A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis

Yalemisew Abgaz[1], Amelie Dorn[2], Barbara Piringer[2], Eveline Wandl-Vogt[2], Andy Way[1]
[1] Dublin City University, [2] Austrian Academy of Sciences
Dublin Ireland, Vienna, Austria
{Yalemisew.Abgaz, Andy.Way}@adaptcentre.ie, {Amelie.Dorn, Barbara.Piringer, Eveline.Wandl-Vogt}@oeaw.ac.at

Abstract

Around the world, there is a wide range of traditional data manually collected for different scientific purposes. A small portion of this data has been digitised, but much of it remains less usable due to a lack of rich semantic models to enable humans and machines to understand, interpret and use these data. This paper presents ongoing work to build a semantic model to enrich and publish traditional data collection questionnaires in particular, and the historical data collection of the Bavarian Dialects in Austria in general. The use of cultural and linguistic concepts identified in the questionnaire questions allow for cultural exploration of the non-standard data (answers) of the collection. The approach focuses on capturing the semantics of the questionnaires dataset using domain analysis and schema analysis. This involves analysing the overall data collection process (domain analysis) and analysing the various schema used at different stages (schema analysis). By starting with modelling the data collection method, the focus is placed on the questionnaires as a gateway to understanding, interlinking and publishing the datasets. A model that describes the semantic structure of the main entities such as questionnaires, questions, answers and their relationships is presented.

Keywords: Ontology, E-lexicography, Semantic uplift

1. Introduction

There is a substantial amount of traditional data available on the internet and intranets of organisations. Traditional data, in this paper, refers to historical, socio-cultural, political, lexicographic and lexical data sets that are collected over an extended period. Public organisations such as museums, national bibliographic centres and libraries are increasingly opening their doors to facilitate access to such data to support research and development beyond their organisational boundaries (Doerr, 2009). This trend enables researchers to access a significant amount of useful primary data of historical, temporal and societal importance (Kansa et al., 2010; Beretta et al., 2014; Meroño-Peñuela et al., 2015). Making these data available, both for humans and machines, however, comes with several shortcomings.

First, in the majority of cases, these traditional data are initially available in bulk of archival formats, providing only a general description of the content of the data. However, they fail to provide detailed information about why, how, when and who collected the data and how the data can be interpreted and used. Often, consumers of such data require additional contextual information to understand and interpret the information contained in the datasets correctly. This is undoubtedly undesirable as it requires a considerable effort to understand and utilise the dataset.

Second, no matter how big and valuable a released dataset is, it is virtually impossible for machines to use the data without proper semantics for interpreting its content. As machines are becoming ever more typical consumers of such datasets, it has become crucial to include standardised machine-readable semantics in addition to the data itself. The limited availability of semantics to describe the data is, therefore, one of the leading obstacles for machines discovering and interpreting legacy data.

Third, interlinking of the data with other available datasets becomes difficult. The lack of semantics, the use of non-standard vocabulary or the absence of schema mapping (Bizer, Heath, & Berners-Lee, 2009) are some of the causes. Traditional data that includes a schema definition or a data dictionary provides useful information to aid the process of speedy utilisation, but often lacks the information about the means of interlinking the data with existing datasets especially with those available on the linked open data (LOD) platform. The interlinking of the data using a data dictionary further requires a mapping from the data dictionary to a standard vocabulary. This not only requires domain knowledge, but also a detailed knowledge of the internal structure of the data.

In this paper, we focus on a historical data collection of the Bavarian Dialects covering almost a century old data (1911-1998) from the present-day Austria. For effective opening up and utilisation of the collection, we present our approach to facilitating the semantic modelling, enrichment and publishing of traditional data, taking the data collection questionnaires and their individual questions as the starting point. The questionnaires and questions are essential parts of the entire collection as they serve as an entry point to access the answers, where typically neither the headword nor the definition are noted as standard terms. The use of linguistic and cultural concepts in the model thus allows for the exploration and exploitation of cultural links, which is one of the main aims of the exploreAT! project. The questionnaires of the "Datenbank der bairischen Mundarten in Österreich (DBÖ/dbo@ema)" within the project exploreAT! (Wandl-Vogt, 2012) is used as a case study to demonstrate the process. The approach is composed of major steps such as domain analysis, schema analysis, semantic model and semantic up-lift. Domain analysis includes the understanding of the rationale of the data collection, the method of data collection, the original documents used, primary agents that produced the data collection methods and those agents who collected the data. By employing this step, it is possible to collect significant semantics that describes the collection. Schema analysis of the dataset at various stages is also a crucial step, which includes a closer inquiry of the structure of the data, the relationship between entities and their attributes and investigation of any inconsistencies and anomalies. The semantic modelling step focuses on representing the structure and the semantics of the entities in the datasets using a well-defined semantic model. It is another essential step especially for domains that lack a suitable vocabulary to describe entities fully. In the absence of such vocabulary, it becomes crucial to build a semantic model of the domain from scratch. Finally, the semantic model is used to up-lift, interlink and integrate the data with other related datasets. It will serve as a means to open up valuable traditional data to support further research and possibly answer various questions involving the evolution of conceptualisations of societies in the past and the present.

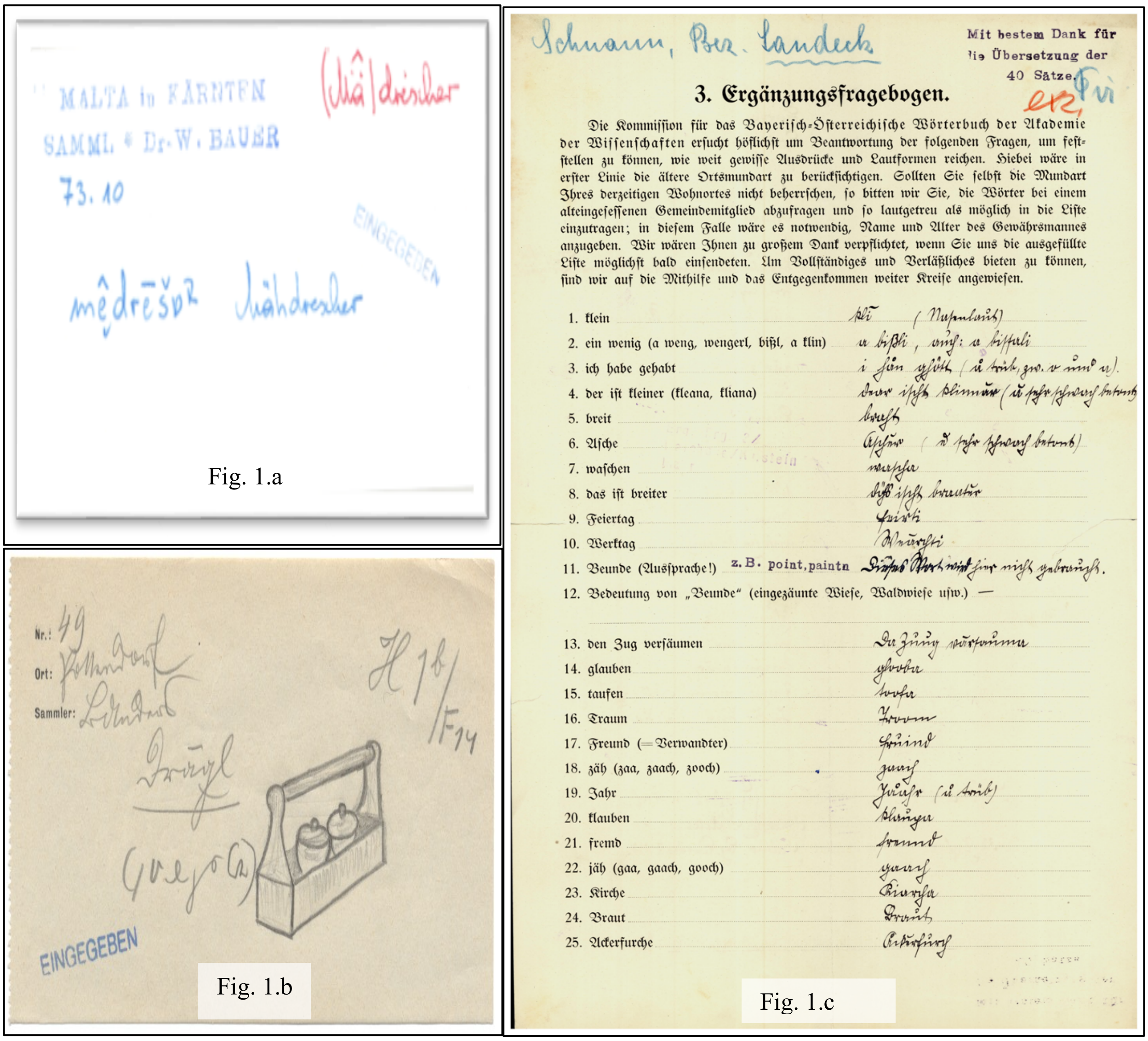
This approach enables organisations to make their datasets not only digitally available but also semantically enrich the dataset to facilitate a common understanding, interpretation and consumption by both machines and humans. The focus of this paper is, thus, to present our approach and the resulting semantic model. Even if the overall semantic model covers various aspects of the data, at this stage, it will focus only on modelling the questionnaires and questions, which provides users with a unique perspective of accessing the data, looking at it from the original questions and navigating to the corresponding answers, collectors or entities of interest. The model will further facilitate conceptual interoperability (Chiarcos et al., 2013) with other LOD repositories.

This paper is structured in the following way: Section 2 sheds light on the domain and describes the nature of the datasets in use. Section 3 presents the approach including domain and schema analysis and Section 4 discusses the core semantic model using the exploreAT! case study of Bavarian Dialects. In Section 5, we present ongoing work to utilise the semantic model towards the publishing of the datasets using LOD principles. Finally, the conclusion and future work are discussed in Section 6.

2. Background

2.1 Database of Bavarian Dialects (DBÖ)

The database of Bavarian Dialects (*Datenbank der bairischen Mundarten in Österreich -DBÖ) [Database of Bavarian Dialects in Austria]* (Wandl-Vogt, 2008) is a historical non-standard language resource. It was originally collected in the Habsburg monarchy with the aim of



MALTA in KÄRNTEN
SAMML. * Dr. W. BAUER
73.10

(Mā)drescher

mêdrēšɒʀ Mähdrescher

EINGEGEBEN

Fig. 1.a

Nr.: 49
Ort: Pottendorf
Sammler: [illegible]

Grüzl

(Grejo(a))

H 1b/F14

EINGEGEBEN

Fig. 1.b

Schnann, Bez. Landeck

Mit bestem Dank für
die Übersetzung der
40 Sätze.

erl.

3. Ergänzungsfragebogen.

Die Kommiſſion für das Bayeriſch-Öſterreichiſche Wörterbuch der Akademie der Wiſſenſchaften erſucht höflichſt um Beantwortung der folgenden Fragen, um feſtſtellen zu können, wie weit gewiſſe Ausdrücke und Lautformen reichen. Hiebei wäre in erſter Linie die ältere Ortsmundart zu berückſichtigen. Sollten Sie ſelbſt die Mundart Ihres derzeitigen Wohnortes nicht beherrſchen, ſo bitten wir Sie, die Wörter bei einem alteingeſeſſenen Gemeindemitglied abzufragen und ſo lautgetreu als möglich in die Liſte einzutragen; in dieſem Falle wäre es notwendig, Name und Alter des Gewährsmannes anzugeben. Wir wären Ihnen zu großem Dank verpflichtet, wenn Sie uns die ausgefüllte Liſte möglichſt bald einſendeten. Um Vollſtändiges und Verläßliches bieten zu können, ſind wir auf die Mithilfe und das Entgegenkommen weiter Kreiſe angewieſen.

1. klein .... klī (Naſenlaut)
2. ein wenig (a weng, wengerl, bißl, a klin) .... a bißli, auch: a biſſali
3. ich habe gehabt .... i hān ghött (ā trüb, zw. a und u).
4. der iſt kleiner (kleana, kliana) .... dear iſcht kliānär (ä ſehr ſchwach betont
5. breit .... broat
6. Aſche .... Aſchär (ä ſehr ſchwach betont)
7. waſchen .... waſcha
8. das iſt breiter .... diß iſcht broatär
9. Feiertag .... Firti
10. Werktag .... Wärchti
11. Beunde (Ausſprache!) z.B. point, paintn .... Dieſes Wort wird hier nicht gebraucht.
12. Bedeutung von „Beunde“ (eingezäunte Wieſe, Waldwieſe uſw.) —

13. den Zug verſäumen .... Da Zug varſauma
14. glauben .... globa
15. taufen .... toafa
16. Traum .... Troam
17. Freund (= Verwandter) .... Freind
18. zäh (zaa, zaach, zooch) .... zach
19. Jahr .... Jauhr (ā trüb)
20. klauben .... klauba
21. fremd .... fremd
22. jäh (gaa, gaach, gooch) .... gach
23. Kirche .... Kircha
24. Braut .... Braut
25. Ackerfurche .... Ackerfurch

Fig. 1.c

Figure 1. Sample paper slips (a, b) and sample filled questionnaire((Ergänzungsfragebogen) (c).

documenting the German language and rural life in Austria from the beginnings of the Bavarian dialect to the current day. The inception of the data collection went back to 1913 and continued until 1998 in present-day Austria, Czech Republic, Slovakia, Hungary and northern Italy, leaving a century-old historical, socio-cultural and lexical data resource. Even if the original aim of the collection was to compile a dictionary and a linguistic atlas of Bavarian dialects (Arbeitsplan, 1912) spoken by the locals, the data includes various socio-cultural aspects of the day-to-day life of the inhabitants, such as traditional customs and beliefs, religious festivities, professions, food and beverages, traditional medicine, and many more (Wandl-Vogt, 2008)

The data was collected using 109 main questionnaires, nine additional questionnaires (*Ergänzungsfragebögen*) and two *Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen questionnaires* and other additional freestyle questionnaires and text excerpts from various sources such as vernacular dictionaries and literature. In total, there are 24,382 individual questions corresponding to the available questionnaires in the collection. In response to the questionnaires over the span of the project, several million (~ 3.6 million) of individual answers noted on paper slips (Fig. 1.a, b) were collected. The answers to the questions include single words, pronunciations, illustrations and explanations of cultural activities on topics such as traditional celebrations, games, plays, dances, food and other topics.

In addition to the primary data, the entire collection also includes biographies of individual collectors and contributors of various roles. 11,157 individuals who had various functions in the project had participated in the data collection process as authors of the questionnaires, data collectors, editors or coordinators, with some having several of these functions at once. Detailed information about the personal background of individual contributors which was also noted in the course of data collection and during the digitisation process in later years is stored in a specific database (*Personendatenbank [person database]*). Persons and their background are thus other important features of the data that offer additional points for the exploration and the systematic opening of the collection.

The data set further contains additional information about the geographic locations and names of places including cities, districts and regions related to the places where the questionnaires were distributed. In rare cases, the paper slips include information about the time of the data collection.

The collected data has been used to produce a dictionary, *Wörterbuch der bairischen Mundarten in Österreich [Dictionary of Bavarian Dialects in Austria]* (WBÖ); up to now five volumes (A–E, P and T) have been published. Today, about three-quarters of the collected paper slips are available in a digital format following several stages of digitisation. The available formats corresponding to the stages include scanned copies of the paper slips, a textual representation of the paper slips in TUSTEP, MySQL (Barabas et al., 2010) and TEI/XML (Schopper, 2015). This is an ongoing effort to make the data accessible and analyse them, including the use of semantic web technologies to make the data suitable for semantic publishing in the LOD platform.

3. Approach

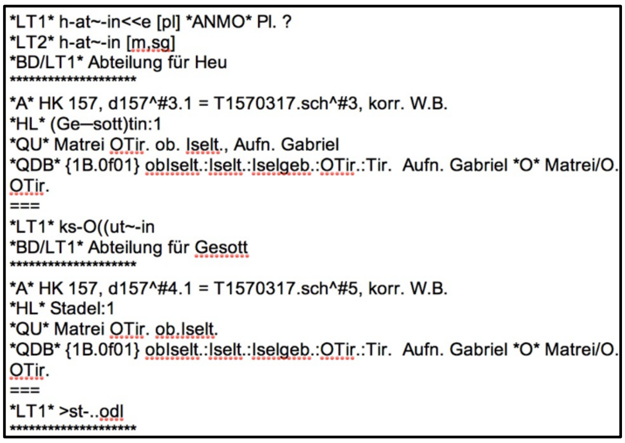
There is an increasing focus on semantic publishing of traditional data using LOD platforms. To support this, different approaches are used to enrich and expose the data stored in legacy databases semantically. One such approach, direct conversion, converts structured databases (usually relational databases and XML files) directly to RDF triples (Berners-Lee, 1998). This approach mainly uses the schema of the legacy system to transform the data. The transformed data, usually in a triple format (subject, predicate, object), is published as a separate service to the legacy data or as a new layer on top of the legacy database. This approach allows a mass conversion of legacy data without the need for analysis beyond the available schema. However, one of the drawbacks of this approach is that it is restricted to the semantics available within the data and adds little semantics other than the one contained in the schema (Simpson & Brown, 2013). This approach is mainly applicable for general collections but requires a detailed analysis when the domain of interest becomes specialised.

The alternative to this approach focuses on the analysis of the domain of interest and generate/select one or more ontologies that describe the semantics of entities and their relationships. This approach is more rigorous in that experts define the semantics of each entity and its properties. Besides, it facilitates inclusion of the domain knowledge of the experts and opens up a way of accommodating entities that are relevant to the domain but not included in the dataset. The downside of this approach is that it requires a certain level of domain-expert involvement and may require more effort and expert agreement. However, this approach provides a robust semantics and significantly contributes to interoperability.

In our work, we merge the two approaches and use schema analysis to identify entities, attributes and their relationships and domain analysis to analyse and describe the domain and to understand the rationale of the data collection method.

3.1. Schema Analysis

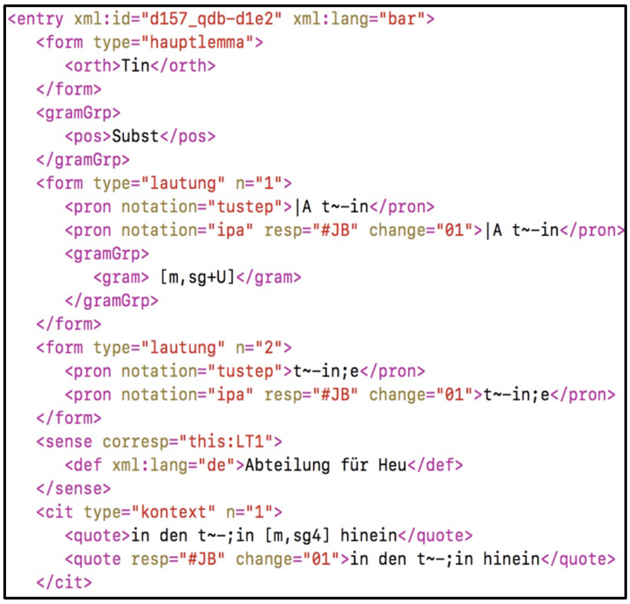
The availability of the dataset in various formats motivates us to look into schema analysis. The questionnaires are available as analogue paper copies, flat text files, in TEI/XML format and a relational table format (dbo@ema). The schema analysis of the available datasets provides us with valuable information to build our semantic model. Research (Ferdinand, C. Zirpins, & D. Trastour, 2004; Deursen et al., 2008; Battle, 2006) has shown that schema analysis provides significant information. The quality of the resulting semantic data, however, depends on the completeness and expressiveness of the available schema and does not reflect the meanings of the entities. In many cases, even if the structural information is available, accurate interpretation of the meaning conveyed by a given schema and its mapping to a standard vocabulary is difficult to achieve. For example, a relational schema which stores the year as "Year" requires accurate interpretation of whether the attribute "Year" refers to the year of publication of the questionnaire or the year it is distributed to data collectors or any other interpretation. Additionally, it requires an accurate description to resolve if "year" can be considered the same as "dcterms:date". Despite these drawbacks, schema analysis plays a significant role in identifying entities, attributes and their relationships.



```
*LT1* h-at~-in<<e [pl] *ANMO* Pl. ?
*LT2* h-at~-in [m,sg]
*BD/LT1* Abteilung für Heu
********************
*A* HK 157, d157^#3.1 = T1570317.sch^#3, korr. W.B.
*HL* (Ge—sott)tin:1
*QU* Matrei OTir. ob. Iselt., Aufn. Gabriel
*QDB* {1B.0f01} obIselt.:Iselt.:Iselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* ks-O((ut~-in
*BD/LT1* Abteilung für Gesott
********************
*A* HK 157, d157^#4.1 = T1570317.sch^#5, korr. W.B.
*HL* Stadel:1
*QU* Matrei OTir. ob.Iselt.
*QDB* {1B.0f01} obIselt.:Iselt.:Iselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* >st-..odl
********************
```

Figure 2. TUSTEP format

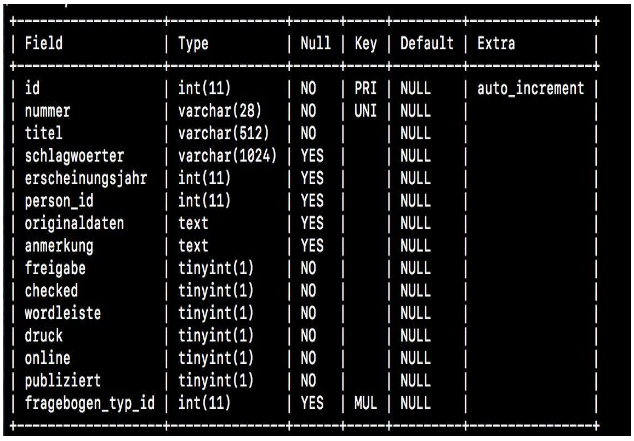
Schema description: through the life of the dataset, various software tools have been used to store and process the data. Currently, the software includes TUSTEP (Fig. 2), XML/TEI (Fig. 3) and MySQL (Fig. 4). Each of these tools keeps some schema of their own to describe the contents of the files. Having studied all these formats to understand the schema, we used the relational database schema as our main source containing 88 relational tables. In this paper, our focus is on the schema which is directly related to questionnaires (4), questions (2), authors (n=7) and answers (n=7).



```
<entry xml:id="d157_qdb-d1e2" xml:lang="bar">
   <form type="hauptlemma">
      <orth>Tin</orth>
   </form>
   <gramGrp>
      <pos>Subst</pos>
   </gramGrp>
   <form type="lautung" n="1">
      <pron notation="tustep">|A t~-in</pron>
      <pron notation="ipa" resp="#JB" change="01">|A t~-in</pron>
      <gramGrp>
         <gram> [m,sg+U]</gram>
      </gramGrp>
   </form>
   <form type="lautung" n="2">
      <pron notation="tustep">t~-in;e</pron>
      <pron notation="ipa" resp="#JB" change="01">t~-in;e</pron>
   </form>
   <sense corresp="this:LT1">
      <def xml:lang="de">Abteilung für Heu</def>
   </sense>
   <cit type="kontext" n="1">
      <quote>in den t~-;in [m,sg4] hinein</quote>
      <quote resp="#JB" change="01">in den t~-;in hinein</quote>
   </cit>
```

Figure 3. XML/TEI format

From the schema analysis, entities such as questionnaire (Fragebogen), types of questionnaires, questions (Frage), answers and authors are identified. Attributes of these entities and their data types are also identified.



```
+-------------------+---------------+------+-----+---------+----------------+
| Field             | Type          | Null | Key | Default | Extra          |
+-------------------+---------------+------+-----+---------+----------------+
| id                | int(11)       | NO   | PRI | NULL    | auto_increment |
| nummer            | varchar(28)   | NO   | UNI | NULL    |                |
| titel             | varchar(512)  | NO   |     | NULL    |                |
| schlagwoerter     | varchar(1024) | YES  |     | NULL    |                |
| erscheinungsjahr  | int(11)       | YES  |     | NULL    |                |
| person_id         | int(11)       | YES  |     | NULL    |                |
| originaldaten     | text          | YES  |     | NULL    |                |
| anmerkung         | text          | YES  |     | NULL    |                |
| freigabe          | tinyint(1)    | NO   |     | NULL    |                |
| checked           | tinyint(1)    | NO   |     | NULL    |                |
| wordleiste        | tinyint(1)    | NO   |     | NULL    |                |
| druck             | tinyint(1)    | NO   |     | NULL    |                |
| online            | tinyint(1)    | NO   |     | NULL    |                |
| publiziert        | tinyint(1)    | NO   |     | NULL    |                |
| fragebogen_typ_id | int(11)       | YES  | MUL | NULL    |                |
+-------------------+---------------+------+-----+---------+----------------+
```

Figure 4. MySQL format

Each attribute of the entities is examined for the relevance of the conveyed information in addition to the availability of usable data. There are attributes that contain null values for all records and columns with redundant information. For example, the attribute "wordleiste" ("MS Word Bar") in Fig. 4 contains empty values across all the records in the table. Such attributes are identified and presented to the domain experts for further analysis. There are also attributes that contain null values for some of the records and are left as they are, as there are possibilities to populate them from other sources. Expert evaluation categorised these attributes as "relevant", "needs further investigation" and "not relevant". We included the first two categories but discarded the "not relevant" ones. Finally, the entities and attributes are used as an input for preparing the semantic model.

3.2. Domain Analysis

Domain analysis serves as another step for understanding the rationale of the data collection and the data collection process itself. It provides a solid foundation about why, how, when and by whom the data was collected, stored and processed. It further provides a solid base for understanding the core entities of the datasets, the relationship among the entities and across other entities of similar purpose. Our approach starts with the study of primary sources of information, investigating and examining original materials, interviewing users and maintainers of the dataset. It also includes secondary sources to complement and clarify the domain knowledge.

Following the approach used by Boyce & Pahl (2007), the domain analysis stage seeks information related to 1) Purpose - the rationale of the data collection, 2) Source - the data collection method used, 3) Domain - the nature of the collected data, and 4) Scope - what are the core entities of interest.

Purpose: The purpose of the data collection is to document the wealth of diversity of rural life and unite it under a Pan-European umbrella with a special focus on German language and diverse nationalities in the late Austro-Hungarian Monarchy (Gura, Piringer, & Wandl-Vogt, forthcoming). The rationale of the data collection serves as a guidance for tuning our objectives and achieving the results. Thus, accordingly, our long-term interest is to capture the lexical data, represent it using standard vocabularies and interlink it with other collections.

Source: The primary data is collected using questionnaires with one or more questions. Questionnaires were distributed to the collectors, and the collectors filled the questionnaires by asking individuals and groups. In some cases, the collectors filled out the questionnaires themselves after observing teams of respondents. Then, collectors sent out the completed questionnaires to the centre where the data was further processed. The questions could be completed by one respondent or a group of respondents. In other cases, questions were filled by the data collectors themselves. Paper slips containing answers arrived at the centre even after several years and are stored in drawers alphabetically.

An interesting aspect of the domain analysis is the identification of the different question types which are not mentioned in any of the available schemas. A closer look at the questions resulted in the identification of patterns of questions used. The data collection is systematic in that it associates certain abbreviations to the questions that have asked similar types of questions. For example, phonological questions have abbreviations such as "*Aussprache, Ausspr.* or *Ltg.*" morphological questions have "*Komp.*" and synonym questions have "*Syn.* or *Synonym*" patterns. However, not all the questions have such abbreviations. The question types and their definitions are represented in detail in the next section. As the questions are linked to the answers, it is also possible to identify the different types of answers provided for a given question. The identification of question types by the domain experts will play a significant role for question-answering systems by exploiting these categories. However, modelling the answers is beyond the scope of this paper.

Domain: The primary data collected is lexical data in direct response to the questions of the questionnaire. It covers various aspects such as names, definitions, pronunciations, illustrations and other categories targeting a linguistic atlas and dictionary compilation (Arbeitsplan, 1912). However, there are other data generated during the process, including details of data collectors, the time and place of the data collection. Regarding the domain, the main interest is the linguistic data of historical and cultural importance.

Scope: From the above steps, we already identified the core entities contained in the datasets. These entities are defined and described by experts. The focus of this exercise is to use the questionnaires as the main entry point to semantically explore the data. Questionnaires contain individual questions of a particular topic which are linked to individual answers. However, in this paper, we will mainly focus on modelling questionnaires and the questions and explore obvious links to answers, authors, collectors and geographic locations. By doing so, we provide additional information which is relevant to answer research questions regarding gender-symmetry or spatiotemporal distributions. However, modelling the answers is complex and will not be discussed in detail in this paper. A pilot for modelling geographic locations is developed and treated separately (Scholz et al., 2016; Scholz, Hrastnig, & Wandl-Vogt, 2018).

4. Semantic Modelling

As a means of semantically enriching the datasets to publish it as a LOD, a semantic model was developed that incorporated the questionnaire model (Fig. 5) and the question model (Fig. 6) with a link to the associated entities. Both models are ontological models built using the Web Ontology Language (OWL2)[1] specification following ontological principles (Noy & Mcguinness, 2001; Edgar & Alexei, 2014). These models provide:

- A succinct definition of the entities and their relations,
- Interoperability with existing semantic resources to support LOD, and
- Extensibility to introduce new classes and relations.

There are many ontologies available to describe data of interest. These ontologies range from general purpose upper ontologies to lower, domain-specific ontologies to describe fine-grained knowledge for describing historical and cultural domains. After deciding the domain and the scope, the next step in the modelling stage is to consider reusing existing ontologies as this is preferable to developing an in-house ontology. However, for domain-specific description of datasets, it is difficult to find a suitable ontology and thus requires preparation either from scratch or extending existing ones.

We searched existing ontologies that can describe our domain of interest. The main repositories searched include LOV[2] ontology repository, Schema.org[3] and other specialised search tools such as Watson semantic web search engine.[4] We found terminologies related to questions, answers and questionnaires, but they do not fit our requirements, and such ontologies are not available yet. However, we will exploit some of the concepts defined in the Ontolex-Lemon model (McCrae et al., 2017) to describe the lexical data in the collection. We will further reuse vocabularies such as FOAF, SKOS and Dublin Core to describe authors, editors, collectors, places and publication. In addition to describing the entities, generic ontological constructs are used to create an interlinking with concepts from other repositories, and to compare our data with other similar data sets using meaningful interoperability.

A combination of top-down and bottom-up approaches as proposed by Uschold & Gruninger, (1996) is used to develop the model. The approach integrated domain analysis as a top-down approach and schema analysis as a bottom-up approach to build the ontology, in order to support our domain-specific requirements. We also used existing standardised vocabularies for entities that already have compatible representations. We developed our

[1] https://www.w3.org/TR/owl2-primer/

[2] http://lov.okfn.org

[3] http://schema.org

[4] http://watson.kmi.open.ac.uk/WatsonWUI/

ontology to represent both the structure and the meaning of the entities of interest.

4.1. Questionnaire Model

The questionnaire model is built based on the detailed analysis of the original and physically compiled book of sets of questionnaires and its electronic version (dbo@ema). Up to now, we have identified three questionnaire types. Each type has its characteristics and differs from the others in its purpose, the type of information it seeks and its format, including its physical appearance. Treating the different sets of questionnaires independently is crucial to preserve the historical importance and the structural and semantic relation each questionnaire set has with the collected data. The questionnaire types are discussed below:

1. Systematic: [*Systematischer Fragebogen*] is a questionnaire that is used to collect the original data. This type of questionnaire is used from the beginning of the data collection process.
2. Additional: [*Ergänzungsfragebogen*] is a questionnaire that is used as a supplementary questionnaire to the systematic questionnaire.
3. Dialectographic *[Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen] is a* questionnaire of the Munich and Vienna Dictionary Commissions.

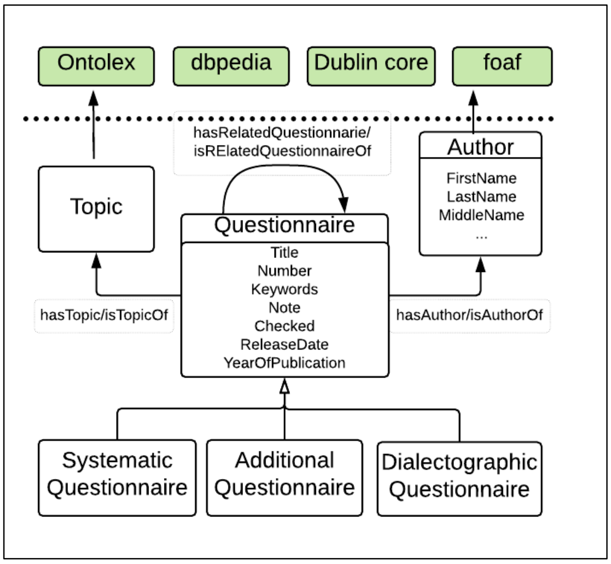


Figure 5. A Semantic model of questionnaire

A questionnaire may have one or more related questionnaires that deal with the same topic. We observed that questionnaires refer to other questionnaires. Such relationships are captured by an object property "hasRelatedQuesitonnaire" with an inverse property "isRelatedQuestionnaireOf". A questionnaire has at least one topic, and this relationship is captured by "hasTopic" with "isTopicOf" inverse object property. Furthermore, a questionnaire has at least one Author, and this relationship is captured by "hasAuthor" object property with "authorOf" inverse object property.

Topics (Questionnaire Topics). A topic is the main subject of the questionnaire or a given question. A questionnaire may focus on a general topic such as "Food" and a question may cover subtopics such as "Traditional Food". This information will be treated as a topic following a proper disambiguation technique and then relate to ontolex:lexicalConcept.

Author/collectors. Authors are defined in FOAF and Dublin Core. We will reuse the definition provided in FOAF Agent/Author classes.

4.2. Question Model

A question is a linguistic expression used to request information, or the request made using such an expression. The information requested is provided in the form of answer.[5] In this ontology, we categorise the questions mainly based on the content, the forms and the expected answers from the respondents. An analysis carried out by the experts, users and ontology engineers identified 12 different types of questions and added two more questions to accommodate future processing of additional questionnaire sets. It is important to note that these question types are not mutually exclusive to one another and there are instances of questions that belong to more than one type of questions, e.g. the question "*Kopf: Kopf/Haupt (in urspr. Bed.) in Vergl./Ra. (Kopf stehn, der Kopf mÌ¦chte einem zerspringen)*" is both semasiological and syntactic. The semantics of the question types are given below:

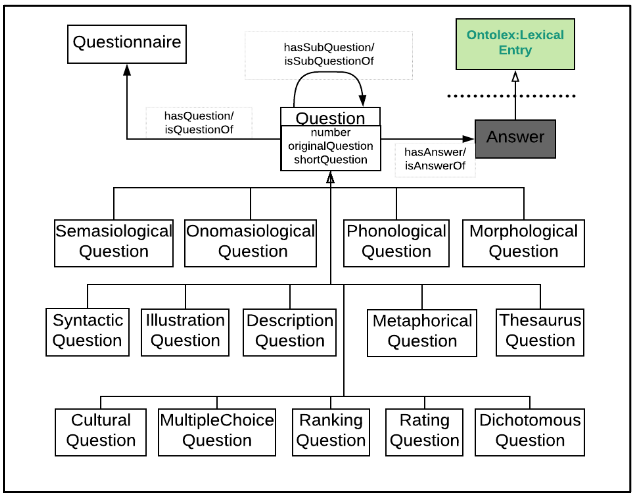


Figure 6. A Semantic model of question

1. *Onomasiological:* asks for the name of a given entity, e.g. "how do you call x?" where x represents an entity.
2. *Semasiological:* asks for the meaning of a given entity, e.g. "what does x mean?".
3. *Dichotomous:* asks for a selection of answers from a binary option. It includes yes/no or agree/disagree types of answers to stated questions.
4. *Description:* asks for a written representation of a given entity, e.g. "What would be the function of x?".

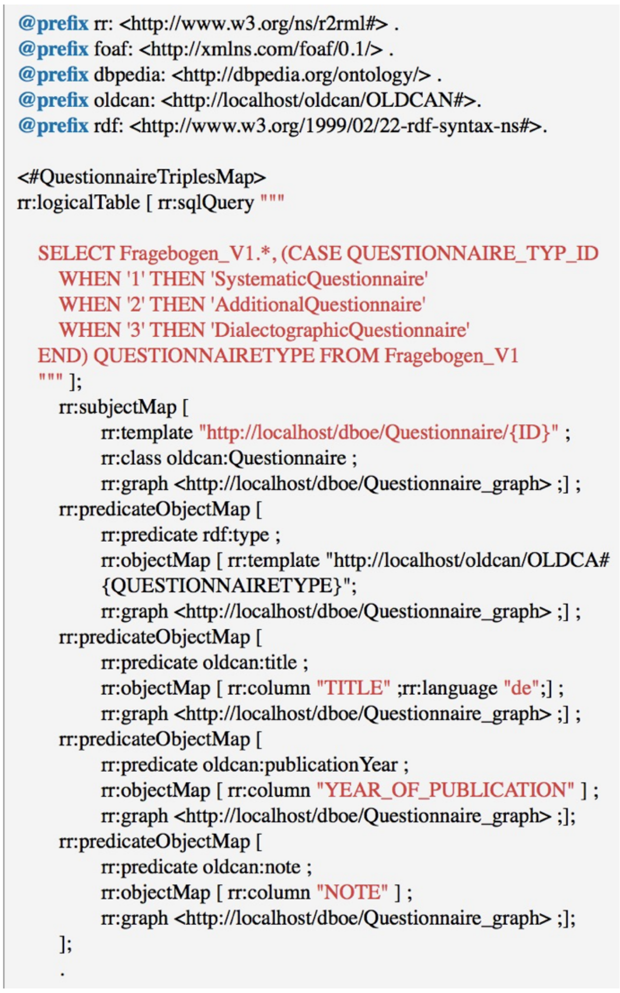
[5] https://en.wikipedia.org/wiki/Question

5. *Illustration:* asks for a pictorial or diagrammatic representation of a given entity, e.g. "What does x look like?".
6. *Morphological:* asks about the structure and the formation of words and parts of words. Based on the structure, morphological questions can take various forms.
7. *Phonological:* asks for the pronunciation, or phonetic representation of words.
8. *Syntactic:* asks for construction of phrases or sentences using a given word or a given idiom, e.g. "Provide a phrase/sentence for/using a word/idiom x".
9. *Metaphorical:* asks for some conveyed meanings given a word or an expression. Metaphorical questions are related to semasiological questions, but they ask for an additional interpretation of the expression beyond its obvious meaning.
10. *Thesaurus:* asks for a list of words or expressions that are used as synonyms (sometimes, antonyms) or contrasts of a given entity.
11. *Cultural:* asks for a belief of societies, procedures on how to make or prepare things and how to play games, contents of cultural songs, poems used for celebrations. Analysis of the existing questions shows that the cultural question type has its subtypes and has instances that significantly overlap with the other question types.
12. M*ultiple Choice:* asks for a selection of one item from a list of three or more potential answers.
13. *Rating:* asks the respondent to assign a rate (degree of excellence) to a given entity based on a predefined range
14. *Ranking Question:* asks the respondent to compare entities and rank them in a certain order.

It is commonly observed that a question may ask several other sub-questions, and this is captured by the "hasSubQuestion" object property. Thus, the object property "hasSubQuestion" relates one question with its subquestions. Each question is linked to its associated answer. A question may have several answers collected from different sources. This is captured by the "hasAnswer" object property with its inverse "isAnswerOf". Finally, a question is related to a questionnaire with the "isQuestionOf" object property where a single question is contained only in one questionnaire.

Answer: An answer is a written, spoken or illustrated response to a question. The different types of questions have answers either in a written, spoken or illustration format. In the case of questions that involve lexical data collection, the answer could be associated with some lexical category. For each types of questions, there are different types of answers including sentences, individual words, multiword expressions, affix, diagrams, etc. Modelling the answers is under investigation. However we will treat answers with single word, multiword expression or affixes as ontolex:lexicalEntries. For example, the answer to a thesaurus question is expected to be a word, or multiword expression in the OntoLex model.

Finally, an initial version of the ontology- Ontology for Lexical Data Collection and ANalysis (OLDCAN)[6] is developed following the approach discussed above. Since the project is at its development stage, a permanent URL has not been yet assigned to either the ontology or to the data. However, the ongoing results are available under a Creative Commons Licensing.[7]



```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia: <http://dbpedia.org/ontology/> .
@prefix oldcan: <http://localhost/oldcan/OLDCAN#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<#QuestionnaireTriplesMap>
rr:logicalTable [ rr:sqlQuery """

  SELECT Fragebogen_V1.*, (CASE QUESTIONNAIRE_TYP_ID
    WHEN '1' THEN 'SystematicQuestionnaire'
    WHEN '2' THEN 'AdditionalQuestionnaire'
    WHEN '3' THEN 'DialectographicQuestionnaire'
  END) QUESTIONNAIRETYPE FROM Fragebogen_V1
  """ ];
    rr:subjectMap [
        rr:template "http://localhost/dboe/Questionnaire/{ID}" ;
        rr:class oldcan:Questionnaire ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate rdf:type ;
        rr:objectMap [ rr:template "http://localhost/oldcan/OLDCA#
        {QUESTIONNAIRETYPE}";
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:title ;
        rr:objectMap [ rr:column "TITLE" ;rr:language "de";] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:publicationYear ;
        rr:objectMap [ rr:column "YEAR_OF_PUBLICATION" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    rr:predicateObjectMap [
        rr:predicate oldcan:note ;
        rr:objectMap [ rr:column "NOTE" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    ];
    .
```

Figure 7. R2RML mapping excerpts

5. Semantic Up-lift

This stage focuses on the use of the semantic model and selected vocabularies to semantically enrich the data. It is used to annotate every data element with semantic information that states what it is, how it should be interpreted and how it is related to other elements within the datasets or across other datasets. There are various methods and tools used to transform relational databases to semantically compatible formats including direct mapping (Berners-Lee, 1998) and domain semantics-driven mapping (Michel, Montagnat, & Faron, 2013). We followed R2RML[8] to annotate our datasets due to its customisability for mapping relational databases into triples. Unlike direct mapping that depends on the database's structure, it is possible to use an ontology of the domain. Since R2RML is a vocabulary by itself, it stores

[6] http://exploreat.adaptcentre.ie/#Semantics

[7] https://creativecommons.org/licenses/by/3.0/at/deed.en

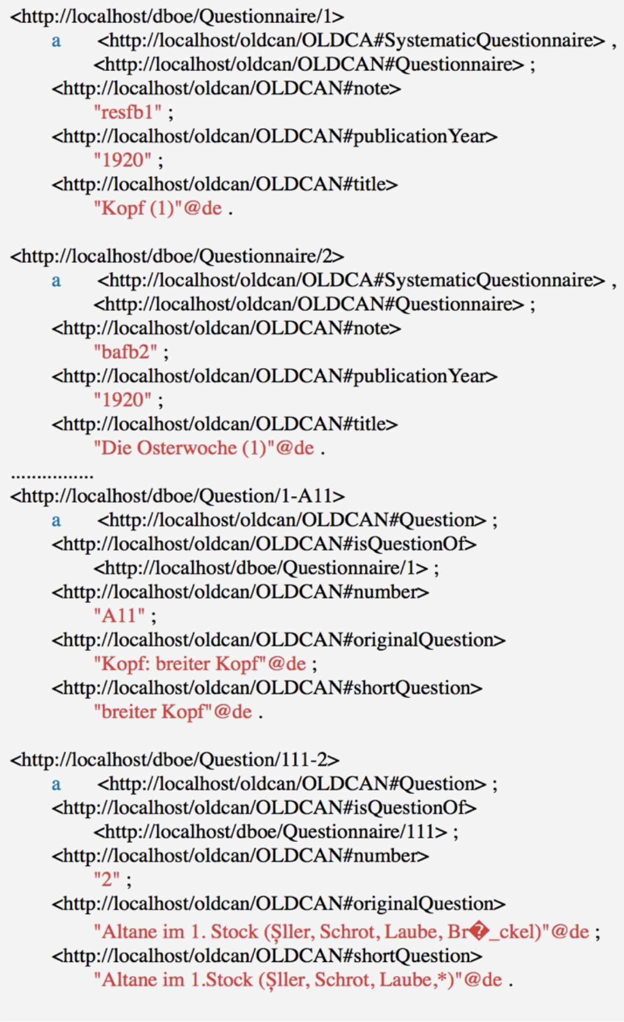
[8] https://www.w3.org/TR/r2rml/

the mappings from a relational database to RDF as RDF files and allows inclusion of provenance information. This facilitates knowledge discovery and reuse of mappings. However, it requires more effort compared to direct mapping. R2RML is used to map the relational data into a LOD. This phase includes the following steps:

1. Converting the major tables into classes,
2. Mapping object property relationships,
3. Mapping data property relationships,
4. Enriching the data with additional semantics.

To demonstrate the envisioned mapping, excerpts of the mapping file for both questionnaire and questions are generated. In the mapping (Fig. 8), each questionnaire is associated to oldcan:Questionnaire class using "a" ("rdf:type") property. The template defines the URL of the specific location of the questionnaire. The selected attributes are mapped to data properties, e.g. title is mapped to oldcan:title and the language of the title is included using a language tag “de”.

The mapping of the questions is done similarly. Here the object property isQuestionOf is used to link the question with its questionnaire. In the ontology, the hasQuestion object property is defined as an inverse of isQuestionOf to achieve both brevity and searchability in the generated data. The different types of the questionnaires and the questions are captured. An excerpt of the resulting triple[9] is presented in Fig. 8.



```
<http://localhost/dboe/Questionnaire/1>
      a       <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
              <http://localhost/oldcan/OLDCAN#Questionnaire> ;
      <http://localhost/oldcan/OLDCAN#note>
              "resfb1" ;
      <http://localhost/oldcan/OLDCAN#publicationYear>
              "1920" ;
      <http://localhost/oldcan/OLDCAN#title>
              "Kopf (1)"@de .

<http://localhost/dboe/Questionnaire/2>
      a       <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
              <http://localhost/oldcan/OLDCAN#Questionnaire> ;
      <http://localhost/oldcan/OLDCAN#note>
              "bafb2" ;
      <http://localhost/oldcan/OLDCAN#publicationYear>
              "1920" ;
      <http://localhost/oldcan/OLDCAN#title>
              "Die Osterwoche (1)"@de .
................
<http://localhost/dboe/Question/1-A11>
      a       <http://localhost/oldcan/OLDCAN#Question> ;
      <http://localhost/oldcan/OLDCAN#isQuestionOf>
              <http://localhost/dboe/Questionnaire/1> ;
      <http://localhost/oldcan/OLDCAN#number>
              "A11" ;
      <http://localhost/oldcan/OLDCAN#originalQuestion>
              "Kopf: breiter Kopf"@de ;
      <http://localhost/oldcan/OLDCAN#shortQuestion>
              "breiter Kopf"@de .

<http://localhost/dboe/Question/111-2>
      a       <http://localhost/oldcan/OLDCAN#Question> ;
      <http://localhost/oldcan/OLDCAN#isQuestionOf>
              <http://localhost/dboe/Questionnaire/111> ;
      <http://localhost/oldcan/OLDCAN#number>
              "2" ;
      <http://localhost/oldcan/OLDCAN#originalQuestion>
              "Altane im 1. Stock (Şller, Schrot, Laube, Br�_ckel)"@de ;
      <http://localhost/oldcan/OLDCAN#shortQuestion>
              "Altane im 1.Stock (Şller, Schrot, Laube,*)"@de .
```

Figure 8. Questionnaire and question triples

6. Conclusion and Future Work

The effort to open up legacy databases to make them accessible, usable and researchable has increased with the development of LOD platforms. Such platforms facilitate publishing legacy data of a wide range of contents and formats. As the content becomes specialised, the need for finding and developing semantic models that describe the domain of interest become crucial. This paper has presented an approach which is currently used for building a semantic model for enriching and publishing traditional data of historical, cultural and lexical importance. It is argued that the use of such an approach for building semantic models to assist with semantic publishing of traditional data on the LOD platform is vital to the exploitation of data of historical importance. It further paves the way for researchers to understand and compare conceptualisation of entities at different times and their evolution through time. As the paper presents work in progress, our immediate focus is the enrichment of the semantic model by in-depth examination of the entities including answers to the questions to enable a strong semantic interlinking that will facilitate efficient question answering and comparison of the different types of questions. Furthermore, additional enrichment to interlink the data with other similar datasets and the visualisation of the dataset will be the next area to tackle.

7. Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! project and the ADAPT Centre for Digital Content Technology at Dublin City University which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

Bibliographical References

Arbeitsplan (1912). *Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912.* Karton 1. Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch. Archive of the Austrian Academy of Sciences. Wien.

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. & Schwaiger, S. (2010). Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In H. Bergmann, M. M. Glauninger, E. Wandl-Vogt, & S. Winterstein (Eds.), *Fokus Dialekt. Analysieren – Dokumebattentieren – Kommunizieren. Festschrift für Ingeborg Geyer zum 60. Geburtstag.* (Germanistische Linguistik 199–201). Hildesheim, Zürich, New York: Olms, (pp. 47–64).

Battle, S. (2006). Gloze: XML to RDF and back again. *First Jena User Conference.* Bristol, UK.

Beretta, F., Ferhod, D., Gedzelman, S. & Vernus, P. (2014). The SyMoGIH project : publishing and sharing historical

[9] http://exploreat.adaptcentre.ie/#APIs

data on the semantic web. *Digital Humanities 2014, July 2014*, Lausanne, Switzerland. (pp. 469–470).

Berners-Lee, T. (1998). *Relational Databases on the Semantic Web.* In *Design Issues for the World Wide Web.* Retrieved January 10, 2018, from https://www.w3.org/DesignIssues/RDB-RDF.html

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web Information Systems, 5*(3), 1–22.

Boyce, S. & Pahl, C. (2007). Developing Domain Ontologies for Course Content. *Educational Technology & Society, 10*, 275–288.

Chiarcos, C., Cimiano, P., Declerck, T. & McCrae, J. P. (2013). Linguistic Linked Open Data (LLOD) – Introduction and Overview. In C. Chiarcos, P. Cimiano, T. Declerck, & J. P. McCrae (Eds.), *2nd Workshop on Linked Data in Linguistics.* Pisa, (pp. i–xi).

Doerr, M. (2009). Ontologies for Cultural Heritage. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies. International Handbooks on Information Systems.* Berlin, Heidelberg: Springer.

Ferdinand, M., Zirpins, C. & Trastour, D. (2004). Lifting XML Schema to OWL. In N. Koch, P. Fraternali, & M. Wirsing (Eds.), *Web Engineering. 4th International Conference, ICWE 2004. Munich, Germany, July 26-30, 2004. Proceedings.* (Lecture Notes in Computer Sciences 3140). Berlin, Heidelberg: Springer, (pp. 354–358).

Gura, C., Piringer, B. & Wandl-Vogt, E. (forthcoming). Nation Building durch Großlandschaftswörterbücher. Das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) als identitätsstiftender Faktor des österreichischen Bewusstseins.

Kansa, E. C., Kansa, S. W., Burton, M. M., & Stankowski, C. (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies, 6*(2), 301–326.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Leiden: Lexical Computing CZ, (pp. 587–597).

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. & Van Harmelen, F. (2015). Semantic Technologies for Historical Research: A Survey. *Semantic Web, 6*(6), 539–564.

Michel, F., Montagnat, J., & Faron Zucker, C. (2013). *A survey of RDB to RDF translation approaches and tools.* Retrieved January 16, 2018, from https://hal.archives-ouvertes.fr/hal-00903568v1

Noy, N. F., & McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology.* Technical, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Scholz, J., Hrastnig, E., & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In P. Fogliaroni, A. Ballatore & E. Clementini (Eds.), *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017).* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 275–282).

Scholz, J., Lampoltshammer, T. J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Indeterminate and Crisp Boundaries. In G. Gartner, M. Jobst, & H. Huang (Eds.), *Progress in Cartography.* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 133–151).

Schopper, D., Bowers, J. & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a data-base of dialects into a rich LOD resource. Retrieved January 17, 2018 from *Text Encoding Initiative Conference and members' meeting 2015, October 28-31, Lyon, France. Papers.*

Serna Montoya, E., & Serna Arenas, A. (2014). Ontology for knowledge management in software maintenance. *International Journal of Information Management, 34*(5), 704–710.

Simpson, J. & Brown, S. (2013). From XML to RDF in the Orlando Project. In *Proceedings. International Conference on Culture and Computing. Culture and Computing 2013. 16-18 September 2013.* Kyoto: IEEE Xplore Digital Library, (pp. 194–195).

Uschold, M. & Gruninger, M. (1996). Ontologies: Principles, methods, and applications. *Knowledge Engineering Review, 11*(2), 93–155.

Van Deursen, D., Poppe, C., Martens, G., Mannens, E. & Van de Walle, R. (2008). XML to RDF Conversion: A Generic Approach. In P. Nesi, K. Ng & J. Delgado (Eds.), *Proceedings. Fourth International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution. Florence, Italy. 17 – 19 November 2008.* IEEE Xplore Digital Library, (pp. 138–144).

Wandl-Vogt, E. (2008). …wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen). In P. Ernst (Ed.), *Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. – 23. September 2006.* Wien: Praesens, (pp. 93–112).

Wandl-Vogt, E. (2012). *Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema).* Retrieved January 17, 2018 from https://wboe.oeaw.ac.at/projekt/beschreibung/

WBÖ (1970–2015). *Wörterbuch der bairischen Mundarten in Österreich. Bayerisches Wörterbuch: I. Österreich.* 5 vols. Ed. by Österreichische Akademie der Wissenschaften. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Language Resource References

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria] (DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010) (Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema). Wien. [Processing status: 2018.01.]