number of companies	Region	License	Data Online	Free Data	Machine-readable	Bulk Download	Regular Updates
Crunchbase[5]	800,000	worldwide	Mixed	✓	Mixed	✓	on request	✓
DBpedia Core[6]	100,000	worldwide	CC BY-SA and GPL	✓	✓	✓	✓	✓
Europages Corporate[7]	2,600,000	Europe	Closed	✓	✓			✓
Global Research Identifier Database (GRID)[8]	49,000	worldwide	CC-BY	✓	✓	✓	✓	✓
Industry Stock[9]		worldwide	Closed	✓				✓
Net Estate[10]		worldwide	Closed	✓				✓
Open Corporates[11]	93,000,000	worldwide	ODbL	✓	✓	✓	on request	✓
OpenCyc[12]	26,000	worldwide	OpenCyc	✓	✓	✓	✓	✓
Thomson Reuters Open PermID[13]	3,500,000	worldwide	CC-BY and CC-NC	✓	✓	✓	✓	✓
SimFin[14]	183	worldwide	CC-BY-NC[15]	✓	✓	✓	✓	✓
Wikidata[16]	100,000	worldwide	CCO	✓	✓	✓	✓	✓
OpenStreet-Maps POI[17]	millions	worldwide	CC BY-AS	✓	✓	✓	✓	✓

Table 2: Company Dataset Providers

The overview in Table 2 lists a number of existing private company data providers. Due to space limitations, this does not summarize company data providers for regions or countries. For an overview of government company data providers, please refer to Section 2.1. Different company data providers use different ways of how the data is made accessible. Some are just wiki-pages (e.g.

[5] https://www.crunchbase.com/
[6] http://wiki.dbpedia.org/
[7] http://corporate.europages.co.uk/
[8] https://www.grid.ac/
[9] http://www.industrystock.com/
[10] https://www.netestate.de/en/information-extraction/data/website-database/
[11] https://opencorporates.com/
[12] http://www.cyc.com/platform/opencyc/
[13] https://permid.org/
[14] https://simfin.com/
[15] alternative commercial license available
[16] https://www.wikidata.org/
[17] http://wiki.openstreetmap.org/wiki/Planet.osm

SimFin) which allow users to download the required information as Excel documents.

Others support APIs which can usually be accessed using an API key. In order to get an API key, in most cases one has to pay a subscription fee to the company data provider. Depending on the chosen API key different amount of requests per day and per month are allowed (e.g. OpenCorporates, Crunchbase, etc.). Only a few company data providers allow bulk downloads (e.g. DBpedia, Wikidata) and sometimes only after a request (e.g. Crunchbase).Since DBpedia and Wikidata contain general information about different kind of entities, the datasets can be large in size. Therefore it is advisable to slice these datasets for the required information in order to reduce the graph traversal and ontology mapping overhead.

The company dataset with the most amount of company information is the OpenCorporates dataset. It contains more than 95 million company entities from all over the world. The OpenCorporates impact report of 2014[18] gives a good overview of what OpenCorporates has achieved and how much effort is required to keep a company dataset up-to-date. The mentioned report even notes that OpenCorporates sometimes receives legal and physical threats. Another company dataset which covers a lot of company entities is the Thomson Reuters Open PermID dataset with more than 3.5 million company entries. It is worth mentioning that it is possible to download the RDF dataset in a bulk download and a big part is under a license which can even be used in a commercial environment. Crunchbase has created quite a big dataset which covers more than 800,000 companies globally. It can be accessed in many different ways and a bulk download is possible on request.

Some company dataset providers have specialized in certain domains. One example is the GRID dataset, which mostly contains information about research institutions. Another one is Industry Stock and werliefertwas.de[19]. These data providers not only provide information about what companies exist, but also about what products they offer. This kind of company data provider is particularly useful, when performing research in conjunction with supply chain management. Unfortunately it is not possible to download the data or to use it via an API.

Since both Wikidata and DBpedia are connected to Wikipedia, the amount of company information is limited, since the Wikipedia community has come up with guidelines about when a Wikipedia article can be created for a company and/or company product.[20] Therefore only a limited number of companies can be found in DBpedia and Wikidata as shown in table 2. Another interesting dataset is the OpenStreetMaps (OSM) Point of Interest (POI) dataset. This dataset contains information like company names, country code, location and associated tags. In order to use the location information, one would have to

[18] http://blog.opencorporates.com/2015/05/26/opencorporates-impact-report-2014/

[19] https://www.wlw.de/

[20] https://en.wikipedia.org/wiki/Wikipedia:Notability_%28organizations_and_companies%29

"translate" the latitude and longitude information into an address. How OSM POI information can be used is shown in [2] and [13]. This dataset is a great source which can be used to find local branches of companies. Since it only stores very limited information per entry, matching an OSM entry to a company entry in a different dataset is not always possible.

2.3 Discussion

Getting access to company data is quite restrictive in most countries. Reasons for the restrictiveness are unfortunately not clear and speculating about them is not part of this paper. Since some governments have decided to open up their company data, expectations were high that the example will trigger a domino effect, which has not yet happened. In addition, company dataset providers like OpenCorporates, Crunchbase and Thomson Reuters PermID help to push towards an open dataset for companies. Since government company data is hard to come by, creating a large company knowledge graph involves a lot of work and resources. Hence, a lot of company data providers are investing a lot of work into writing crawlers, convince governments to make the data public and to deal with the law suits.[21] As a result, these company data providers have to invest a lot of financial resources to extend and maintain the company datasets. Hence, free access to the company data can fall under licenses which are free for private and research purposes, but set different conditions for commercial use. If the provided company dataset is going to be used in a commercial environment, companies have to pay monthly subscription fees in order to be able to access a REST API or a bulk download.

Since no unified company identifier exists which is shared between different datasets, it is hard to link company entities in different dataset which represent the same real world company. Unfortunately information about company subsidiaries and acquisitions are rarely part of the company datasets too.

In the domain of supply chain management it is often important to find potential suppliers of a particular product type. As highlighted in 2.2, most of the data providers do not offer this data for free. The only exceptions are DBpedia and Wikidata, which only provide limited amount of information about company products. If one requires product information, it might be best to collaborate with a company dataset provider which can provide access to the required information.

3 Description of Identifiers

When linking different dataset into one big one, having a unique URI/ID scheme is important. In [14] the authors compare different company identifier schemes which are ranging from DUNS numbers[22], tax IDs, DBpedia URIs, company

[21] `http://blog.opencorporates.com/2015/05/26/opencorporates-impact-report-2014/`
[22] `https://fedgov.dnb.com/webform/pages/dunsnumber.jsp`

websites, etc. Unfortunately most existing identification schemes are either not unique, can only be used in a certain region or are a subject to being changed at some point. Hence the author of the paper introduces a unique number which is associated with an entity (e.g. company, person, etc.). In the context of this work the following URI scheme is suggested for the company dataset:

1. namespace
2. country code where a company is based[23]
3. name of the company
4. unique ID (e.g. PermID)

Following this scheme, the URI for the Siemens AG could look like this:

http://corp.dbpedia.org/resource/DE/Siemens_AG__4295869238

The source dataset identifiers (URI of origin or integer value) will be stored in the dataset as well and thus provide a link back to the original dataset.

4 Available company vocabularies

For the specific domain of companies, their internal and external relationships to other companies, products, geographical locations and changes over time has not yet been comprehensively modelled. However, a number of related ontologies exist. The Organization ontology[24] models organizations, their geographical sites and members, including member posts and their respective roles in these organizations. Changes regarding organizations over time is addressed by `ChangeEvents`, which can result in new organizations being formed. In regards to organization members, `memberships` are associated with `intervals` that describe the timespan a certain agent was a valid member of the organization. The model can be easily adapted to the industry domain, evidenced by the fact that the Registered Organization Vocabulary[25], a profile of the Organization Ontology defining a number of more company specific properties, already exists. While these vocabularies are fit to describe companies themselves, their products and technologies are not yet covered.

`Schema.org`[26], is a collection of schemas for structured data on the Internet that provides a granular model of organizations, corporations and business in general, as well as products and product attributes. Being included in large content management systems like Drupal and adopted for search engine optimization purposes are strong incentives to use `schema.org` to model data targeting the industrial domain. However, `schema.org` is not easily transferable to an OWL ontology and does not use RDFS properties to formalize its class and type hierarchy (cf. [12]).

[23] `https://www.iso.org/obp/ui`
[24] `http://www.w3.org/TR/vocab-org/`
[25] `http://www.w3.org/TR/vocab-regorg/`
[26] `http://schema.org`

GoodRelations[27] is a vocabulary for E-Commerce that granularly models `BusinessEntities`, points of sale, products, services and their prices. It is well-established as a tool to provide RDFa markup for semantic search engine optimization. The ontology is aligned with `schema.org` using `owl:equivalentClass` statements. Although products and services are modelled in a detailed way in the GoodRelations ontology, organizational structures of businesses can not be expressed.

Lastly, the DBpedia ontology[28] also defines organizations, locations, events and contains properties to describe products of companies. However, historically, the ontology was directly derived from Wikipedia infoboxes and only currently the DBpedia Ontology Working Group has started to reengineer[29] the ontology for general and domain-specific data integration. In its current form, the DBpedia ontology was shaped by the mapping process that translates Wikipedia infobox templates to DBpedia classes and properties. Thus, DBpedia ontology classes feature a varying degree of granularity reflecting the non-systematic granularity of the infoboxes in Wikipedia. For example, there are currently nine subtypes of company in the DBpedia ontology[30] none of which deal with the type of incorporation of the company that would be an important feature for a classification. Products of companies are represented using the `http://dbpedia.org/ontology/product` property. The property does not have an `rdfs:range` defined. This results in under-specified semantics, as the property in most cases links to DBpedia resources describing the concepts or categories of products instead of a specific product instance directly.

5 Company Data Fusion Workflow

Each linked dataset normally goes through the life-cycle which is described in [1]. In brief this process encompasses the following steps. It starts with a search step where one has to find relevant datasets which can be linked. The identified dataset have to go through a linking process where entities of different datasets are linked. This process goes in hand with a classification process where source datasets are integrated with upper-level ontologies. Then data quality of the new dataset is checked and identified errors have to be fixed. Due to the large amount of company information which potentially has to be merged, the resulting puts many requirements on the ingestion and runtime environment. Figure 1 depicts our workflow to process such a challenging task. In order to streamline this process, we separate it into four different stages:

1. Resource change management (cf. Section 5.1)
2. RDF conversion and normalization (cf. Section 5.2)

[27] `http://www.heppnetz.de/ontologies/goodrelations/v1.html`
[28] `http://dbpedia.org/ontology/`
[29] `http://sourceforge.net/p/dbpedia/mailman/dbpedia-ontology/`
[30] Airline, Bank, Brewery, BusCompany, Caterer, LawFirm, Publisher, RecordLabel and Winery

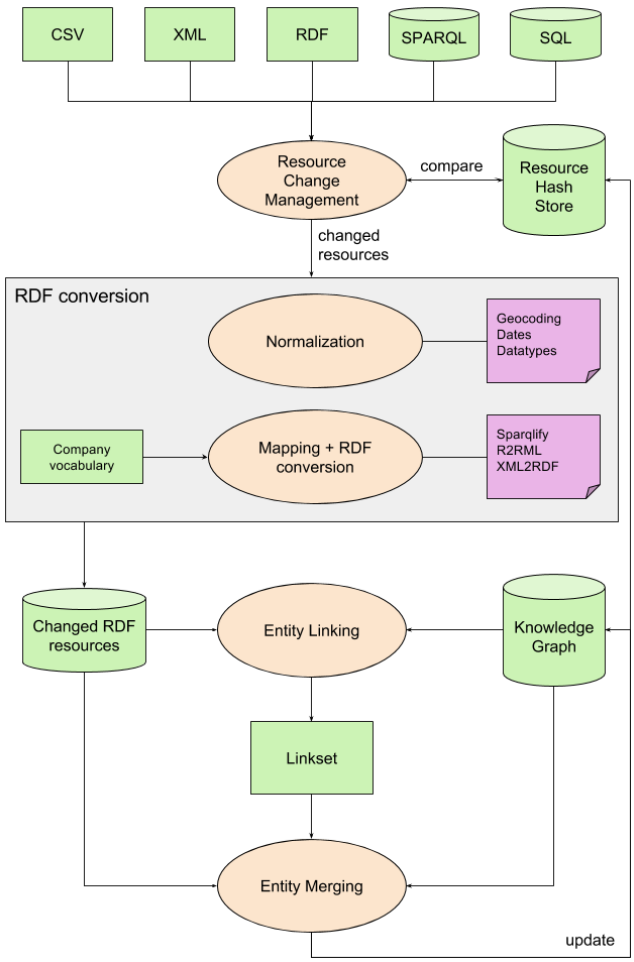

Fig. 1: Company Data Fusion Workflow Diagram

3. Entity linking and mapping (cf. Section 5.3)
4. Fusion (cf. Section 5.4)

All of the workflow steps require a lot of processing and where possible the processing should be executed in parallel. Over the last years several new big data analysis frameworks have emerged. Two of the most popular ones are Apache Flink[31] and Apache Spark[32]. Both frameworks are suitable to implement and execute the described workflow.

[31] `https://flink.apache.org/`
[32] `http://spark.apache.org/`

Different company data provider offer their data in different data formats. The formats will range from CSV and Excel files, databases, SPARQL endpoints to turtle, n-triples, N3, etc. In order to simplify the proposed workflow, the input file formats will be converted into a common RDF compatible file format. The mapping of the source file format into the common RDF format will be described by a mapping configuration which will exist for each source dataset. After the normalization step is finished, the resulting RDF file will be loaded into an SPARQL backend.

5.1 Resource Change Management

This part of the workflow is used to gather information about new, changed and deleted entities. Before checking with the system whether an entity exists, a URI is created which is unique for each entity within a dataset. How the URI is created depends on the source dataset. Sometimes the original dataset source entity URI can be used and sometimes a new URI has to be created (e.g. CSV file). In order to simplify the detection of changes for a known entity, a hash over all properties of an entity is created (e.g. MD5). For an entity which is described with the help of a CSV file, the hash is created on the whole input line. When the entity is described using a different input format, a configuration entry can specify what entity properties should be used to create the hash value. Once the entity information is gathered, it is checked against a corpus of known entities. This corpus stores information like:

- knowledge graph URI
- source entity URI
- entity hash over all entity properties
- source name
- flag on whether the known entity is valid

Once new, changed or deleted entities have been identified, they are passed to the next processing step. This ensures that the system does not have to process all the entities after a dataset update.

5.2 RDF Conversion and Normalization

With the help of a pre-defined mapping file, the input data is converted into a common RDF format. The common RDF format matches RDF source entity and property types to the entity and property types of the DBpedia-based company ontology. This generated RDF file will be used as the basis for further processing.

RDFUnit [8] is a Unit Testing Suite for RDF datasets and we plan to incorporate it into the RDF Conversion step to verify the generated data. RDFUnit executes SPARQL-based tests against datasets to spot data quality issues. These tests can be auto-generated from RDFS & OWL axioms (with a Closed World Assumption) and complemented by manually crafted constraints. The manual RDFUnit constraints can, for instance, require the minimum information that

must exist for a company entity within an input dataset. If for example, the company name property is the only one which is associated with a particular company entity, then this is an example of an entity which could be filtered out. The minimum amount of information required per entity can be specified in a config-file which is associated with the input dataset. In case an error or invalid entity is found with the help of RDFUnit, an entry is added to the Resource Management. The invalid entity is then filtered out from further processing.

Once the validated input data has been identified, a normalization process has to ensure that all the stored property values conform with the same data type format. The normalization step is important in order to increase the recall when trying to link entities in other datasets or when performing a search query on the dataset. Unfortunately the format of property values often differs even within a dataset. Since the generated RDF is already mapped to the company ontology, it is possible to apply the normalization step on the appropriate property values. Here are some examples:

- **Company Name:** Some company names use abbreviations (e.g. corp, inc) for the forms of enterprise, whilst other use their full forms. Here a set of dictionaries for different languages combined with rules and/or regular expression patterns can be used to create alternative names for a company. Therefore, the company name "Example Corp" could be expanded to "Example Corporation" and "Example Corp.". If a company name is an abbreviation, then it can be expanded from "EXAMPLE Corp." to "E.X.A.M.P.L.E. Corp", "E.X.A.M.P.L.E. Corporation", "EXAMPLE Corp.", etc. In order to avoid the creation of wrong naming variants very conservative regular expressions or language rules have to be used.
- **Address:** Since addresses can be written in different ways, it is important to normalize the address format as well. A possible way of normalizing the company address is to use a web-service like OpenStreetMaps Nominatim [5]. Nominatim supports a wide range of address formats and it returns a response with structured and normalized address information. The normalized address information can be used to store a normalized version of the address in the graph. In addition it is possible to extract additional information like town, region, country and even coordinates. This extra information can be used to enhance the existing dataset with additional geo-information where required.
- **URL:** Since the URL is a very important feature when matching two company entities, the guidelines which were outlined in this paper [9] can be used to normalize them into a common format.
- **Data Types:** Numbers, dates and other data types can be stored in different ways, it has to be ensured that they conform with the standards and are stored in the same data type format.

5.3 Linking different Datasets

Since there are many different company datasets, the linking and mapping process is important in order to link information about an entity between different

datasets. As mentioned before in Section 3 no unique company identifier exists which is shared between all the different datasets. Therefore it is important to have a linking strategy for each dataset pair. In order to ensure that the linking strategy can be executed, an entity linking tooling is needed. Two main link discovery frameworks for Semantic Web Data have been developed: LIMES[33] and Silk[34]. Both frameworks allow the definition of matching rules for entities. In addition, both support entity linking machine learning techniques which have been described in [10] and [7]. Although the company name is an important feature, it is not enough when linking two company entities. Two entities in different countries or different regions of the same country could potentially have the same company name. Therefore, the company name has to be combined with other property information like an address, company homepage URL, IDs, etc. The matching strategy differs between pairs of datasets and it depends on the amount of properties which are available per entity. If the number of properties which can be matched is small, then the possibility of creating false positive matches is very high.

It is important to realize that some entity properties are more important during the linking process then others. One of the easiest ways of matching two company entities is to use the company homepage URL. Although this linking strategy can create false positives by matching subsidiaries with its parent company, it can be used as a strong property feature in order to cluster a group of related company entities. Combined with other properties, like geo-information, a more reliable entity linking match can be produced. From this simple thought experiment one could assume that URL and IDs have a higher entity linking weight than a company name or address, since multiple companies can share the same name and multiple companies can share the same address. In order to achieve the weighted linking strategy, one can execute the entity linking frameworks multiple times. Each time with a different way of linking company names. The weighted intersection between the results of all these runs can then be used to create a more reliable mapping of entities. At the end of this workflow step a linkset is produced which can be merged with the existing knowledge graph.

5.4 Dataset Fusion

As shown in Figure 1 three different sources have to be used when merging the data into the common company dataset. The first part is the newly created linkset which contains all the linked entities between the source and the target datasets. The second part is the changeset with the newly created or deleted entities. Last but not least the original dataset has to be used, since existing entity properties might have to be updated. Once the merged entity data is created, the existing knowledge graph is updated. Then, the final entity URI is stored in the Resource Management Database together with the entity hash.

[33] `http://aksw.org/Projects/LIMES.html`
[34] `http://silkframework.org/`

6 Research questions and conclusion

In this work a survey of company data was described which reviews existing government and non-government company data providers. As the survey has shown, company data and in particular company product data is not easy to obtain. Despite the fact that not all governments have opened up their company data registers a considerable amount of company data still exists which has to be linked and made available for different business purposes. Conducting this process some along with many different challenges and opens up the discussion for several generally applicable research questions.

The first research question is related to the storage of provenance information. Due to the fact that the newly created company dataset consists of information which come from different source datasets, storing provenance information is very important. Through the works of the W3C paper [15] and projects like PROV-O ([3]) the topic has gained more attention within the Semantic Web research community. A common way of storing provenance information is via reification. Others have proposed the Singleton [11] or Quadruple [6] approach. Furthermore, this work has shown that company datasets are updated frequently. It is possible that company information changes. One example is the change of a company's CEO. There are multiple ways on how such a change could be represented in a dataset. The simplest and very common way is to replace the triple with the new information. Despite the fact that changing the information is very simple, time-variant information is lost. In this scenario it is not possible to query for previous CEOs of a company. Therefore, we will evaluate different models on how to store provenance, time, geo- and other metadata information in a graph without jeopardizing performance and the ability to query for facts and metadata information.[35]

The second research question is related to entities in the dataset to mapping industry or product domains. Different regions have different industry and product classification systems, which should all be reflected in the fused data in some way. Most of the company datasets do not provide any or insufficient information about what industry a company is operating in. Since many company related use cases require this information, it is important to add it to the dataset where required. In order to obtain this data other data sources like company websites have to be crawled. The crawled data can then be used to create some kind of topic model which can be matched to an industry domain. How to build such a gold standard and how to obtain information for the mapping process are interesting research questions.

Another research question is related to combining RDF knowledge graphs with REST backends. As highlighted in this paper many company data providers only offer their data via REST APIs which can be accessed using API keys. This problem is very likely applicable for other domains as well. Since it is not possible to download a bulk of data which can then be linked to a target dataset other means of integration have to be identified. Furthermore licensing issues with

[35] http://www.w3.org/2009/12/rdf-ws/papers/ws07

source datasets might complicate the integration in a target dataset. Integrating the source dataset at a service level might solve the problems. Hence we are planing to investigate to combine REST APIs of knowledge data providers with RDF. Although some work has been done in this area (e.g. [4]), there is still a lot of work to do in order to support a seamless integration of a REST service or remote SPARQL endpoint with a target dataset.

As one can see the domain of company data has a lot of room for very interesting research which is applicable to many other domains which have to deal with a lot of heterogeneous data. One of the major advantage of this domain is the large amount of potential commercial partners which are interested scientific advancements in this domain.

Acknowledgments.

This work was supported by grants from the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) for the SmartDataWeb Project[36]

References

1. S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the Life-Cycle of Linked Data with the LOD2 Stack. In *ISWC '12*, 2012.
2. S. Auer, J. Lehmann, and S. Hellmann. LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In *ISWC '09*, 2009.
3. K. Behajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV Ontology, W3C Recommendation. *W3C Consortium*, 2013.
4. P.-A. Champin. RDF-REST: A Unifying Framework for Web APIs and Linked Data. In *SALAD@ESWC '13*, CEUR, pages 10–19, May 2013.
5. K. Clemens. Geocoding with OpenStreetMap Data. In *GEOProcessing '15*, 2015.
6. G. Flouris, I. Fundulaki, P. Pediaditis, Y. Theoharis, and V. Christophides. Coloring RDF Triples to Capture Provenance. In *ISWC '09*, 2009.
7. R. Isele and C. Bizer. Learning Linkage Rules using Genetic Programming. In *OM '11*, 2011.
8. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, and R. Cornelissen. Test-driven Evaluation of Linked Data Quality. In *WWW '14*.
9. S. H. Lee, S. J. Kim, and S. H. Hong. On URL normalization. In *ICCSA '05*, 2005.
10. A.-C. N. Ngomo, M. A. Sherif, and K. Lyko. Unsupervised Link Discovery Through Knowledge Base Repair. In *ESWC '14*, 2014.
11. V. Nguyen, O. Bodenreider, and A. Sheth. Don't Like RDF Reification? Making Statements about Statements Using Singleton Property. In *ISWC '14*, 2014.
12. P. F. Patel-Schneider. Analyzing Schema.org. In *ISWC '14*, 2014.
13. C. Stadler, J. Lehmann, K. Höffner, and S. Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012.
14. B. Ulicny. Constructing Knowledge Graphs with Trust. In *METHOD '15*, 2015.
15. J. Zhao, C. Bizer, Y. Gil, P. Missier, and S. Sahoo. Provenance Requirements for the Next Version of RDF. In *W3C workshop RDF Next Steps*, 2010.

[36] `http://smartdataweb.de/`