



Article

VC-SLAM—A Handcrafted Data Corpus for the Construction of Semantic Models

Andreas Burgdorf ^{*,†}, Alexander Paulus [†], André Pomp [†]  and Tobias Meisen [†] 

Chair of Technologies and Management of Digital Transformation, University of Wuppertal, 42119 Wuppertal, Germany; paulus@uni-wuppertal.de (A.P.); pomp@uni-wuppertal.de (A.P.); meisen@uni-wuppertal.de (T.M.)

* Correspondence: burgdorf@uni-wuppertal.de

† These authors contributed equally to this work.

Abstract: Ontology-based data management and knowledge graphs have emerged in recent years as efficient approaches for managing and utilizing diverse and large data sets. In this regard, research on algorithms for automatic semantic labeling and modeling as a prerequisite for both has made steady progress in the form of new approaches. The range of algorithms varies in the type of information used (data schema, values, or metadata), as well as in the underlying methodology (e.g., use of different machine learning methods or external knowledge bases). Approaches that have been established over the years, however, still come with various weaknesses. Most approaches are evaluated on few small data corpora specific to the approach. This reduces comparability and also limits statements for the general applicability and performance of those approaches. Other research areas, such as computer vision or natural language processing solve this problem by providing unified data corpora for the evaluation of specific algorithms and tasks. In this paper, we present and publish VC-SLAM to lay the necessary foundation for future research. This corpus allows the evaluation and comparison of semantic labeling and modeling approaches across different methodologies, and it is the first corpus that additionally allows to leverage textual data documentations for semantic labeling and modeling. Each of the contained 101 data sets consists of labels, data and metadata, as well as corresponding semantic labels and a semantic model that were manually created by human experts using an ontology that was explicitly built for the corpus. We provide statistical information about the corpus as well as a critical discussion of its strengths and shortcomings, and test the corpus with existing methods for labeling and modeling.

Keywords: semantic labeling; semantic modeling; semantic mapping; data corpus



Citation: Burgdorf, A.; Paulus, A.; Pomp, A.; Meisen, T. VC-SLAM—A Handcrafted Data Corpus for the Construction of Semantic Models. *Data* **2022**, *7*, 17. <https://doi.org/10.3390/data7020017>

Academic Editor: Craig A. Knoblock

Received: 15 September 2021

Accepted: 22 January 2022

Published: 25 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Semantic mapping as an essential component of ontology-based data management (OBDM) and knowledge graph creation has received increased attention in recent years. In this context, we understand semantic mapping as the linking of data attributes of a data set with elements of an ontology. Since this is a time-consuming task, a research focus in recent years has been on the automation of the semantic mapping process, consisting of semantic labeling and semantic modeling [1]. Recent approaches for automated semantic mapping utilize labels and data values of the data set to assign semantic labels and infer relations in order to create semantic models. Figure 1 provides an example for a semantic model. An adequate data corpus that can be used to train and evaluate automated mapping algorithms is fundamental in this regard, especially if they rely on machine learning techniques.

In recent years, various corpora have emerged. However, none of them has yet become a standard that is reused in a variety of scientific papers. One limiting factor of the corpora used is that a significant percentage is suitable only for the semantic labeling part, but not for the full semantic modeling process. Furthermore, the used ontologies are small in size (or only a fraction of the available elements is used) and the corpora consist of, if covered,

only a few handcrafted semantic models. Moreover, the specified target semantic models are often small and do not represent data sets encountered in real-world data platforms, both in terms of the size and complexity of the model. However, rich models have great relevance in the conceptual description of data sets and offer potentials in the subsequent application of algorithms.

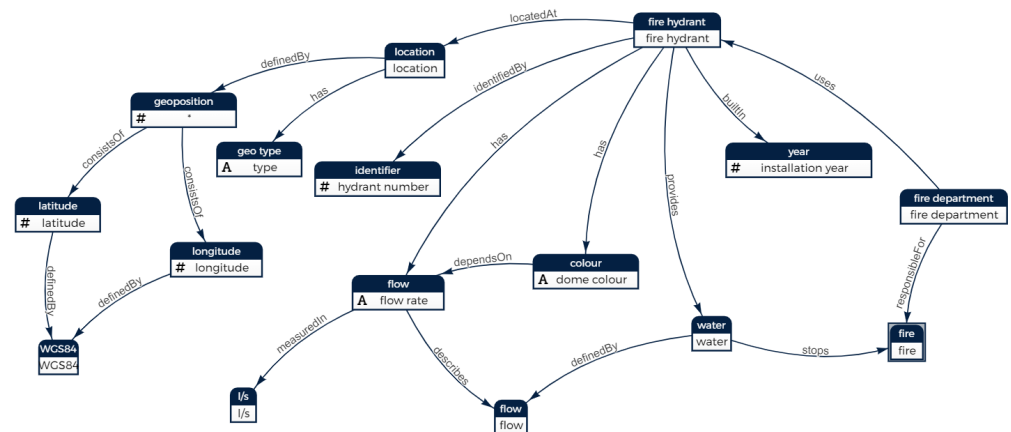


Figure 1. Example semantic model created for the corpus developed in this paper. This model describes water hydrants in Edmonton (ID 39). Each node consists of a concept and, if directly mapped to a data attribute, the attribute name accompanied by its data type.

It is apparent that previous data corpora are basically used for either labeling or modeling. Furthermore, existing sets are adapted to the previously used methods for semantic modeling (schema- and data-driven) and accordingly contain only necessary labels and values, since there was no need for more data. This hinders the investigation of modeling methodologies that have not been in focus of research in the past, but offer further potential, such as meta-data-based semantic modeling [2]. In fact, there is no corpus that allows to use textual data documentations for semantic modeling.

In order to promote the research and especially the comparability of new methods for semantic labeling and modeling, a new central corpus is of utter importance. With this paper and the published corpus, we are making a meaningful contribution. Our corpus has two important design goals: First, to have a rich ontology suitable for real-world use cases, and second, to include a large number of rich semantic models. We also want our corpus to support both semantic labeling and modeling, and to enable comparability of different approaches through benchmarks on the same data sets. Using a shared corpus enables researchers to compare approaches on a common evaluation basis for which algorithms yield comparable results if evaluated using the same metrics.

In the Section 2, we first give a survey of the current status of semantic mapping and the landscape of used corpora. Here, we especially identify obstacles in the evaluation of the recent developed approaches before we define objectives for a new corpus. In the Section 3, we describe how the data corpus was created. This includes data acquisition as well as ontology creation and manual modeling. In the Section 4, we provide statistical information about the corpus, test two algorithms for semantic labeling and semantic modeling exemplarily, and discuss strengths and weaknesses. Finally, we provide an outlook regarding future work and a conclusion in the Section 5. The described corpus is available on GitHub¹ and Zenodo².

2. Landscape of Semantic Corpora

In this section, we address our assessment that the current data corpora do not meet the requirements for general scientific comparison and reinforce the need for a new corpus. For that, we review recent approaches and take a closer look at the data corpora that were utilized by them. We identify obstacles that limit the comparability of the approaches.

Based on the identified obstacles of the existing data corpora, we subsequently specify objectives for a new corpus that mitigates the identified obstacles.

In order to explore the use of data corpora in the field of semantic mapping, we first distinguish between the different methods, as they depend on different types of data that need to be available in a data corpus. In the following, we distinguish between the previously mentioned steps of semantic labeling on the one hand, and semantic modeling on the other.

2.1. Semantic Labeling and Modeling

We divide semantic labeling into three different directions: (1) schema-driven semantic labeling, (2) data-driven semantic labeling, and (3) metadata-driven semantic labeling.

Schema-driven semantic labeling approaches, such as [3–7], use the labels of the schema of each data point within a data set, respectively. Matching or mapping in these approaches refers to the process of describing how schema characteristics, like the data labels of an input data set, are linked to concepts of a target ontology. In order to develop and evaluate algorithms, it is essential to have meaningful labels (i.e., non-cryptic) in the raw data; otherwise, those cannot be matched properly. Further, external knowledge, for example, in the form of knowledge bases, is often required.

By contrast, data-driven semantic labeling approaches, such as [8–12], use the actual data values contained in a data set to assign fitting concepts. Those approaches mostly rely on reference databases, statistics or machine-learning-based classification and, of course, require data values.

The metadata-based semantic labeling is so far only a conceptual approach (cf. [2]). Here, all available additional pieces of information on a data set are used that might contribute to the semantic labeling of the data. Examples include structured data like CKAN standards, metadata within a database, like comments, or unstructured data like textual data documentations, as they are available in documentation tools like wikis.

Independently of which semantic labeling approaches established an initial mapping, there exist various approaches, such as [13–19], that focus on the creation of a sophisticated semantic model by adding additional context so that the data set can be specified in more detail [19]. Besides the created labels, these approaches require a corpus that already covers semantic models.

2.2. Utilized Corpora and Resulting Obstacles

To gain an understanding of the problems associated with the corpora used so far, we consider the following scenario that we identified when comparing data-driven approaches. The Table 1 shows five approaches with their publication year and the corpora that were used for evaluation, whereas each corpus contains a different number of data sets. In 2012, Goel et al. [8] crawled websites from different domains and created three data corpora, namely, *weather forecast*, *flight status* and *geocoding*. They used these data corpora in order to evaluate their approach. Three years later, Ramnandan et al. [9] developed a new approach, which was evaluated using data corpora with the same name, indicating that these data corpora are the same. However, it is not explicitly stated whether these are the exact same data corpora, and no statistics are provided alongside the evaluation. While at first sight, the approach of Goel et al. achieved higher accuracy (0.97 vs. 0.64 for flight status), Ramnandan et al. ran the algorithm of Goel et al. for comparison themselves and evaluated it on their (probably) different data. They came to significantly worse results: an accuracy of 0.88 (weather forecast) and 0.42 (flight status). In addition, Ramnandan et al. created another data corpus, called the phone directory, and also evaluated their approach on it. Although there are several reasons why the re-implementation of the algorithm presented by Goel et al. [8] performed poorly, such as a different use of configuration variables or hyperparameters, it leads to a result that is difficult to compare. Another question that one might ask is why Ramnandan et al. [9] did not use the geocoding data corpus, but proposed a new data corpus containing phone directory data. However, in

subsequent years, only the weather data corpus was used to evaluate further approaches, with Pham et al. [10] achieving an accuracy of 0.95–0.96, and Ruummele et al. [11], an accuracy of 0.98. Here, the data corpora appear to have been the same, so we could assume that they are fairly comparable. However, instead of picking up the other already existing data corpora, additional shared data corpora were introduced and respectively used by Pham et al. and Ruummele et al. These are a *city*, a *soccer*, and a *museum* data corpus, which are not listed in the Table 1 to keep the table clear, as the different versions of the used corpora (cf. the Table 2) would significantly increase the size of the table. Although all four data corpora were used by these approaches, Ruummele et al. also used yet another corpus, called *weapons*, for semantic labeling, without explaining why it was needed or what gap it filled. While the approaches just mentioned—even if partly confusing—tried to evaluate their approaches in a comparable way, there are also plenty of approaches that evaluate on completely different data corpora (cf. [6,20]).

Table 1. Accuracies achieved for semantic mapping on the stated data corpora.

Publication	Year	Data Corpora			
		Weather Forecast	Flight Status	Geocoding	Phone Directory
Goel [8]	2012	0.89	0.97	0.98	
Ramnandan [9]	2015	0.96	0.64		0.83
Goel * [9]	2015	0.88	0.42		0.7
Pham [10]	2016	0.95–0.96			
Ruummele [11]	2018	0.98			

*: denotes that the approach of Goel et al. [8] was re-evaluated by Ramnandan et al. [9].

From scenarios like this, we see that using an approach developed by others does bear a risk of getting ambiguous results, and should therefore be avoided if possible. In addition, we see that the use of different corpora or the introduction of new corpora further reduces comprehensibility and comparability. However, if evaluation results are not comparable or cannot be verified for clarification, the credibility of those performance indicators is limited.

Therefore, this scenario essentially leads to two obstacles. First, the scenario demonstrated that it may be necessary or desirable to revisit other developers' approaches to compare one's own approach against them. For developers of a new approach, this means setting up and running all the other approaches they want to compare themselves with. However, the set up of foreign approaches may either contain some pitfalls, or even be completely impossible (e.g., as not all components have been made publicly available). Furthermore, from our experience, it takes a lot of time to achieve a working state when setting up other approaches' environments from scratch and without prior knowledge. Still, even after a setup has been successful, configuration is often an issue, especially when processes like feature engineering and hyperparameter tuning have to be carried out. If the original configuration of the environment has not been published by the original authors, re-configuring the algorithm's instance to yield the same results can be cumbersome. A different configuration of the respective approaches can lead to errors and accidentally affect the evaluation results. Thus, even if another algorithm can be run on a new data corpus and the same evaluation metrics are applied, results may not be trustworthy or comparable to the performance of the algorithm during its original evaluation. We define this as *Obstacle 1: Re-evaluation of other approaches is time-consuming*.

A related issue is the need to modify different approaches to evaluate them in a new environment. This mostly occurs if a different data corpus, which does not match the original evaluation data corpora characteristics, is to be used to compare both approaches. Altering the existing approach's algorithms or data ingestion may lead to unforeseen errors that are hard to detect, which might lead to the modified algorithm to underperform. As there exists no reference value for that algorithm on the new data corpus, it is hard to distinguish between bad performance or technical error. We define this as *Obstacle 2: Modification of other approaches is error-prone*.

However, if we now do not only consider semantic labeling, but also semantic modeling, the existing approaches are even more complex and difficult to compare. the Table 2 provides, for an excerpt of the above approaches, an overview of both the data corpora, as well as the ontologies used by these approaches. the Table 2 illustrates that there is no pair of approaches that uses the same data corpora and ontologies. While some approaches share a data corpus and an ontology, others have quite no intersection at all. When we inspect these sources deeper, we especially note that the museum data corpus finds a high degree of re-usage in recent publications. However, an even closer inspection of the evaluation setups shows that there is no clear versioning of the used data corpus. For instance, Taheriyani et al. [15] and Vu et al. [19] both utilized the same data set. However, Vu et al. stated that they modified the corpus for their evaluation. In this way, without explicit versions, the results obtained on slightly different corpora may quickly be compared with each other. We define this as *Obstacle 3: No uniform data corpus including an explicit versioning used.*

Table 2. An overview of recent approaches in semantic labeling and modeling and the data corpora and ontologies they utilize. Data corpus names occurring in different versions are denoted with the number of sources they consist of.

Publication	Linking Open Drug Data	City (17)	Museum (6)	Museum (29)	Museum (28)	Weapon Ads	Weather	City (10)	Soccer	Weather Forecast	Flight Status	Geocoding	Phone Directory	AAC	ACE	CIDOC-CRM	DBpedia	Dublin Core	EDM	ElementsGr2	FOAF	FRBR	GeoNames	ORE	Schema.org	SKOS	WGS84
[8]										x	x	x															
[9]				x						x	x		x				x		x								
[10]			x				x	x	x								x										
[11]				x		x	x	x	x																		
[13]	x														x												
[14]		x	x														x		x				x		x	x	
[15]			x											x					x	x	x			x		x	
[16]			x											x		x	x	x	x	x	x	x		x		x	
[17]			x			x										x			x						x		
[18]			x						x																		
[19]					x									x		x		x	x	x						x	

Furthermore, it is not only the use of different data corpora and their non-existent versioning that lead to poor comparability. Through the use of different ontologies, comparability becomes even more complex. As there are many different ontologies available for mapping (cf. the Table 2), their structure and size also influences the accuracy of a mapping algorithm's output. We define this as *Obstacle 4: A variety of ontologies are used inconsistently.*

In terms of future research and development, we also note that many existing data corpora cannot be used for metadata-driven approaches, as they do not contain appropriate metadata and, in particular, textual data documentations. In the related domain of table-matching, there are corpora such as *SemTab*³ [21], *VizNet*⁴ [22] or *T2Dv2*⁵, which also contain texts, but these are associated with high uncertainty. Texts in these corpora may be irrelevant, and in many cases, have no relation to the table data at all. We summarize this as *Obstacle 5: No data corpus with included metadata available.*

In essence, the above obstacles show that the approaches within a methodology (schema-driven, data-driven, metadata-driven) are not comparable even today. One solution would be to build a separate corpus for each method. However, it would still not be possible to compare approaches from different methodologies with each other. This becomes even more important when approaches work not only on one methodology, but on several (hybrid) approaches. When these additional issues are not considered, results will inevitably differ, which means that real comparability cannot be achieved.

2.3. Objectives for Semantic Mapping Corpora

In order to mitigate the identified obstacles, it is worth looking at how this specific problem is solved in other fields of research. In other domains, certain tasks are already clearly linked to a specific corpora. The *MNIST*-database⁶ [23] is used for digit recognition in computer vision, *Imagenet*⁷ [24] is connected to the task of object recognition and in the area of natural language processing, and the *Natural Questions*⁸ [25] data corpus is used for evaluating question answering approaches. In similar domains, there are the previously briefly introduced corpora, which target the area of data-driven semantic labeling for tabular data (*SemTab* [21], *VizNet* [22] or *T2Dv2*). The specific focus of these corpora is on data-driven semantic labeling, and they were especially used in the context of entity linking for knowledge graphs. The corpora are not suitable for evaluating hybrid approaches as well as cross-methodology approaches for general semantic labeling and modeling tasks. In addition, due to their lacking metadata, it is not possible to create rich semantic models for them.

The analysis of these corpora shows that there is a need for a unified corpus for semantic labeling and modeling to enable the comparability of approaches. The quality of a new algorithm compared to other algorithms can be measured by the results it achieves on a certain version of the common evaluation corpus. Especially when using different metrics, the advantages and disadvantages of the various algorithms can be more easily identified. In order to keep up with new developments, such data corpora are extended by the community, and versioning is taken into account (cf. [Obstacle 3](#)). This ensures that, for the purposes of comparability and clarity, it can be explicitly stated on which exact data corpus the presented results were achieved. This also means that when comparing with other algorithms, these do not have to be modified first, as the same data corpus is usable regardless of its version, which counters [Obstacle 2](#). However, those standard corpora do not avoid dealing with [Obstacle 1](#), as older algorithms might have to be re-evaluated on newer versions of the same data corpus.

Based on the identified obstacles and the standards for corpora in other domains, we define the following objectives that a data corpus for semantic labeling and modeling must satisfy:

Objective 1: provide data for all mapping methodologies: In order to work with all methodologies of semantic labeling, the corpus must provide schemata for schema-driven approaches, data values for data-driven semantic labeling, and metadata to support metadata-driven methods in every data set. If fulfilled, this will partially overcome [Obstacle 3](#) by forming a shared evaluation base for existing and new approaches. In case the corpus is actively maintained, having a proper versioning is required to fully remove [Obstacle 3](#). It is important to note that in case the corpus will, at any point in time, be used to train machine learning models, any available data values need to be included. Metadata must consist of at least a short description of the data, but is not limited in its level of detail. Metadata can be reached from a textual data documentation written by humans for a human target group to structured information like database documentation or data catalog standards. This should comprise any kind of metadata that can potentially be leveraged to generate semantic labels. In addition, for each data set within the corpus, it is not only established semantic labeling, but also a corresponding semantic model that needs to be included to support semantic labeling as well as semantic modeling. To support semantic refinement following automated semantic modeling, semantic models contained in the corpus should be rich in supplementary concepts and should outreach minimal models. This also means that users viewing the data can get a holistic understanding solely by inspecting the semantic model. This enables to get a quick overview about all the pieces of information that are relevant for the data set, but which are not explicitly included in the raw data.

Objective 2: enforce common data structures: To ensure any of the data sets contained in the corpus can be used for any algorithm, the structure in which the contents, that is, raw data and semantic models, are provided has to be well-defined. This allows approaches to use future parts of the corpus for evaluations that were not available at the time of development. If the data are later added, the approach can process the whole corpus without further modifications, avoiding [Obstacle 2](#). Similarly, when adding an additional data set to the corpus, the formats used need to be commonly known, and individual files should follow the standardized format.

Objective 3: provide used ontologies: To reduce the variety of different ontologies used ([Obstacle 4](#)) and simultaneously prevent unpredictable concepts, as they are missing in the target ontology, the corpus should provide a dedicated ontology containing concepts that are used in the semantic labels and models.

Finally, to allow comparability between the different methodologies, the named objectives must be fulfilled for every data set of the corpus: schema, the raw data, and metadata must be provided. Only then can approaches from each methodology use the corpus and methodologies be mixed up.

3. The VC-SLAM Corpus

To satisfy the objectives defined in the Section 2.3, we present the *Versatile Corpus for Semantic Labeling And Modeling (VC-SLAM)*. We describe the main features of the corpus, information on how the corpus was created, and how it can be used.

3.1. Description of the Corpus

In the following, we state the general characteristics of VC-SLAM. This initial version of the corpus is intended as a first cornerstone and should be continuously extended by the community while specifying a shared format (cf. [Objective 2](#)) and keeping a clean versioning (cf. [Objective 1](#)). Our composed corpus \mathcal{C} consists of a set of data sets \mathcal{D} and an ontology \mathcal{O} . Each data set $d_i \in \mathcal{D}$ must have the following content according to the defined objectives, as well as the requirements of the different methodologies:

- sdp_i : JSON-based samples with attribute names and values;
- raw_i : the raw data set including all its values;
- $note_i$: a short textual description of the data set content;
- $meta_i$: a textual data documentation of the data set, especially covering its meaning and its structure, as well as all other available metadata in an unstructured way;
- \mathcal{SM}_i : a set of semantic models, $\mathcal{SM}_{ij} \in \mathcal{SM}_i$, describing this data source where $|\mathcal{SM}_i| \geq 1$;
- For each concept $c \in \mathcal{SM}_{ij}$, it also holds that $c \in \mathcal{O}$.

We explicitly allow multiple semantic models per data set as the semantic models might differ between various users.

3.2. Data Set Identification and Acquisition

VC-SLAM has been built using data from the (*smart*) city context, a rapidly evolving field that requires being able to deal with different heterogeneous data sets. Available data sets originate in the domains of education, traffic, mobility, administration, and many more. To limit the context but include data from different domains, we limited the collection to data sets which have a geo-reference, but can be from every city-related context, like speed limits, public restrooms, or air pollution. This essentially allows to achieve an overlap between the semantic models and to still keep diversity across sub-domains. To find fitting data sets, we scanned 190 Open Data Portals (ODP)⁹, as those offer a great variety of data sets that match the target domain specifications. An exclusion criterion for the initial version of the corpus was that most portals did not contain textual data documentations, or that these documentations were not available in the English language. For this initial step, we acquired a total of 101 data sets from 23 different English language ODPs. We

provide an overview of the used sources and how many data sets we obtained from the respective ODP in our public repository.

3.3. Modeling Setup

The semantic model creation process was realized using the open-source tool *PLASMA*¹⁰ [26], which is a modular platform that enables users to create semantic models for the data sets using a dedicated user interface and underlying recommendations. However, for the creation of the models in VC-SLAM, no supportive technologies have been used in order to prevent any automation bias. For the integration of data sets in *PLASMA*, we needed to perform some pre-processing. The name and short description of the respective data sets were taken directly from the ODP into *PLASMA*. This allowed us to fix formatting errors that occurred during the transfer from the website. To perform the syntactic analysis of the data as preparation for the semantic modeling in *PLASMA*, three sample data points of the raw data were used to identify the structure of the data set. For data sets originally provided in the *GEOJSON*-format, we had to pre-process a data sample to reduce the level of nesting specified by the standard due to a limitation of *PLASMA*. We created a JSON file for each data source in addition to the raw data, which consists of a JSON array and exemplary data points. An ontology engineer screened all data sets that had to be integrated and created an ontology covering the geographic aspects of the data sets, since they are the intersection of all data sets. During the modeling itself, the ontology was extended to include specific concepts of the respective data sets in close consultation with the ontology engineer who monitored the ontology extensions.

3.4. Modeling

To achieve the best possible reference semantic labels and models, the whole modeling process was performed manually by five data modeling experts without the support of any algorithms suggesting a semantic model, or even an initial labeling. For this purpose, *PLASMA* was used, which presents the syntactic model of the data set to the user within the modeling user interface. In addition, the user can select semantic concepts and relations from a predefined list. This list includes the concepts from the initially loaded and continuously extended ontology. Modeling users were advised to leave attributes out during modeling that are describing metadata like references to the ODP the data set came from. Examples of this are automatically calculated grids within the ODP or internal identifiers of the ODP. Further modelers were advised to leave attributes out that were neither documented or should be left out in agreement with the ontology engineer. This particularly applies to those attributes with which ODPs enrich their JSON export, but not their raw data. For each data set, the list of removed attributes is explicitly available in our repository. The Figure 2 provides an overview of the modeling process and the interactions during the process. In our view, since there is a large number of different valid models for each data set, which differ mainly in the level of detail, quality assurance tools, such as inter-annotator agreements, cannot be used in the modeling process. That is why each model was discussed qualitatively by the group of modeling experts before finishing it. It was also ensured that each model is meaningful in itself and enriched by any meta-models, such as units, categories, and so forth, which on the one hand supports later consumers in quickly understanding the data, but is also the basis for the further application of semantic models, for example, for the purpose of semantic processing.

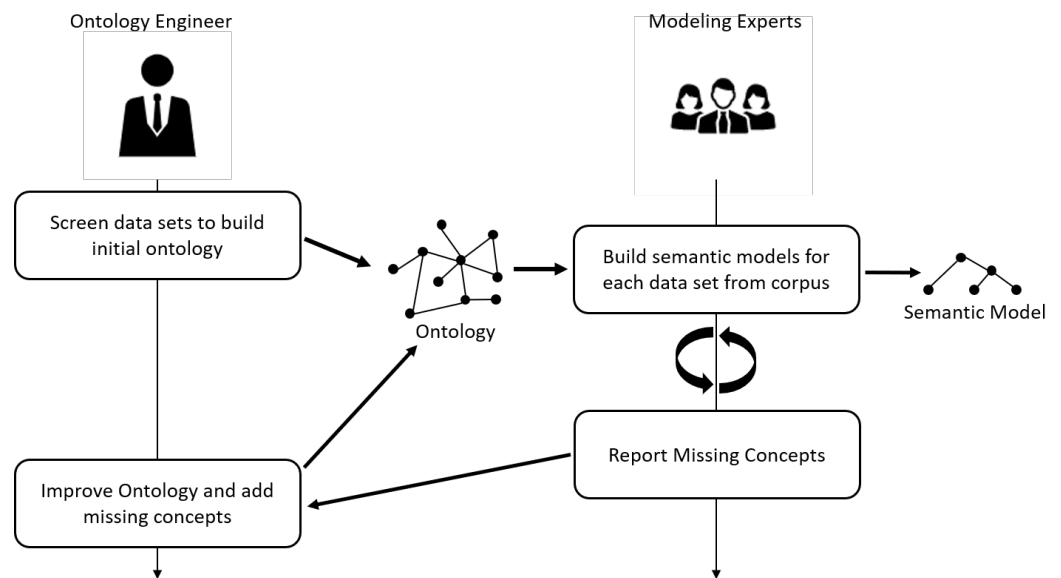


Figure 2. The process of ontology and semantic model creation.

After modeling, all models including the target ontology were exported. Our repository contains these files in the RDF-Turtle format. Thereby, we provide the ontology in two different ways to meet the requirements of different approaches for semantic modeling. The corpus contains the ontology once with distinct relations, where the same relation can basically only exist between the two same concepts (*ontology_dist.ttl*), and once with open relations where the same relation can exist between different concepts (*ontology.ttl*). The Figure 3 presents the connection of the different components of our corpus. In order to allow users of data-driven methods to use the values in the form of RDF literals, we recommend the use of a generic property “*vcslam:hasValue*”, which here allows a link according to the original label, or the paths represented in the mappings. A more detailed explanation of how to use the files can be found in the Git repository.

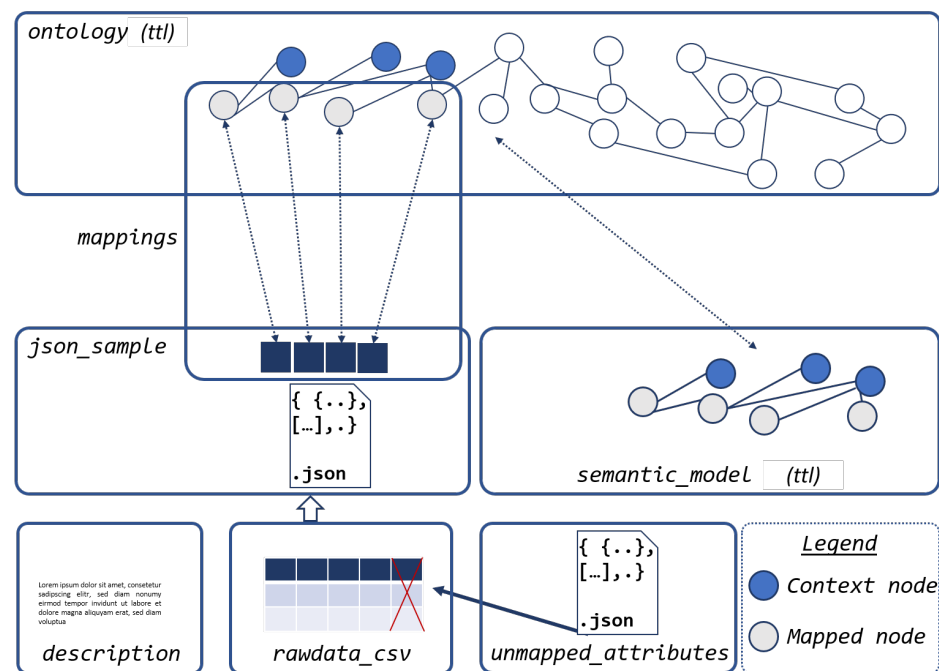


Figure 3. Overview of the different components of VC-SLAM and their interaction for semantic modeling.

4. Statistics and Discussion

The resulting corpus consists, in total, of 101 different data sets. In the following, we present a statistical description of the created corpus. For this purpose, we address the raw data or samples, the textual data documentations, semantic models, and the final ontology. Based on extracted statistics, we discuss to what extent the corpus meets our defined objectives. More in-depth statistics on the individual data sets of the corpus can be found in the Github repository.

4.1. Corpus Statistics

4.1.1. Raw Data

The raw data consists of 101 data sets that are in CSV or JSON format. However, as PLASMA requires data to be available in JSON format, we generated extracts of the data sets that were only available in CSV on the ODP. The Figure 4 illustrates the number of data points as well as the number of data attributes that are present in each data set.

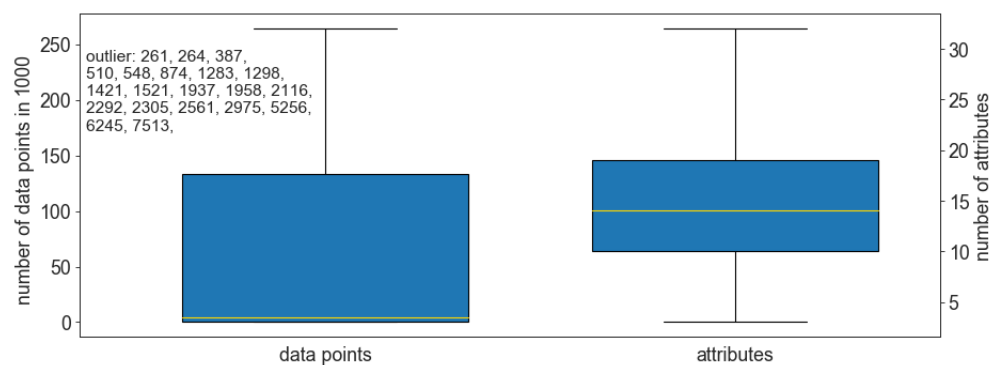


Figure 4. The boxplot visualizes for the individual data sets how many data points they consist of, and how many attributes each raw data set contains.

Each raw data set inside the corpus consists of attributes that have between 3 and 32 data attributes at a mean of 14.88. In comparison, the introduced data sets (cf. the Section 2) had an average attribute count of 11 (weather), 11.66 (weapons), 15.27 (museum), 11.5 (soccer) and 52 (city) [11]. The mean value of data points per data set is 451,325.6, with the smallest data set consisting of only four data points, and the largest consisting of 7,513,202. These values are important for the data-driven methodologies.

4.1.2. Metadata

The attached metadata required for the metadata-driven methodologies consist of unstructured data taken from the web pages of the respective data sets in the ODP. To assess the extent of the metadata, we first counted the words of each data documentation. Since the metadata focuses on the descriptions of each attribute, we also calculated a normalized factor indicating the words per attribute. To make a basic statement about the suitability of the metadata for semantic mapping, we also determined how many of the original attribute names and how many of the concepts used by the modeling experts occurred in the metadata. The Table 3 gives an overview of the acquired values.

The maximum value and the 75% percentile state that there are 44.91 (20.12) words of documentation per attribute. From this we conclude that there are several data sets in the corpus that could be sufficient to support metadata-driven modeling by terms of quantity. Even 13.09 words per attribute (50% percentile) could be sufficient and form at least one describing sentence per attribute. On average, 58% of the mapped concepts and 50% of the original attributes are present in the metadata, so at least partial linkage between metadata and ontology or original data are ensured. However, referring to the words per attribute, the minimum (3.72) and the 25% percentile (7.57) indicate that metadata are not sufficient for all data sets.

Table 3. The table presents the extent of metadata and their linkage to the original data and semantic models. The scope is represented by both the number of words and a normalized count (divided by the number of attributes in the original data points). The last four columns indicate how many attribute names from the raw data and how many of the concept names used in the semantic model occur in the metadata (absolute and relative with respect to included attributes or used concepts).

	Words	Words Attributes Attributes	Attributes in Metadata	Attributes in Metadata Attributes	Concepts in Metadata	Concepts in Metadata Used Concepts
mean	261.98	14.61	9.06	0.51	10.83	0.52
std	180.68	8.53	6.56	0.29	4.20	0.13
min	37.00	3.72	0.00	0.00	3.00	0.24
25%	128.00	7.57	4.00	0.25	8.00	0.43
50%	211.00	13.09	8.00	0.50	10.00	0.53
75%	337.00	20.12	12.00	0.79	13.00	0.62
max	948.00	44.91	30.00	1.00	21.00	0.82

4.1.3. Semantic Models

For each data set, we created a semantic model. The Figure 1 shows an example of the resulting semantic models. The Table 4 presents statistics on the properties of the models. We describe how many concepts and relations are used in a model and how many of them are unique. Furthermore, the table contains information about how many of the concepts are linked to attributes in the original data and which concepts are additionally present as nodes not directly linked to data, but providing further context. In addition, the table presents the number of attributes of the original data that are not mapped to any concept at all (cf. the Section 3.4).

Table 4. This table summarizes for the created models how many *concepts* and how many *unique concepts* are used. *Context concepts* denote the number of concepts mapped independently of original attributes, and *data concepts* denote concepts directly mapped to attributes from the data set. *Unmodeled attributes* refer to attributes of the original data that are not connected to any concept. The number of used relations is denoted by *relations* and *unique relations*.

	Concepts	Unique Concepts	Context Concepts	Data Concepts	Unmodeled Attributes	Relations	Unique Relations
mean	23.74	20.77	8.93	14.81	3.43	28.10	11.27
std	8.40	5.98	3.78	6.58	2.79	11.34	3.36
min	10.00	10.00	3.00	4.00	0.00	10.00	5.00
25%	17.00	17.00	6.00	10.00	1.00	20.00	9.00
50%	22.00	20.00	8.00	13.00	3.00	26.00	11.00
75%	29.00	25.00	11.00	19.00	5.00	34.00	13.00
max	55.00	36.00	23.00	32.00	11.00	68.00	20.00

Looking at original data and semantic models, it can be seen that the models are enriched with additional information. While there is a mean of 18.28 for the number of attributes in original data, there is a mean of 23.74 for the number of used concepts. The Figure 5 visualizes the number of concepts different models share. We see that all models share at least three concepts. Due to the choice of data sets, these come from the geo-sector. In addition, there are 40 models that share more than ten concepts, indicating similarities in the data sets beyond just geo-data.

With regard to [Objective 1](#), we see that the VC-SLAM corpus contains the necessary data for all presented mapping methodologies. The Figure 5 emphasizes that included data sets have a mean value of 451,325.6 data points that can serve for evaluating approaches from the data-driven methodology. Furthermore, we have provided the schema including attribute names of all data sets, which is the necessary prerequisite to evaluate schema-

driven approaches. The labels that occur are not only written-out attribute names, but domain- or provider-specific abbreviations are also possible. These are detailed in the accompanying metadata available for each data set, and form the basis for metadata-based approaches. Each data set is mapped to a semantic model that can be used to evaluate the results of the automated semantic labeling and modeling algorithms.

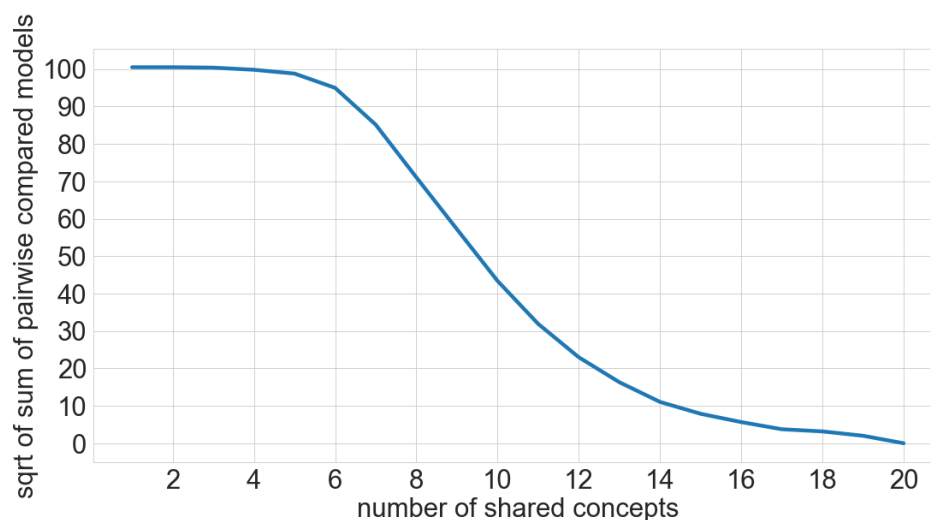


Figure 5. The line chart visualizes the number of shared concepts between models. On the x-axis we defined a threshold, and on the y-axis, the square root of the sum of all models having at least as many shared concepts defined by x.

In order to make access to the corpus as easy as possible, data are available in a uniform way for each of the included data sets, as defined by [Objective 2](#). The adherence to and reuse of existing standards ensures reusability for third parties. The use of RDF-Turtle format for models, TXT for metadata, and CSV, JSON or GEOJSON for raw data allows users to process the data with familiar tools and libraries. Apart from any knowledge bases that may be required, developers of different algorithms have all the data they need available in a common form.

4.1.4. Final Ontology

The ontology after modeling consists of 483 different concepts and 117 different relations. The frequency of reuse of concepts and relations is presented in the [Table 5](#). The large differences in the frequency of use of different concepts can be explained in the arrangement of the data. Due to the design of the corpus, the most frequently used concepts were latitude and longitude (each used 98 times), since almost all data sets in the corpus contained geo-data. Therefore, such basic concepts were applied in all related models. In contrast, the data sets differ enormously apart from the geo-context. For example, only one data set deals with graffiti, so the corresponding graffiti-concept does not find any reuse. In this way, the VC-SLAM corpus even allows to evaluate how well semantic model creation can work on unknown data sets. Individual statistics on used concepts and relations are available in the repository. The final ontology contains all concepts and relations that are used in semantic models. This reduces the amount of ontologies needed and, above all, the effort to unify them. To comply with existing standards, the ontology is also part of the corpus as an RDF file in Turtle format, as targeted by [Objective 3](#).

Apart from our defined objectives, the corpus also stands out for its size (101 data sets), especially in comparison to the repeatedly used museum (29 data sets) and soccer (12 data sets) corpus. Another novelty of VC-SLAM is that contained data have a common denominator with the city- and geo-domain, but in the underlying subject matter they originate from diverse domains. The [Figure 6](#) demonstrates the size of the VC-SLAM

ontology, which is many times larger than the number of classes and properties used by previous approaches.

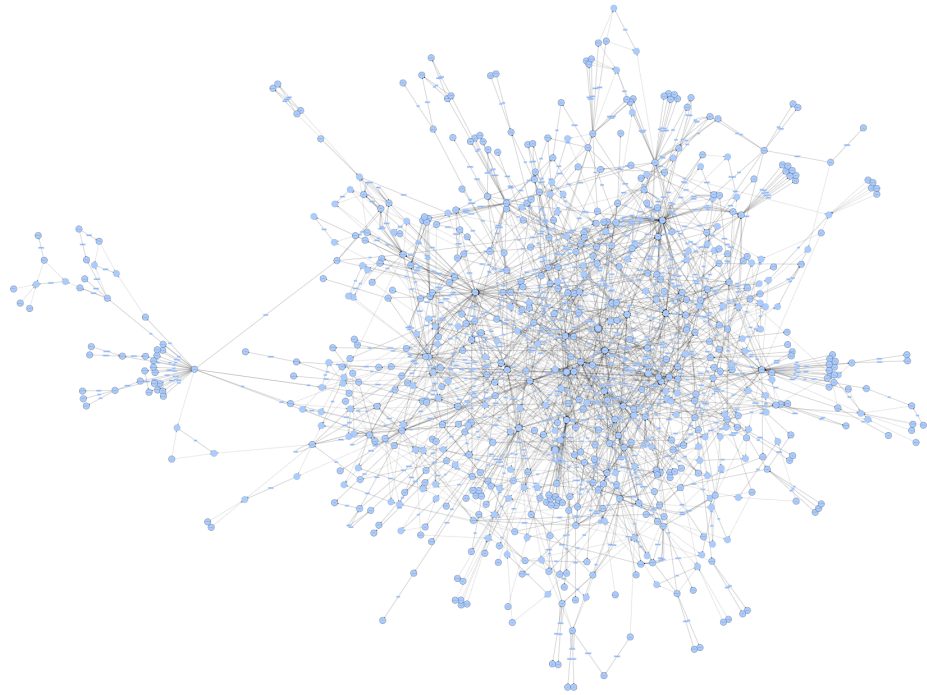


Figure 6. Visualization of the resulting ontology demonstrating the high degree of connectivity and the large number of clusters.

Table 5. The number of different concepts and relations within the ontology and statistics about the frequency of use.

	Concepts	Relations
count	483.00	117.00
mean	5.04	24.26
std	13.78	73.72
min	1.00	1.00
25%	1.00	1.00
50%	1.00	4.00
75%	3.00	10.00
max	113.00	535.00

4.2. Testing VC-SLAM on Existing Algorithms

We tested VC-SLAM on two approaches, one for semantic labeling and one for modeling in order to estimate the usability of VC-SLAM for those types of problems, as well as to evaluate the performance of said approaches on larger and more complex models.

4.2.1. VC-SLAM for Semantic Labeling

As a proof of concept of semantic labeling, we evaluated the Domain Independent Semantic Labeling (DSL) presented by Pham et al. [10]. In carrying out the evaluation, the obstacles addressed in the Section 2 again became apparent. The original implementation was outdated, which is why we used the new implementation by Rümmele et al. [11]. Even after adjustments to the software, we were only able to apply the approach to 78 of the 101 data sets from VC-SLAM. Although all the data used by VC-SLAM comply with the csv standard, we could not find any specific error during the tests. In order to get an overview of the achievable results, we started a leave-out-1 evaluation and examined with which probability the correct labels are found, on average, among the Top 1 to Top 9 results. The

Figure 7 provides an overview of the resulting accuracies. As one can see, the obtained results are worse than those reported in the original paper (cf. the Table 1), which is due to the fact that VC-SLAM involves significantly more extensive data sets with respect to the number of attributes and a much larger target ontology (17 concepts used in DSL vs. 483 concepts available in VC-SLAM). This confirms on one side [Obstacle 1](#) and [Obstacle 2](#), and shows on the other side that VC-SLAM fulfills the formulated objectives.

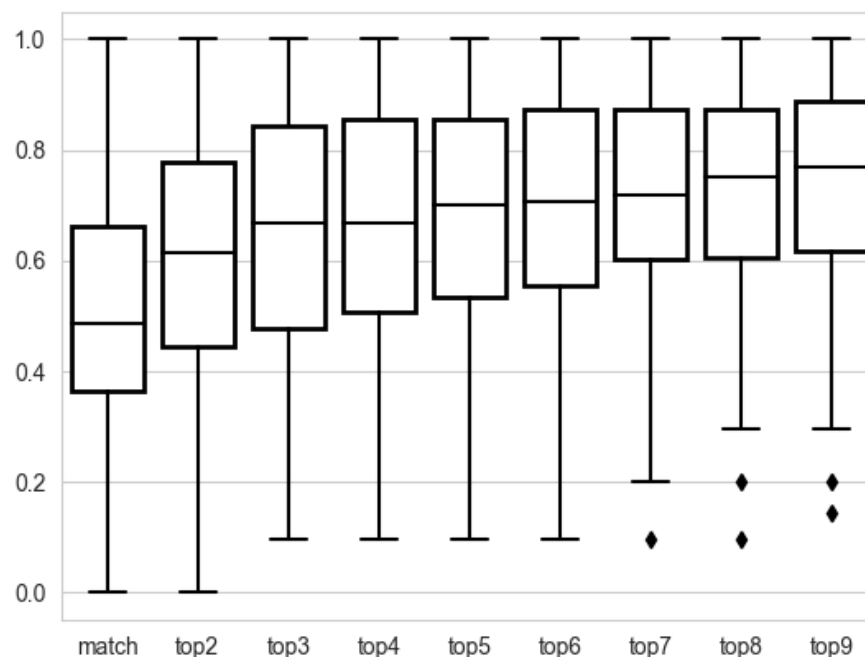


Figure 7. Resulting accuracy of the DSL approach on VC-SLAM.

4.2.2. VC-SLAM for Semantic Modeling

In order to test a semantic modeling approach on VC-SLAM, we used the approach by Futia et al. [27] that is available online¹¹. Since we only intended to test the semantic relation inference, we presumed existing labels. We converted the mappings provided by VC-SLAM to the format expected by SEMi and generated the extended graphs. We then used the provided Steiner tree algorithm to identify an initial semantic model. Once the Steiner tree models (in the form of minimal spanning trees) were generated, we attempted to train the graph neural network proposed by the authors in order to refine the semantic modeling. For the training, we used a leave-one-out setting, assuming that learning from 100 models for each model could suffice. However, this did not succeed, and we could not generate embeddings that were able to meaningfully refine the semantic model, presumably due to the lack of training data. This shows that our corpus is not yet usable for approaches that rely on a large number of training triples, which is due to our initial corpus goal of having a limited number of rich semantic models instead of many minimal models.

4.3. Limitations

Since the presented corpus is only an initial version, it still has weaknesses that we will overcome in future releases. We are aware that the amount of 101 data sets still is too small, especially if machine learning models have to be trained with the data and each underlying domain is to be present multiple times in the corpus. However, we believe that the quality already represented in VC-SLAM will encourage the community to use the corpus and expand it with more data that matches current domains. Up to now, the corpus has only been evaluated on two existing methods for semantic labeling or modeling. This is already a first step towards comparative benchmarks and shows the possibilities that VC-SLAM offers. However, additional approaches need to be evaluated using VC-SLAM

as a corpus. Along with the corpus, we will provide published benchmarks created based on VC-SLAM and link them to the associated source code if available. As discussed in the Section 2, this requires very intense effort of technical commissioning of individual software stacks. For machine-learning-based methods especially, we cannot guarantee that all parameters were chosen correctly to achieve an optimal result. Furthermore, the number of models in the corpus needed to be increased in order to allow machine-learning-based methods to achieve fair results. We plan to evaluate further selected approaches as one of the subsequent steps, but also hope developers of individual approaches will use the corpus in order to build up a valuable benchmark over time by either evaluating their old or new approaches. A limiting factor in obtaining suitable data sets was the availability of metadata in terms of textual data documentations. For the initial corpus, we only focused on using data that is already available in open data portals. However, our research has shown that there are a lot of ODPs that provide suitable texts, but which first have to be translated, because they are available, such as in Dutch, French or German. This also has the consequence that attributes in original data and, if necessary, even values must be consistently renamed.

5. Conclusions and Outlook

In this paper, we introduced VC-SLAM, a versatile corpus for semantic labeling and modeling. VC-SLAM aims to fill the gap that so far no common evaluation corpus exists in the field of semantic mapping. The corpus is structured to contain data for data-driven, schema-driven and metadata-driven semantic labeling approaches as well as semantic modeling approaches, and thus supports evaluation of approaches from different methodologies. In order to ensure evaluability of different approaches, the corpus consists of a common target ontology and manually created semantic models for each of its 101 data sets originating from the city and geo-domain. We hope that VC-SLAM can define a first milestone for a common benchmark in the semantic mapping research community.

In the future, we plan to constantly expand the corpus in new versions. A concrete maintenance plan and especially major enhancements depend on the acceptance of the community, but we will continuously improve the quality of the existing data sets as best as possible. Another important step we will take in the near future is to extend the corpus with additional data sets containing meaningful textual metadata to support future research in metadata-driven semantic mapping. For this, we will also take non-English sources into account where appropriate and set up a suitable translation pipeline. We further plan to establish one or more focus domains within the city domain by collecting data more specifically, so that the corpus allows us to distinguish between the achieved modeling quality within narrow boundaries as well as across domains during evaluation. Each extension to the corpus is also made available to the community via the repository, following a transparent versioning. For that, we plan that the corpus is made available in a new version every six months. In addition, we will publish the results of further evaluated semantic labeling and modeling methods as a benchmark in combination with the corpus. Thus, initial benchmark results are available and can serve as a comparison for future approaches.

Author Contributions: Conceptualization, A.B., A.P. (Alexander Paulus), A.P. (André Pomp) and T.M.; methodology, A.B., A.P. (Alexander Paulus) and A.P. (André Pomp); software, A.B., A.P. (Alexander Paulus) and A.P. (André Pomp); validation, A.B. and A.P. (Alexander Paulus); formal analysis, A.B. and A.P. (Alexander Paulus); investigation, A.B. and A.P. (Alexander Paulus); resources, A.B.; data curation, A.B. and A.P. (Alexander Paulus); writing—original draft preparation, A.B., A.P. (Alexander Paulus) and A.P. (André Pomp); writing—review and editing, T.M.; visualization, A.B.; supervision, T.M.; project administration, A.P. (André Pomp) and T.M.; funding acquisition, T.M. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge support from the Open Access Publication Fund of the University of Wuppertal and the Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The described corpus is available on GitHub <https://github.com/tmdt-buw/vc-slam> (accessed on 14 September 2021) and Zenodo <https://zenodo.org/record/5782764> (accessed on 14 September 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Notes

- 1 <https://github.com/tmdt-buw/vc-slam> (accessed on 14 September 2021).
- 2 <https://zenodo.org/record/5782764> (accessed on 14 September 2021).
- 3 <https://www.cs.ox.ac.uk/isg/challenges/sem-tab/> (accessed on 14 September 2021).
- 4 <https://viznet.media.mit.edu/> (accessed on 14 September 2021).
- 5 <http://webdatacommons.org/webtables/goldstandardV2.html> (accessed on 14 September 2021).
- 6 <http://yann.lecun.com/exdb/mnist/> (accessed on 14 September 2021).
- 7 <http://www.image-net.org> (accessed on 14 September 2021).
- 8 <https://ai.google.com/research/NaturalQuestions> (accessed on 14 September 2021).
- 9 A list of the concrete ODPs that were scanned is available at our repository: <https://github.com/tmdt-buw/vc-slam> (accessed on 14 September 2021).
- 10 <https://github.com/tmdt-buw/plasma> (accessed on 14 September 2021).
- 11 <https://github.com/giuseppéfutia/semi> (accessed on 14 September 2021).

References

1. Paulus, A.; Burgdorf, A.; Pomp, A.; Meisen, T. Recent Advances and Future Challenges of Semantic Modeling. In Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 27–29 January 2021; pp. 70–75.
2. Burgdorf, A.; Pomp, A.; Meisen, T. Towards NLP-supported Semantic Data Management. *arXiv* **2020**, arXiv:2005.06916.
3. Polfliet, S.; Ichise, R. Automated Mapping Generation for Converting Databases into Linked Data. In Proceedings of the 2010 International Conference on Posters & Demonstrations Track, Shanghai, China, 9 November 2010; pp. 173–176.
4. Pinkel, C.; Binnig, C.; Kharlamov, E.; Haase, P. IncMap: Pay as You Go Matching of Relational Schemata to OWL Ontologies. In Proceedings of the 8th International Conference on Ontology Matching, Sydney, Australia, 21 October 2013; Volume 1111, pp. 37–48.
5. Pinkel, C.; Binnig, C.; Jiménez-Ruiz, E.; Kharlamov, E.; Nikolov, A.; Schwarte, A.; Heupel, C.; Kraska, T. IncMap: A Journey Towards Ontology-based Data Integration. In Proceedings of the Datenbanksysteme für Business, Technologie und Web (BTW 2017), Stuttgart, Germany, 6–10 March 2017.
6. Paulus, A.; Pomp, A.; Poth, L.; Lipp, J.; Meisen, T. Gathering and Combining Semantic Concepts from Multiple Knowledge Bases. In Proceedings of the ICEIS 2018, Funchal, Madeira, Portugal, 21–24 March 2018; pp. 69–80.
7. Papanagiotou, P.; Katsioulis, P.; Katsioulis, P.; Tsetsos, V.; Anagnostopoulos, C.; Hadjiefthymiades, S. RONTTO: Relational to Ontology Schema Matching. *AIS Sigsemis Bull.* **2006**, *3*, 32–36.
8. Goel, A.; Knoblock, C.; Lerman, K. Exploiting Structure within Data for Accurate Labeling Using Conditional Random Fields. In Proceedings of the 14th International Conference on Artificial Intelligence (ICAI), Las Vegas, NV, USA, 16–19 July 2012.
9. Ramnandan, S.K.; Mittal, A.; Knoblock, C.A.; Szekely, P. Assigning semantic labels to data sources. In Proceedings of the European Semantic Web Conference, Portoroz, Slovenia, 31 May–4 June 2015; pp. 403–417.
10. Pham, M.; Alse, S.; Knoblock, C.A.; Szekely, P. Semantic Labeling: A Domain-Independent Approach. In Proceedings of the Semantic Web—ISWC 2016, Kobe, Japan, 17–21 October 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 446–462.
11. Rümmele, N.; Tyshetskiy, Y.; Collins, A. Evaluating approaches for supervised semantic labeling. In Proceedings of the Workshop on Linked Data on the Web, Lyon, France, 23 April 2018.
12. Abdelmageed, N.; Schindler, S. JenTab: Matching Tabular Data to Knowledge Graphs. In Proceedings of the 19th International Semantic Web Conference (ISWC) 2020, Athens, Greece, 2–6 November 2020.
13. Knoblock, C.A.; Szekely, P.; Ambite, J.L.; Goel, A.; Gupta, S.; Lerman, K.; Muslea, M.; Taheriyani, M.; Mallick, P. Semi-automatically mapping structured sources into the semantic web. In Proceedings of the Extended Semantic Web Conference, Crete, Greece, 27–31 May 2012; pp. 375–390.
14. Taheriyani, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. A Graph-Based Approach to Learn Semantic Descriptions of Data Sources. In Proceedings of the Semantic Web—ISWC 2013, Sydney, Australia, 21–25 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 607–623.

15. Taheriyani, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. A Scalable Approach to Learn Semantic Models of Structured Sources. In Proceedings of the 2014 IEEE International Conference on Semantic Computing, Newport Beach, CA, USA, 16–18 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 183–190.
16. Taheriyani, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. Learning the Semantics of Structured Data Sources. *J. Web Semant.* **2016**, *37–38*, 152–169. [[CrossRef](#)]
17. Taheriyani, M.; Knoblock, C.A.; Szekely, P.; Ambite, J.L. Leveraging Linked Data to Discover Semantic Relations within Data Sources. In Proceedings of the ISWC 2016—15th International Semantic Web Conference, Kobe, Japan, 17–21 October 2016.
18. Uña, D.D.; Rümmele, N.; Gange, G.; Schachte, P.; Stuckey, P.J. Machine Learning and Constraint Programming for Relational-To-Ontology Schema Mapping. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 1277–1283.
19. Vu, B.; Knoblock, C.; Pujara, J. Learning Semantic Models of Data Sources Using Probabilistic Graphical Models. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; ACM: New York, NY, USA, 2019; pp. 1944–1953.
20. Hulsebos, M.; Hu, K.; Bakker, M.; Zraggen, E.; Satyanarayan, A.; Kraska, T.; Demiralp, Ç.; Hidalgo, C. Sherlock: A deep learning approach to semantic data type detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1500–1508.
21. Jiménez-Ruiz, E.; Hassanzadeh, O.; Efthymiou, V.; Chen, J.; Srinivas, K. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In Proceedings of the European Semantic Web Conference, Athens, Greece, 2–6 November 2020; pp. 514–530.
22. Hu, K.; Gaikwad, S.; Hulsebos, M.; Bakker, M.A.; Zraggen, E.; Hidalgo, C.; Kraska, T.; Li, G.; Satyanarayan, A.; Demiralp, Ç. Viznet: Towards a large-scale visualization learning and benchmarking repository. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–12.
23. LeCun, Y. *The MNIST Database of Handwritten Digits*; NEC Research Institute: Princeton, NJ, USA, 1998.
24. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the CVPR09, Miami Beach, FL, USA, 20–26 June 2009.
25. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [[CrossRef](#)]
26. Paulus, A.; Burgdorf, A.; Puleikis, L.; Langer, T.; Pomp, A.; Meisen, T. PLASMA: Platform for Auxiliary Semantic Modeling Approaches. In Proceedings of the 23rd International Conference on Enterprise Information Systems, Online, 26–28 April 2021; pp. 403–412. [[CrossRef](#)]
27. Futia, G.; Vetrò, A.; De Martin, J.C. SeMi: A SEmantic Modeling machINE to build Knowledge Graphs with graph neural networks. *SoftwareX* **2020**, *12*, 100516. [[CrossRef](#)]